Chapter 17
Facial Retouching and Alteration Detection

Puspita Majumdar, Akshay Agarwal, Mayank Vatsa, and Richa Singh

Abstract On the social media platforms, the *filters* for digital retouching and face beautification have become a common trend. With the availability of easy-to-use image editing tools, the generation of altered images has become an effortless task. Apart from this, advancements in the Generative Adversarial Network (GAN) leads to creation of realistic facial images and alteration of facial images based on the attributes. While the majority of these images are created for fun and beautification purposes, they may be used with malicious intent for negative applications such as deepnude or spreading visual fake news. Therefore, it is important to detect digital alterations in images and videos. This chapter presents a comprehensive survey of existing algorithms for retouched and altered image detection. Further, multiple experiments are performed to highlight the open challenges of alteration detection.

17.1 Introduction

Social media platforms have become the new source of information, and millions of images and videos are uploaded and shared on these platforms on a daily basis.

P. Majumdar performed the experiments and analyzed the results. A. Agarwal helped in writing the chapter. All authors have reviewed and updated the chapter

P. Majumdar
Puspita Majumdar IIIT-Delhi, Delhi, India
e-mail: pushpitam@iiitd.ac.in

A. Agarwal
Akshay Agarwal SUNY, University at Buffalo, Buffalo, NY, USA
e-mail: aa298@buffalo.edu

M. Vatsa
Mayank Vatsa IIT Jodhpur, Jodhpur, India
e-mail: mvatsa@iitj.ac.in

R. Singh (✉)
Richa Singh IIT Jodhpur, Jodhpur, India
e-mail: richa@iitj.ac.in

C. Rathgeb et al. (eds.), *Handbook of Digital Face Manipulation and Detection*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-87664-7_17

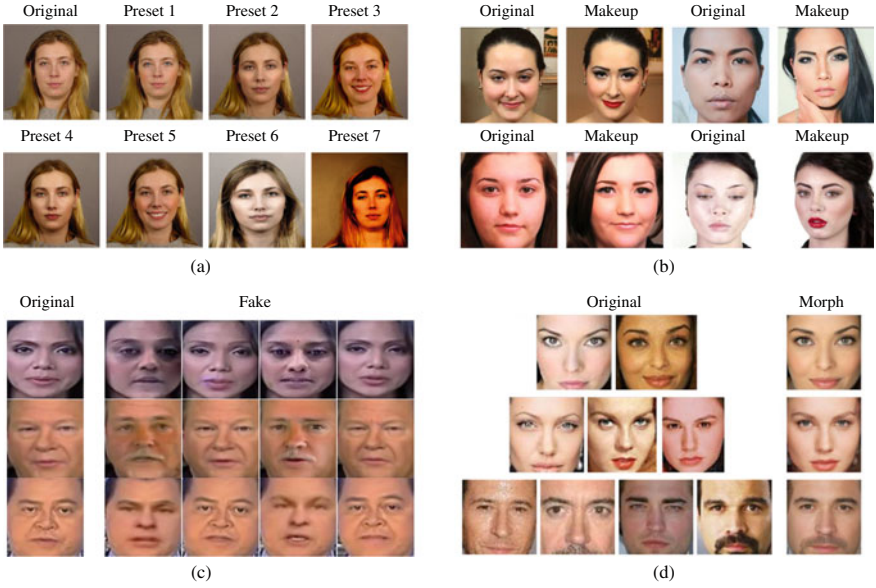

Fig. 17.1 Samples of different facial alterations. **a** Retouching **b** Makeup **c** DeepFakes, and **d** Morphing

While uploading images or sharing them among individuals, the face images are generally retouched/altered to make them look more beautiful or appealing due to the fascination toward few societal factors such as fair complexion and flawless skin [1]. As shown in Fig. 17.1, these alterations can either be in the form of simple retouchings such as removal of pimples, age spots, and wrinkles to complex alterations such as *morphing* or *deepfake* that change the geometric properties.

In cosmetic industries, facial retouching/alteration is commonly used to sell beauty products by making the seller (model) look more appealing in advertisements. These advertisements convey the wrong information of obtaining a flawless appearance upon using their beauty products, which in turn mislead people to use their beauty products. Digitally retouched images can also adversely affect the mindset of the general population and can lead to mental stress [57]. It negatively affects the self-esteem of the viewers by trying to follow the societal norm of pleasant appearance. This leads to body dissatisfaction amongst women and sets unrealistic expectations among them, which leads to various psychological and sociological issues. To cope with the situation, some countries have enacted the "Photoshop Law" to label retouched advertisement photos as retouched [48].

The effect of retouching on face recognition algorithms cannot be ignored. Several countries require hard copy of photographs on identification documents such as driver's license and passports. Generally, people digitally retouch their images and use the prints for application. These images are used to create the identification

documents and may serve as an enrollment image to be matched with real-time query images of a subject. The real-time original images, when matched with enrolled retouched images, degrade the identification performance [9, 53].

Apart from digital retouching, alterations on face images can be in the form of (i) morphing, (ii) attribute modification via GANs, and (iii) deepfakes. In morphing, a new face image is generated using the information available from two or more source face images to conceal own identity or gain the identity of others [5, 44, 60]. GANs based techniques alter the local or global facial attribute of the input face images [32, 33]. In deepfakes, altered videos are generated by face swapping or facial reenactment techniques [56]. With the availability of online tools and apps for performing these alterations flawlessly and effortlessly, anyone can create altered samples.

The effect of altered images in facial recognition algorithms and their use for spreading fake news is a major concern. It is shown that morphed images significantly reduce the performance of face recognition algorithms, including commercial systems and deep neural networks-based models [23, 44]. Their adverse effect can be seen in the application of automatic border access through e-passport. Generally, while issuing e-passports, a hard copy of the photograph is required. The user can provide the morphed photograph to fool both human examiner and automatic face recognition algorithms. Apart from this, spreading fake news using deepfakes is a serious challenge. For example, deepfakes [55] can be used to create fake videos that show celebrities in pornographic content by generating an individual's face that closely matches with another face in the video. Fake videos of Mr. Barack Obama were widely circulated on the Internet [64]. Often, generative models are used for creating such content and can be done in real-time by swapping faces along with their facial expressions [70]. The problem becomes severe when these altered images/videos are presented as evidence in the courts or are used during political campaigns. It is therefore important to detect the altered face images [10, 32, 33, 54].

