# DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales and Javier Ortega-Garcia
Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid, Spain
{ruben.tolosana, ruben.vera, julian.fierrez, aythami.morales, javier.ortega}@uam.es



***Abstract*—The free access to large-scale public databases, together with the fast progress of deep learning techniques, in particular Generative Adversarial Networks, have led to the generation of very realistic fake content with its corresponding implications towards society in this era of fake news.**

**This survey provides a thorough review of techniques for manipulating face images including DeepFake methods, and methods to detect such manipulations. In particular, four types of facial manipulation are reviewed: *i)* entire face synthesis, *ii)* identity swap (DeepFakes), *iii)* attribute manipulation, and *iv)* expression swap. For each manipulation group, we provide details regarding manipulation techniques, existing public databases, and key benchmarks for technology evaluation of fake detection methods, including a summary of results from those evaluations. Among all the aspects discussed in the survey, we pay special attention to the latest generation of DeepFakes, highlighting its improvements and challenges for fake detection.**

**In addition to the survey information, we also discuss open issues and future trends that should be considered to advance in the field.**

***Index Terms*—Fake News, DeepFakes, Media Forensics, Face Manipulation, Face Recognition, Biometrics, Databases, Benchmark**

## I. Introduction

FAKE images and videos including facial information generated by digital manipulation, in particular with DeepFake methods [1], have become a great public concern recently [2], [3]. The very popular term "DeepFake" is referred to a deep learning based technique able to create fake videos by swapping the face of a person by the face of another person. This term was originated after a Reddit user named "deepfakes" claimed in late 2017 to have developed a machine learning algorithm that helped him to transpose celebrity faces into porn videos [4]. In addition to fake pornography, some of the more harmful usages of such fake content include fake news, hoaxes, and financial fraud. As a result, the area of research traditionally dedicated to general media forensics [5]–[11], is being invigorated and is now dedicating growing efforts for detecting facial manipulation in image and video [12]. Part of these renewed efforts in fake face detection are built around past research in biometric anti-spoofing [13]–[15] and modern data-driven deep learning [16], [17]. The growing interest in fake face detection is demonstrated through the increasing number of workshops in top conferences [18]–[22], international projects such as MediFor funded by the Defense Advanced Research Project Agency (DARPA), and competitions such as the recent Media Forensics Challenge (MFC2018)[1] and the Deepfake Detection Challenge (DFDC)[2] launched by the National Institute of Standards and Technology (NIST) and Facebook, respectively.

Traditionally, the number and realism of facial manipulations have been limited by the lack of sophisticated editing tools, the domain expertise required, and the complex and time-consuming process involved. For example, an early work in this topic [23] was able to modify the lip motion of a person speaking using a different audio track, by making connections between the sounds of the audio track and the shape of the subject's face. However, from these early works up to date, many things have rapidly evolved in the last years. Nowadays, it is becoming increasingly easy to automatically synthesise non-existent faces or manipulate a real face of one person in an image/video, thanks to: *i)* the accessibility to large-scale public data, and *ii)* the evolution of deep learning techniques that eliminate many manual editing steps such as Autoencoders (AE) and Generative Adversarial Networks (GAN) [24], [25]. As a result, open software and mobile application such as ZAO[3] and FaceApp[4] have been released opening the door to anyone to create fake images and videos, without any experience in the field needed.

In response to those increasingly sophisticated and realistic manipulated content, large efforts are being carried out by the research community to design improved methods for face manipulation detection. Traditional fake detection methods in media forensics have been commonly based on: *i)* in-camera fingerprints, the analysis of the intrinsic fingerprints introduced by the camera device, both hardware and software, such as the optical lens [27], colour filter array and interpolation [28], [29], and compression [30], [31], among others, and *ii)* out-camera fingerprints, the analysis of the external fingerprints introduced by editing software, such as copy-paste or copy-move different elements of the image [32], [33], reduce the frame rate in a video [34], [35], etc. However, most of the features considered in traditional fake detection methods are highly dependent on the specific training scenario, being therefore not robust against unseen conditions [6], [8], [16]. This is of special importance in the era we live in as most media fake content is usually shared on social networks, whose platforms automatically modify the original image/video, for example, through compression and resize operations [12].