The outline of this chapter is as follows. Section 17.2 discusses the literature of different algorithms proposed for the detection of retouched and altered images. This section further provides the details of the databases proposed for retouching and alteration detection. A thorough experimental evaluation of the performance of existing algorithms to detect retouched and altered images in cross-domain/manipulation settings is discussed in Sect. 17.3. In Sect. 17.4, we highlight the open challenges that require the attention of the research community and focused research efforts, followed by the conclusion in Sect. 17.5.

17.2 Retouching and Alteration Detection—Review

In the literature, researchers have proposed different techniques for detecting facial retouching and alterations. While retouching is done for an appealing appearance without any ill intent, alterations such as morphing and face swap are generally

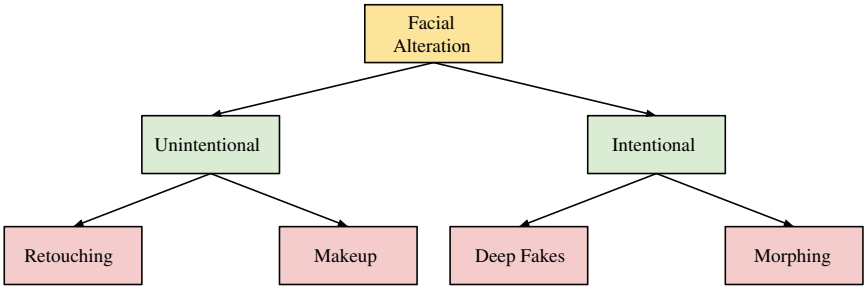

Fig. 17.2 Categorization of facial alterations into unintentional and intentional adversary

done with malicious intent. Therefore, as shown in Fig. 17.2, we have segregated the literature into unintentional and intentional adversary detection. In the following subsections, we discuss the algorithms proposed for the detection of retouched and altered images, followed by the details of the publicly available databases for the same.

17.2.1 Digital Retouching Detection

Retouching on facial images can be performed digitally using easy-to-use image editing tools or physically by applying facial makeups. Retouching is done for beautification purposes, generally, without any malicious intent and can be categorized as *unintentional adversary*. However, due to the adverse effect of self-acclaimed ideal face complexion and an appealing appearance by retouching of face images on social media applications, countries such as Israel, UK, and USA [25, 63, 65] have enacted laws to regulate the use of retouched images. For the strict adhesion of such laws, the successful detection of digitally retouched images is important. To facilitate research in this direction, researchers have proposed different algorithms to create retouched images and analyzed their effect on face recognition algorithms, followed by designing different algorithms for its detection.

In 2011, Kee et al. [36] proposed an amalgamation of photometric and geometric features for an effective retouching of face and body images. Later, Ferrara et al. [22, 23] evaluated the impact of face retouching or beautification on commercial and handcrafted features based face recognition algorithms. In the earlier work, Ferrara et al. [23] have performed multiple levels of beautification and studied its impact on the equal error rate (EER) of the commercial face recognition systems. It is shown that even with the slight beautification, the EER of the system changes by $\sim$2%, whereas heavy retouching can increase the EER by $\sim$17%. In 2016, Bharati et al. [9] created one of the largest databases both in terms of the number of subjects and

type of retouching mediums. The performance of commercial face recognition systems is evaluated on the proposed database. The authors have reported a difference of ∼7.50% and ∼11% in the rank-1 recognition performance of the commercial system and openBR [37], respectively. Further, an algorithm is proposed for detecting retouched images using face patches as input in the deep Boltzmann machine (DBM) for feature extraction and support vector machine (SVM) for binary classification. Experiments are performed on two databases and the proposed algorithm achieved an overall accuracy of 87.10% on the ND-IIITD database and 96.20% on the Celebrity database. These preliminary works highlight the challenges of recognizing retouched face images. Bharati et al. [10] have further created a demography-based retouched face database. The database contains subjects belonging to different gender groups and ethnicity. The authors have also proposed a retouching face detection algorithm based on supervised autoencoder. The experiments are performed with both seen and unseen demographic ethnicity in the training and testing sets. The Caucasian demographic subset yields the lowest detection performance even under seen demographic experimental setting. The performance of the detection algorithm is at-least 2% lower under unseen demographic experimental scenario than the seen demographic scenario. Jain et al. [33] have used the softmax probabilities as the features in the SVM classifier for retouched face detection. Recently, the authors [32] have proposed a multi-level hierarchical framework for the detection of original and altered images. Altered images are further classified into retouched and GANs generated images. Rathgeb et al. [54] have proposed a differential detection approach based on the assumption that while detecting a retouched image, a counter trusted original image is also available. Three difference vectors are computed using texture features, facial landmarks, and featured from deep neural networks. A support vector machine classifier is trained on each difference vector, and a weighted fusion is performed for decision. A critical drawback is the assumption of the availability of the trusted source and its characteristics. While the previous works performed the binary classification of original and retouched images, a recent work by Wang et al. [68] have proposed a framework to first perform the retouching detection and later suggested a possible undo operation to develop the unaltered image. For binary classification dilated residual network is trained using heavy data augmentation techniques. On the detected manipulated images, optical flow field is calculated for measuring the pixel warping effect.

Another related field to digital retouching is facial cosmetics or makeup, i.e., physical retouching. According to multiple market reports, the business of cosmetics is growing exponentially. For example, the US market growth is at the rate of CAGR of 2.47% from the year 2015 to 2020 [67]. Makeups drastically alter the facial appearance of a person and are applied to various facial regions such as eyes, skin, and lip. Similar to digital retouching, makeups also affect the performance of face recognition algorithms. Several researchers have shown the impact of facial makeup in the performance degradation of face recognition algorithms, including commercial systems [18, 29, 62, 69]. To counter the impact of facial makeup on recognition, several algorithms have been proposed to detect makeup images. Chen et al. [12] have utilized the SVM and AdaBoost classifier trained on the fusion of shape and color

features for detecting makeup images. Kose et al. [38] proposed an ensemble-based technique, and Liu et al. [42] have used the entropy information combined with SVM for makeup image detection. Kotwal et al. [39] have utilized the intermediate layer features of deep convolutional neural network (CNN) for age-induced makeup detection. The authors have also proposed a new facial makeup database with both male and female individuals. It is shown that the age-induced makeup can significantly degrade the performance of face recognition network, namely LightCNN [72]. Apart from the simple classification of images as with and without makeup, research works have also been proposed for the removal of makeup to obtain non-makeup images. Cao et al. [11] have proposed a generative adversarial network, namely, bidirectional tunable de-makeup network (BTD-Net) for makeup removal. Arab et al. [6] have proposed a two-level defense against the makeup-based alteration. In the first level, images are first detected for makeup or non-makeup. Later, the makeup removal algorithm is proposed utilizing Cycle generative adversarial network (Cycle GAN) [74]. The authors have shown a significant improvement in the rank-1 face matching accuracy through their makeup removal technique, surpassing several existing algorithms, including BTD-Net. Rathgeb et al. [53] presented a survey of the impact of beautification on face recognition algorithms and different detection techniques.

17.2.2 Digital Alteration Detection

Digital alterations, including morphing, GANs based alterations, and deepfakes are performed with malicious intent and fall under the category of the *intentional adversary*. With the advancements in computer vision and deep learning algorithms, digitally altering/manipulating an image/video has become an easy task. Altered/manipulated images raise serious concerns when used for illegal access, spreading fake news during political campaigns, or as evidence in court. This has attracted the attention of the research community, and several algorithms have been proposed for the generation and detection of altered images. Agarwal et al. [5] have prepared a large scale video-based face swap database using Snapchat. Face swap is an alteration technique in which more than one individual can share a single identity. The authors have shown the vulnerabilities of commercial face recognition systems, and mobile unlocking algorithms against face swapped images. A novel feature descriptor is also proposed to highlight the minute inconsistencies near eyes, nose, and mouth regions. The feature descriptor is then fed into the SVM classifier for binary classification. Other types of alterations include the creation of a new face image by blending multiple faces based on the measurement of facial landmarks [13, 58, 71]. The detection and blending of facial landmarks are performed using different algorithms. In an early attempt to secure the face recognition algorithms against such alterations, researchers have proposed different image features based detection algorithms [34, 59, 61]. Recent detection algorithms against such alterations are based on the characteristics of facial landmarks, head pose [3, 73] and

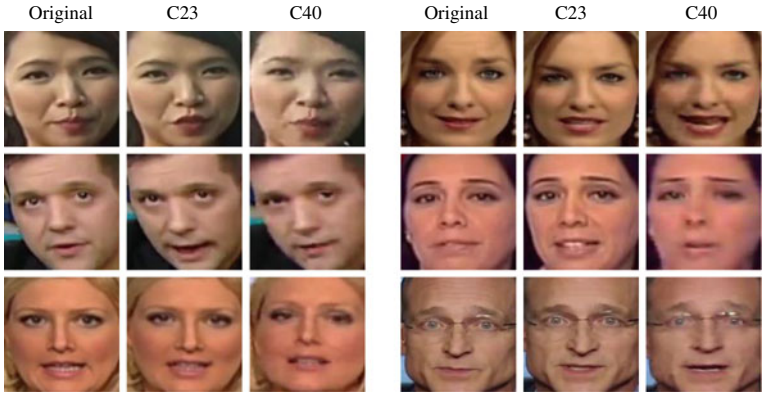

Fig. 17.3 Illustrating the difference between the compressed and uncompressed frames extracted from the original videos of the FaceForensics++ dataset [56]

eye blinking [35]. For detecting GANs based alterations, Jain et al. [32] proposed a three-level hierarchical network, Digital Alteration Detection using Hierarchical Convolutional Neural Network (DAD-HCNN). The proposed network not only distinguishes altered images from original ones but also classifies the images generated using different models of GANs.

With the advancement of generative adversarial networks (GANs), the generation of face swapping and morphing became an easy task. GANs lead to the generation of high resolution manipulated face images such as deepfakes. In deepfakes, the face of a person in a video is swapped with another person (face swapping), or someone's expression is animated over the person in the video (facial reenactment). Face swapping techniques can be broadly divided into two groups: *(i) computer graphics-based techniques* and *(ii) deep neural network-based techniques*. Computer graphics techniques are based on detecting facial landmarks and merging these landmarks for the generation of swapped faces. Deep neural network-based techniques automatically identify the pose and other related information for swapped face generation. To motivate research toward the detection of deepfakes, Facebook has recently organized the Deepfake Detection Challenge (DFDC) [19]. Rossler et al. [56] have proposed one of the largest databases (FaceForensics++) covering different manipulation types generated using computer graphics-based techniques and GANs. The videos in the proposed database are available in three different qualities. Figure 17.3 shows the difference between the compressed and uncompressed frames extracted from the original videos. Authors have evaluated the performance of existing alteration detection algorithms and deep CNN models on the FaceForensics++ database. It is found that XceptionNet [14] outperformed existing algorithms. It is observed that the detection of altered, compressed videos are challenging than uncompressed videos. Dang et al. [17] have proposed an attention-based network utilizing the features of CNN networks for fake detection. Kumar et al. [40] have utilized the patch-based ResNet architecture for the detection of face manipulation videos. Recently, Ciftci et al. [15]

have proposed to use biological signals for fake detection. However, the detection algorithms developed to filter out the manipulated videos are itself observed to be vulnerable against different alterations [3, 4, 27, 31]. This demands the need for the development of robust fake detection algorithms. A detailed survey on deepfakes is given in [47, 66].

17.2.3 Publicly Available Databases

Researchers have proposed multiple facial retouching and deepfake databases to encourage research toward detection of altered images. The following discusses the details of the databases proposed in the literature for retouching and deepfake detection.

Facial Retouching Databases

Bharati et al. [9] have prepared one of the largest database, the ND-IIITD database, covering seven presets of retouching. Different preset variations are applied using professional software, namely, Portraitpro Studio Max [50]. Retouching is applied to important facial landmark regions such as eyes, lips, nose, and skin texture. Also, relevant retouching operations are applied based on the gender of a person. For example, in preset-1, some of the characteristics of retouching applied to females include skin blush, smooth lips, eyes blue, and nose shorten. For males, the characteristics of retouching include pulp lips, nose slim, shorten wrinkles, and forehead-sculpt. The database contains original images of 325 identities of UND-B [24], on top of that, seven presets are applied for the generation of a variety of retouched face images. In total, the database contains 2600 original and 2275 retouched face images. The authors also created a Celebrity database by downloading images from the Internet. Images pairs labeled with retouched and non-retouched are used to create the database. The database contains 330 images belonging to 165 celebrities. Later, Bharati et al. [10] developed a demography based retouched face database using two tools, namely, BeautyPlus [8] and Potraitpro Studio Max [50]. The database contains subjects belonging to two gender groups, male and female, and three ethnicities, Indian, Chinese, and Caucasian. In total, the database contains 1200 original and 2400 retouched images. Recently, Rathgeb et al. [52] proposed a retouched face database with 800 retouched and 100 original images. Retouched images are created using five different mobile apps. Table 17.1 summarizes the details of the existing facial retouching databases.