[1] https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018
[2] https://deepfakedetectionchallenge.ai/
[3] https://apps.apple.com/cn/app/id1465199127
[4] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341

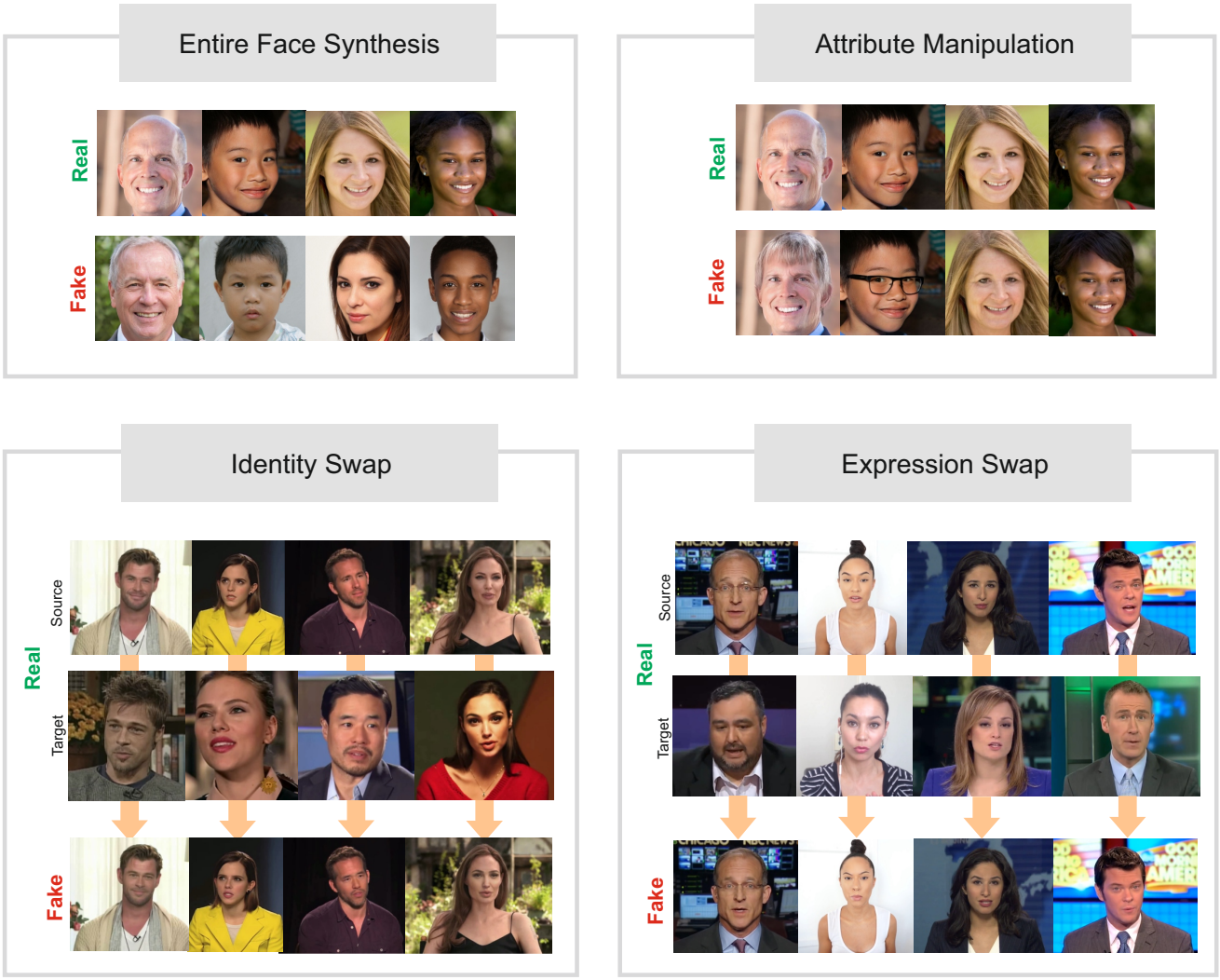


Fig. 1. Real and fake examples of each facial manipulation group. For *Entire Face Synthesis*, real images are extracted from http://www.whichfaceisreal.com/ and fake images from https://thispersondoesnotexist.com. For *Identity Swap*, face images are extracted from Celeb-DF database [26]. For *Attribute Manipulation*, real images are extracted from http://www.whichfaceisreal.com/ and fake images are generated using FaceApp. Finally, for *Expression Swap*, images are extracted from FaceForensics++ [12].