DeepFake Databases

In 2017, Agarwal et al. [5] proposed SWAPPED—Digital Attack Video Face database. The database is prepared using Snapchat that swaps/stitches two faces to create fake videos. The database contains 129 real and 612 fake videos of 110

Table 17.1 Details of existing facial retouching databases

Database	Images		Subjects		Retouching Tool
	Real	Retouched	Male	Female	
ND-IIITD [9]	2600	2275	211	114	PortraitPro Studio Max
Celebrity [9]	165	165	25	140	Unknown (Online Sources)
MDRF [10]	1200	2400	300	300	BeautyPlus and Potraitpro Studio Max (v12)
Rathgeb et al. [52]	100	800	50	50	Multiple Mobile Apps

and 31 subjects, respectively. Li et al. [41] proposed the UADFV database with 49 real and 49 fake videos. The database is created using FakeApp mobile application. A large scale database, namely, FaceForensics++ is proposed by Rossler et al. [56]. The database contains 1000 real videos (downloaded from YouTube). Different manipulation techniques are applied to the real videos to generate 4000 fake videos. The database contains four different subsets of manipulated videos that are generated using (i) computer graphics-based techniques and (ii) learning-based techniques. Computer graphics-based techniques include *FaceSwap (FS)* and *Face2Face (F2F)* while learning-based techniques include *DeepFakes (DF)* and *NeuralTextures (NT)*. Each of the manipulation methods requires the source and target videos for the generation of fake/altered videos. FaceSwap utilizes facial landmarks for the generation of a 3D shape model and swaps the facial regions by minimizing the difference between the landmarks in the source and target subject. Post-processing is required to smoothen out the blended regions and for color correction. While FaceSwap blends two faces together, the Face2Face method transfers the expression from the source video to the target video. Therefore, the swapped videos generated using FaceSwap contains the identity of both source and target subjects while the target identity is preserved in Face2Face. DeepFakes is an autoencoder based manipulation technique with a shared encoder that is trained to reconstruct the source and target faces. GAN loss is applied in the NeuralTextures method, and the mouth region is altered. This method relies on tracked geometry for effective manipulation of the expression at the mouth region. Later, a more advanced version of the database is released with more realistic settings of the real-world scenario [28]. By utilizing 363 real videos of 28 paid actors, 3068 deepfake videos are generated. Both the above databases cover the videos in three different qualities: (i) uncompressed (raw), (ii) low compression with quantization factor set to 23 (high quality), and (iii) high compression with quantization factor set to 40 (low quality). Li et al. [43] presented a large scale DeepFake video dataset, termed CelebDF, with high-quality DeepFake videos of celebrities. The fake videos are generated using an advanced version of face swap algorithms. The dataset contains a total of 590 real and 5639 fake videos. Recently, Facebook

Table 17.2 Details of existing deepfake databases

Database	Real		Fake	
	Videos	Source	Videos	Source
SWAPPED [5]	129	Real-time	612	Snapchat
UADFV [41]	49	Youtube	49	FakeApp
FaceForensics++ [56]	1000	Youtube	4000	FaceSwap, Face2Face, NeuralTexture, DeepFake
DeepFake Detection [28]	363	Real-time	3068	DeepFake
Celeb-DF [43]	590	Youtube	5639	DeepFake
DFDC [20]	21154	Actors	102000	DeepFake
WildDeepFake [75]	3805	Online	3509	DeepFake (Online)

has released the Deepfake Detection Challenge (DFDC) [20] database. It is one of the largest databases containing more than 100,000 fake videos of 3426 actors. Zi et al. [75] created the WildDeepfake database by collecting images from the Internet. Table 17.2 summarizes the details of the existing deepfake databases.

17.3 Experimental Evaluation and Observations

In the literature, algorithms proposed for detecting retouched and deepfake images have shown high accuracy when the models are trained on a specific type of alteration and evaluated on similar alterations. For instance, Jain et al. [33] have proposed a convolutional neural network framework for retouching detection by training the framework on retouched and original images. The proposed framework is evaluated on the ND-IIITD database. As reported in Table 17.3, the framework achieved more than 99% accuracy. Similarly, in [56], we observe that existing algorithms perform well when the models are trained on a specific type of manipulation (Table 17.4). Here, the authors used the FaceForensics++ database for detecting manipulated images.

The high performance of deep learning models to detect retouched and altered images (Tables 17.3 and 17.4) in the same domain/manipulation settings indicate that deep models are able to learn distinguishable features when the distribution of the evaluation dataset is similar to the training dataset. In other words, high performance is observed when the training of deep models is done with some apriori knowledge about the type of alterations performed on the images. However, in a real-world scenario, it is not practical to assume such apriori knowledge. Therefore, in this chapter, we highlight the challenges of retouching and alteration detection in a real-world cross train-test alteration detection scenarios (i.e., when trained on one and test on another).

Table 17.3 Classification accuracy (%) for retouching detection on the ND-IIITD database and comparison with existing reported results in literature [33]

Algorithm	Accuracy
Kee and Farid [36]	48.80
Bharati et al. [9] (Unsupervised DBM)	81.90
Bharati et al. [9] (Supervised DBM)	87.10
Jain et al. [33] (Thresholding)—(64, 64, 3)	99.70
Jain et al. [33] (SVM)—(64, 64, 3)	99.42
Jain et al. [33] (Thresholding)—(128, 128, 3)	99.48
Jain et al. [33] (SVM)—(128, 128, 3)	99.65

- **Cross-domain**: Detecting altered images belonging to different domains (retouched and manipulated).
- **Cross manipulation**: Detecting images generated using different types of manipulations.
- **Cross ethnicity**: Detecting altered images belonging to different ethnicities.

Multiple experiments are performed to evaluate the performance of deep models for retouching and alteration detection in the above three experimental settings. Experiments are performed using two state-of-the-art deep models, namely, ResNet50 [30] and XceptionNet [14]. Two popular databases from the literature, namely, ND-IIITD face retouching database and FaceForensics++ database, are used for the experiments. We have also used the IndianForensics database [46] for the cross ethnicity experiment. Figure 17.4 shows some sample images of the databases. Protocols to perform the experiments and the implementation details are discussed below:

Experimental Protocol and Implementation Details: For the experiments, the ND-IIITD database is divided into non-overlapping training and testing sets with 50% subject-wise partitioning corresponding to each retouched and original preset [9]. Training sets of all the presets are combined to create a single training set. Similarly, all the testing sets are merged together into a single testing set. For the FaceForensics++ database, pre-defined protocol is followed for training, validation, and testing partitioning [56]. The IndianForensics database [46] is divided into 50% train-test splits for the experiments. Videos of the FaceForensics++ and IndianForensics databases are divided into frames. 10 frames per video are extracted, and the results are reported using frame based accuracy.

Pre-trained ResNet50 and XceptionNet models are fine-tuned by adding two fully connected dense layers of 512 dimensions after the final convolutional layer. Models are trained using Adam optimizer for 20 epochs with a batch size of 32. For the initial 10 epochs, the learning rate is set to 0.0001 and reduced by 0.1 after every 5 epochs. Frames are extracted from the videos of the FaceForensics++ database and resized to 128×128 resolution. The images of the ND-IIITD database are also resized to 128×128 resolution. All the experiments are performed under TensorFlow 2.0

Table 17.4 Classification accuracy (%) of manipulation-specific forgery detectors on the FaceForensics++ database (from [56])

	No Compression				Compressed 23				Compressed 40			
	DF	F2F	FS	NT	DF	F2F	FS	NT	DF	F2F	FS	NT
Steg. Features + SVM [26]	99.03	99.13	98.27	**99.88**	77.12	74.68	79.51	76.94	65.58	57.55	60.58	60.69
Cozzolino et al. [16]	98.83	98.56	98.89	**99.88**	81.78	85.32	85.69	80.60	68.26	59.38	62.08	62.42
Bayar and Stamm [7]	99.28	98.79	98.98	98.78	90.18	94.93	93.14	86.04	80.95	77.30	76.83	72.38
Rahmouni et al. [51]	98.03	98.96	98.94	96.06	82.16	93.48	92.51	75.18	73.25	62.33	67.08	62.59
MesoNet [2]	98.41	97.96	96.07	97.05	95.26	95.84	93.43	85.96	89.52	84.44	83.56	75.74
XceptionNet [14]	**99.59**	**99.61**	**99.14**	99.36	**98.85**	**98.36**	**98.23**	**94.50**	**94.28**	**91.56**	**93.70**	**82.11**

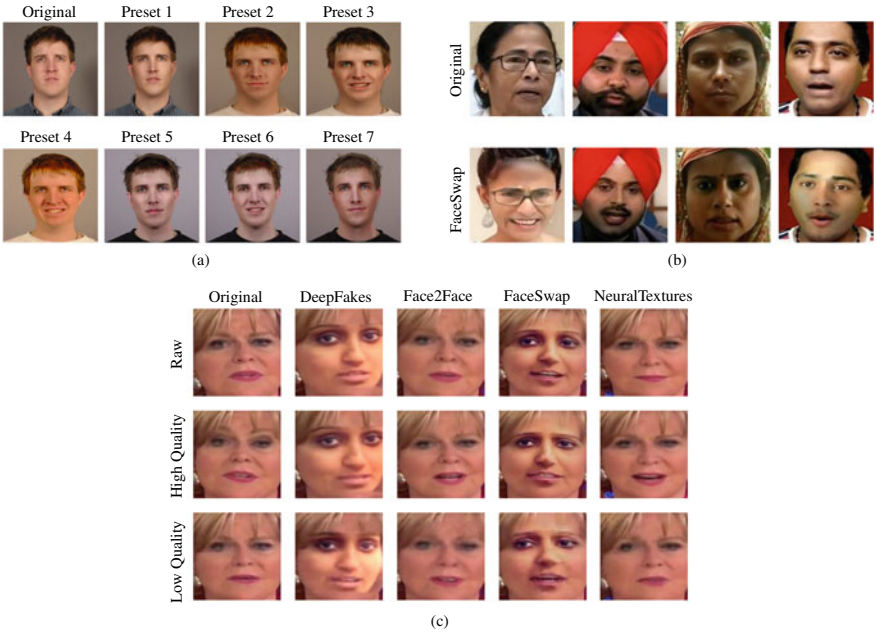

Fig. 17.4 Sample images of the **a** ND-IIITD [9] **b** IndianForensics [46], and **c** FaceForensics++ [56] databases

environment on a DGX station with Intel Xeon CPU, 256 GB RAM, and four 32 GB Nvidia V100 GPU cards.

17.3.1 Cross-Domain Alteration Detection

The aim of these experiments is to evaluate the generalizability of deep models to detect altered images across different domains. In these experiments, models trained on the ND-IIITD database are separately evaluated on the four face manipulation subsets of the FaceForensics++ database (Deepfakes, Face2Face, FaceSwap, and NeuralTextures), and vice versa. Experiments are performed on the uncompressed subsets of the FaceForensics++ database to maintain uniformity with respect to the compression factor of the images in both the databases. Compression introduces artifacts that pose additional challenges to the detection algorithms. Therefore, to solely analyze the challenges due to unseen alterations across different domains, the compression factor of the images is kept consistent during the experiments.

Table 17.5 shows the classification accuracy of deep models trained on different manipulation types of the FaceForensics++ database and evaluated on the ND-IIITD

Table 17.5 Classification accuracy (%) of the models trained on the FaceForensics++ database and evaluated on the ND-IIITD database

	DF	F2F	FS	NT
ResNet50	49.95	49.95	48.59	50.21
XceptionNet	56.22	49.86	46.23	52.89

Table 17.6 Classification accuracy (%) of the model trained on the ND-IIITD database and evaluated on different manipulation types of the FaceForensics++ database

	DF	F2F	FS	NT
ResNet50	50.43	50.18	50.11	50.54
XceptionNet	53.50	52.00	48.68	53.11

database. It is observed that the models do not perform well and yield almost random accuracy for retouching detection. Models trained on FaceSwap achieve the lowest accuracy of 48.59% and 46.23%, with ResNet50 and XceptionNet, respectively. The classification accuracy of the model trained on the ND-IIITD database and evaluated separately on different subsets of the FaceForensics++ database is shown in Table 17.6. Similar to the previous scenario, it is observed that deep models do not perform well in cross-domain settings. The degradation in performance is due to the effect of the domain shift from the training set to the evaluation set.

17.3.2 Cross Manipulation Alteration Detection

To observe the performance of deep models for unseen manipulation detection, experiments are performed on the FaceForensics++ database. This experiment is performed to analyze the robustness of deep models by training them on a specific manipulation type and evaluating on others. We have used four subsets of manipulated videos (with different quality levels) of the FaceForensics++ database for the experiments. Training and evaluation of the models are performed on a fixed quality level. For example, models trained on the uncompressed videos of a specific manipulation type are evaluated on the uncompressed videos of other manipulation types.

Table 17.7 shows the classification performance of deep models for unseen manipulation detection. It is observed that most of the models do not perform well in cross manipulation detection settings. Interestingly, there is minimal effect of compression observed on the performance of deep models. Rather in some cases, it is observed that the performance of deep models on the compressed videos is better than uncompressed videos. For instance, models trained on FaceSwap (FS) when evaluated on DeepFakes (DF) achieves 58.57% and 61.89% accuracy using ResNet50 and XceptionNet, respectively, on high compressed videos (compressed 40), while these models achieve 50.82% and 51.00% accuracy on uncompressed videos. It is our assertion

Table 17.7 Classification accuracy (%) of the models trained on a specific type of manipulation and evaluated on others of the FaceForensics++ database

Trained on		No Compression			Compressed 23			Compressed 40		
		F2F	**FS**	**NT**	**F2F**	**FS**	**NT**	**F2F**	**FS**	**NT**
DF	ResNet50	53.54	49.46	58.57	51.18	50.04	51.71	54.04	53.75	52.86
	XceptionNet	56.36	49.64	63.57	51.86	49.93	54.07	53.36	55.64	52.89
		DF	**FS**	**NT**	**DF**	**FS**	**NT**	**DF**	**FS**	**NT**
F2F	ResNet50	58.11	50.75	51.57	55.79	52.18	50.79	58.86	53.32	54.25
	XceptionNet	63.11	51.18	51.61	59.96	51.68	52.86	58.96	52.96	54.25
		DF	**F2F**	**NT**	**DF**	**F2F**	**NT**	**DF**	**F2F**	**NT**
FS	ResNet50	50.82	53.07	50.00	51.04	52.32	50.18	58.57	51.36	50.29
	XceptionNet	51.00	52.39	49.93	52.36	53.61	48.96	61.89	51.64	51.00
		DF	**F2F**	**FS**	**DF**	**F2F**	**FS**	**DF**	**F2F**	**FS**
NT	ResNet50	86.89	56.68	49.39	74.43	56.39	48.04	61.29	60.11	52.32
	XceptionNet	91.32	67.00	49.75	76.68	58.57	48.61	61.89	61.50	50.79

that instead of learning the discriminative features to distinguish manipulated videos from original ones, the models are learning the compression artifacts in compressed videos for discrimination. Therefore better performance is achieved for low-quality videos. It is also important to observe that the models trained on NeuralTextures (NT) achieves high accuracy when evaluated on DeepFakes (DF), while the opposite is not true. This raises several questions about the kind of information learned by deep models for discrimination. All these observations open new research threads toward developing sophisticated algorithms for unseen manipulation detection. It further emphasizes the importance of the interpretability of deep models for a better understanding of the obtained results.

17.3.3 Cross Ethnicity Alteration Detection

To observe the fairness of detection algorithms, experiments are performed on the FaceForensics++ and IndianForensics databases, to analyze the performance of deep models in cross ethnicity settings. The IndianForensics database contains 200 original and 234 fake videos of Indian people. Fake videos are created by face swapping using FSGAN [49]. Experiments are performed by training the models on the IndianForensics database and evaluating on FaceSwap manipulated videos of the FaceForensics++ database and vice versa. The aim of these experiments is to evaluate the performance of detection algorithms across Indian and non-Indian ethnicities. Figure 17.5 shows the classification accuracy for the same. ResNet50 and XceptionNet models trained on the IndianForensics database yield an accuracy of 39.29% and 39.79%, respectively, on detecting FaceSwap manipulated videos of the FaceForensics++ database. On the other hand, models trained on the FaceForensics++

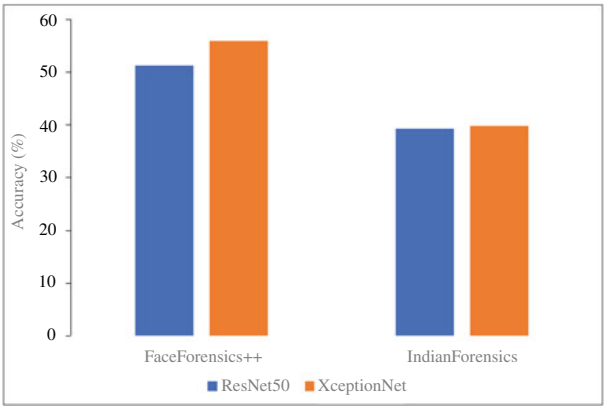

Fig. 17.5 Classification accuracy (%) of the models trained on the IndianForensics database and evaluated on the FaceForensics++ database and vice versa

database yields an accuracy of 51.35% and 55.95% on the IndianForensics database corresponding to ResNet50 and XceptionNet, respectively. The low detection accuracy indicates the effect of ethnicity on the performance of detection algorithms. A similar effect of ethnicity on the alteration detection algorithms has been recently shown by Mehra et al. [46].

17.4 Open Challenges

To develop robust alteration detection algorithms/systems which can be deployed in the real world, we believe that the challenges discussed below require the attention of the research community.

Generalizability of Detection Algorithms Across Different Domains: Retouching and deepfakes are different types of facial alterations that belong to different domains of adversaries (unintentional and intentional). In the literature, various algorithms/deep models have been proposed for their detection, and high performance is achieved by training them separately, either for the task of retouching detection or deepfakes detection. However, as mentioned in the previous section, in a real-world scenario, the apriori knowledge of the type of alteration is not available. It is possible that the images in the evaluation dataset are altered using some other image editing tools and techniques which are not seen during the training process. The experiments performed to evaluate the generalizability of deep models for cross-domain alteration detection indicate that deep models do not perform well for detecting altered images belonging to different domains of adversaries. Therefore, it is important to develop generalizable algorithms that could handle the effect of domain shift between different types of alterations.

Robustness of Detection Algorithms Across Different Types of Manipulations: Manipulations are performed using different computer vision-based techniques, learning-based techniques, and using simple mobile applications. Due to the ease of creating manipulated images/videos, social media platforms are now flooded with altered content. With the advancement of technology, different types of manipulated images are created on a daily basis and shared through social media platforms. It is therefore important that the detection algorithms deployed on these platforms must detect the altered images generated using new techniques. In a real-world scenario, it is impractical to regularly update the deployed models with new types of manipulated images/videos. Thus, the detection algorithms/models should be robust to unseen manipulations as well.

Effect of Ethnicity on Detection Algorithms: Fairness in model predictions with respect to different demographic groups or protected attributes (such as gender and race) is important for the trustability and dependability of deep learning algorithms [23, 42]. Therefore, in a real-world scenario, the detection algorithms must be fair across different demographic groups. In other words, the performance of detection algorithms/deep models should be equal across different demographic groups. Experiments performed to detect altered images in cross ethnicity settings indicate that the performance of deep models degrades significantly when the altered images belong to different ethnicities. This highlights the need for sophisticated detection algorithms to overcome the challenges of cross ethnicity effect.

17.5 Conclusion

Face image alterations have a very diverse usage, ranging from beautification, to getting unauthorized access, to even spreading fake news. Based on the intent, alterations can be broadly classified into two categories: unintentional manipulations which include makeup and retouching/beautification, and intentional manipulations which includes deepfakes. Both these alterations significantly degrade the performance of face recognition algorithms and have several adverse effects when used with malicious intent. In this chapter, as the first contribution, we have provided a comprehensive survey of the literature toward these manipulations. For both the alterations, a summary of the relevant databases and detection techniques is provided. The survey can help the research community to progress in the field of altered image detection and to develop secure face recognition algorithms/systems.

The second contribution of this chapter aims at highlighting the open challenges in facial alteration detection. In the literature, the detection algorithms are generally evaluated by training and testing under the same domain (for instance, same alteration type), and the algorithms have shown high detection accuracy. In this chapter, we showcase more diverse usage of the algorithms and performed several experiments to evaluate the performance of two state-of-the-art deep convolutional network models under those challenging unseen alteration detection settings. It is found that the models that reported high accuracy for seen alteration settings failed miserably under

unseen alteration settings. We assert that the challenges discussed in this chapter and the experimental results will help the research community in building robust and generalizable detection algorithms deployable in the real world.

Acknowledgements R. Singh and M. Vatsa are partially supported through a research grants from MeitY, and MHA, Government of India. P. Majumdar is partly supported by DST INSPIRE Ph.D. Fellowship. M. Vatsa is also partially supported through Swarnajayanti Fellowship by the Government of India.

References

1. 68 percent of adults edit their selfies before sharing them with anyone. https://fstoppers.com/mobile/68-percent-adults-edit-their-selfies-sharing-them-anyone-95417. Accessed 29 January 2021
2. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: IEEE international workshop on information forensics and security (WIFS), pp 1–7
3. Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H (2019) Protecting world leaders against deep fakes. In: CVPR workshops, pp 38–45
4. Agarwal A, Sehwag A, Vatsa M, Singh R (2019) Deceiving the protector: fooling face presentation attack detection algorithms. In: IEEE international conference on biometrics (ICB), pp 1–6
5. Agarwal A, Singh R, Vatsa M, Noore A (2017) Swapped! digital face presentation attack detection via weighted local magnitude pattern. In: IEEE international joint conference on biometrics (IJCB), pp 659–665
6. Arab MA, Azadi Moghadam P, Hussein M, Abd-Almageed W, Hefeeda M (2020) Revealing true identity: detecting makeup attacks in face-based biometric systems. In: ACM international conference on multimedia, pp 3568–3576
7. Bayar B, Stamm MC (2016) A deep learning approach to universal image manipulation detection using a new convolutional layer. In: ACM workshop on information hiding and multimedia security, pp 5–10
8. Beautyplus. https://www.beautyplus.com/. Accessed 29 January 2021
9. Bharati A, Singh R, Vatsa MBowyer KW (2016) Detecting facial retouching using supervised deep learning. IEEE Trans Inform Forensics Secur 11(9):1903–1913
10. Bharati A, Vatsa M, Singh R, Bowyer KW, Tong X (2017) Demography-based facial retouching detection using subclass supervised sparse autoencoder. In: 2017 IEEE international joint conference on biometrics (IJCB), pp 474–482
11. Cao C, Lu F, Li C, Lin S, Shen X (2019) Makeup removal via bidirectional tunable de-makeup network. IEEE Trans Multimedia 21(11):2750–2761
12. Chen C, Dantcheva A, Ross A (2013) Automatic facial makeup detection with application in face recognition. In: 2013 international conference on biometrics (ICB), pp 1–8
13. Choi D, Hwang C (2011) Image morphing using mass-spring system. In: International conference on computer graphics and virtual reality, pp 156–159
14. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: IEEE conference on computer vision and pattern recognition, pp 1251–1258
15. Ciftci UA, Demir I, Yin L (2020) Fakecatcher: detection of synthetic portrait videos using biological signals. IEEE transactions on pattern analysis and machine intelligence
16. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: ACM workshop on information hiding and multimedia security, pp 159–164

17. Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: IEEE/CVF conference on computer vision and pattern recognition, pp 5781–5790
18. Dantcheva A, Chen C, Ross A (2012) Can facial cosmetics affect the matching accuracy of face recognition systems? In: IEEE fifth international conference on biometrics: theory, applications and systems (BTAS), pp 391–398
19. Deepfake Detection Challenge. https://deepfakedetectionchallenge.ai
20. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC (2020) The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397
21. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp 214–226
22. Ferrara M, Franco A, Maltoni D (2016) On the effects of image alterations on face recognition accuracy. In: Face recognition across the imaging spectrum, pp 195–222. Springer
23. Ferrara M, Franco A, Maltoni D, Sun Y (2013) On the impact of alterations on face photo recognition accuracy. In: International conference on image analysis and processing, pp 743–751
24. Flynn PJ, Bowyer KW, Phillips PJ (2003) Assessment of time dependency in face recognition: An initial study. In: International conference on audio-and video-based biometric person authentication, pp 44–51
25. French law on photoshopped images. https://www.huffpost.com/entry/france-photoshop-models-law$_$n$_$59d0dcc6e4b05f005d34c309. Accessed 29 January 2021
26. Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inform Forensics Secur 7(3):868–882
27. Gandhi A, Jain S (2020) Adversarial perturbations fool deepfake detectors. In: IEEE international joint conference on neural networks (IJCNN), pp 1–8
28. Google AI, contributing data to deepfake detection research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed 29 January 2021
29. Guo G, Wen L, Yan S (2013) Face authentication with makeup changes. IEEE Trans Circuit Syst Video Technol 24(5):814–825
30. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778
31. Hussain S, Neekhara P, Jere M, Koushanfar F, McAuley J (2021) Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In: IEEE/CVF winter conference on applications of computer vision, pp 3348–3357
32. Jain A, Majumdar P, Singh R, Vatsa M (2020) Detecting GANs and retouching based digital alterations via DAD-HCNN. In: IEEE/CVF conference on computer vision and pattern recognition workshops, pp 672–673
33. Jain A, Singh R, Vatsa M (2018) On detecting GANs and retouching based synthetic alterations. In: 2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS), pp 1–7
34. Jassim S, Asaad A (2018) Automatic detection of image morphing by topology-based analysis. In: IEEE European signal processing conference (EUSIPCO), pp 1007–1011
35. Jung T, Kim S, Kim K (2020) Deepvision: Deepfakes detection using human eye blinking pattern. IEEE Access 8:83144–83154
36. Kee E, Farid H (2011) A perceptual metric for photo retouching. Natl Acad Sci 108(50):19907–19912
37. Klontz JC, Klare BF, Klum S, Jain AK, Burge MJ (2013) Open source biometric recognition. In: IEEE sixth international conference on biometrics: theory, applications and systems, pp 1–8
38. Kose N, Apvrille L, Dugelay JL (2015) Facial makeup detection technique based on texture and shape analysis. In: IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 1, pp 1–7
39. Kotwal K, Mostaani Z, Marcel S (2019) Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features. IEEE Trans Biometr Behav Identity Sci 2(1):15–25

40. Kumar P, Vatsa M, Singh R (2020) Detecting face2face facial reenactment in videos. In: IEEE/CVF winter conference on applications of computer vision, p. 2589–2597
41. Li Y, Chang MC, Lyu S (2018) In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. In: IEEE international workshop on information forensics and security
42. Liu KH, Liu TJ, Liu HH, Pei SC (2015) Facial makeup detection via selected gradient orientation of entropy information. In: IEEE international conference on image processing (ICIP), pp 4067–4071
43. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-df: a large-scale challenging dataset for deepfake forensics. In: IEEE/CVF conference on computer vision and pattern recognition, pp 3207–3216
44. Majumdar P, Agarwal A, Singh R, Vatsa M (2019) Evading face recognition via partial tampering of faces. In: IEEE/CVF conference on computer vision and pattern recognition workshops, pp 11–20
45. Majumdar P, Chhabra S, Singh R, Vatsa M (2020) Subgroup invariant perturbation for unbiased pre-trained model prediction. Frontiers Big Data 3:52
46. Mehra A, Agarwal A, Vatsa M, Singh R (2021) Detection of digital manipulation in facial images (student abstract). In: AAAI conference on artificial intelligence
47. Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. ACM Comput Surv (CSUR) 54(1):1–41
48. New Israeli law bans use of too-skinny models in ads. https://cnn.it/1mNTiY1. Accessed: 9 February 2021
49. Nirkin Y, Keller Y, Hassner T (2019) FSGAN: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7184–7193
50. Portrait pro21. https://www.anthropics.com/portraitpro/. Accessed 29 January 2021
51. Rahmouni N, Nozick V, Yamagishi J, Echizen I (2017) Distinguishing computer graphics from natural images using convolution neural networks. In: IEEE workshop on information forensics and security, pp 1–6
52. Rathgeb C, Botaljov A, Stockhardt F, Isadskiy S, Debiasi L, Uhl A, Busch C (2020) Prnu-based detection of facial retouching. IET Biometrics 9(4):154–164
53. Rathgeb C, Dantcheva A, Busch C (2019) Impact and detection of facial beautification in face recognition: an overview. IEEE Access 7:152667–152678
54. Rathgeb C, Satnoianu CI, Haryanto N, Bernardo K, Busch C (2020) Differential detection of facial retouching: a multi-biometric approach. IEEE Access 8:106373–106385
55. Reddit bans deepfake porn videos. http://www.bbc.com/news/technology-42984127. Accessed 9 February 2021
56. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to detect manipulated facial images. In: IEEE/CVF international conference on computer vision, pp 1–11
57. Russello S (2009) The impact of media exposure on self-esteem and body satisfaction in men and women. J Interdisciplinary Undergrad Res 1(1):4
58. Sadu C, Das PK (2020) Swapping face images based on augmented facial landmarks and its detection. In: IEEE region 10 conference (TENCON), pp 456–461
59. Scherhag U, Budhrani D, Gomez-Barrero M, Busch C (2018) Detecting morphed face images using facial landmarks. In: International conference on image and signal processing, pp 444–452
60. Singh R, Agarwal A, Singh M, Nagpal S, Vatsa M (2020) On the robustness of face recognition algorithms against attacks and bias. In: AAAI conference on artificial intelligence, vol 34, pp 13583–13589
61. Spreeuwers L, Schils M, Veldhuis R (2018) Towards robust evaluation of face morphing detection. In: IEEE european signal processing conference (EUSIPCO), pp 1027–1031
62. Sun Y, Ren L, Wei Z, Liu B, Zhai Y, Liu S (2017) A weakly supervised method for makeup-invariant face verification. Pattern Recogn 66:153–159
63. Supermodels without photoshop: Israel photoshop law. https://www.ibtimes.com/supermodels-without-photoshop-israels-photoshop-law-puts-focus-digitally-altered-images-photos. Accessed 29 January 2021

64. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing obama: Learning lip sync from audio. ACM Trans Graph 36(4):95:1–95:13
65. The self esteem act. https://www.dailymail.co.uk/femail/article-2048375/Self-Esteem-Act-US-parents-push-anti-Photoshop-laws-advertising.html. Accessed 29 January 2021
66. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inform Fusion 64:131–148
67. United states beauty and personal care products. https://www.mordorintelligence.com/industry-reports/united-states-cosmetics-products-market-industry. Accessed 29 January 2021
68. Wang SY, Wang O, Owens A, Zhang R, Efros AA (2019) Detecting photoshopped faces by scripting photoshop. In: IEEE/CVF international conference on computer vision, pp 10072–10081
69. Wang S, Fu Y (2016) Face behind makeup. In: AAAI conference on artificial intelligence, vol 30
70. Watch a man manipulate George Bush face in real time. https://bit.ly/2wVgNN4. Accessed 9 February 2021
71. Wu J (2011) Face recognition jammer using image morphing. Boston Univ., USA, Tech. Rep. ECE-2011
72. Wu X, He R, Sun Z, Tan T (2018) A light cnn for deep face representation with noisy labels. IEEE Trans Inform Forensics Secur 13(11):2884–2896
73. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 8261–8265
74. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision, pp 2223–2232
75. Zi B, Chang M, Chen J, Ma X, Jiang YG (2020) Wilddeepfake: a challenging real-world dataset for deepfake detection. In: 28th ACM international conference on multimedia, pp 2382–2390