This survey provides an in-depth review of digital manipulation techniques applied to facial content due to the large number of possible harmful applications, e.g., the generation of fake news that would provide misinformation in political elections and security threats [36], [37]. Specifically, we cover four types of manipulations: *i)* entire face synthesis, *ii)* identity swap, *iii)* attribute manipulation, and *iv)* expression swap. These four main types of face manipulation are well established by the research community, receiving most attention in the last few years. Besides, we also review in this survey some other challenging and dangerous face manipulation techniques that are not so popular yet like face morphing.

Finally, for completeness, we would like to highlight other recent surveys in the field. In [38], the authors cover the topic of DeepFakes from a general perspective, proposing the R.E.A.L framework to manage DeepFake risks. In addition, Verdoliva has recently surveyed in [39] traditional manipulation and fake detection approaches considered in general media forensics, and also the latest deep learning techniques. The present survey complements [38] and [39] with a more detailed review of each facial manipulation group, including manipulation techniques, existing public databases, and key benchmarks for technology evaluation of fake detection methods, including a summary of results from those evaluations. In addition, we pay special attention to the latest generation of DeepFakes, highlighting its improvements and challenges for fake detection.

The remainder of the article is organised as follows. We first provide in Sec. II a general description of different types of facial manipulation. Then, from Sec. III to Sec. VI we describe the key aspects of each type of facial manipulation including public databases for research, detection methods, and benchmark results. Sec. VII focuses on other interesting types of face manipulation techniques not covered in previous sections. Finally, we provide in Sec. VIII our concluding remarks, highlighting open issues and future trends.

## II. Types of Facial Manipulations

Facial manipulations can be categorised in four main different groups regarding the level of manipulation. Fig. 1 graphically summarises each facial manipulation group. A description of each of them is provided below, from higher to lower level of manipulation:

- **Entire Face Synthesis:** this manipulation creates entire non-existent face images, usually through powerful GAN, e.g., through the recent StyleGAN approach proposed in [40]. These techniques achieve astonishing results, generating high-quality facial images with a high level of realism. Fig. 1 shows some examples for entire face synthesis generated using StyleGAN[5]. This manipulation could benefit many different sectors such as the video game and 3D-modelling industries, but it could also be used for harmful applications such as the creation of very realistic fake profiles in social networks in order to generate misinformation.
- **Identity Swap:** this manipulation consists of replacing the face of one person in a video with the face of another person. Two different approaches are usually considered: *i)* classical computer graphics-based techniques such as FaceSwap[6], and *ii)* novel deep learning techniques known as DeepFakes[7], e.g., the recent ZAO mobile application. Very realistic videos of this type of manipulation can be seen on Youtube[8]. This type of manipulation could benefit many different sectors, in particular the film industry. However, in the other side, it could also be used for bad purposes such as the creation of celebrity pornographic videos, hoaxes, and financial fraud, among many others.
- **Attribute Manipulation:** this manipulation, also known as face editing or face retouching, consists of modifying some attributes of the face such as the colour of the hair or the skin, the gender, the age, adding glasses, etc [41]. This manipulation process is usually carried out through GAN such as the StarGAN approach proposed in [42]. One example of this type of manipulation is the popular FaceApp mobile application. Consumers could use this technology to try on a broad range of products such as cosmetics and makeup, glasses, or hairstyles in a virtual environment.
- **Expression Swap:** this manipulation, also known as face reenactment, consists of modifying the facial expression of the person. Although different manipulation techniques are proposed in the literature, e.g., at image level through popular GAN architectures [43], in this group we focus on the most popular techniques Face2Face and NeuralTextures [44], [45], which replaces the facial expression of one person in a video with the facial expression of another person. This type of manipulation could be used with serious consequences, e.g., the popular video of Mark Zuckerberg saying things he never said[9].

[5]https://thispersondoesnotexist.com
[6]https://github.com/MarekKowalski/FaceSwap
[7]https://github.com/deepfakes/faceswap
[8]https://www.youtube.com/watch?v=UlvoEW7l5rs
[9]https://www.bbc.com/news/technology-48607673

TABLE I
**Entire Face Synthesis:** Publicly available databases.

| Database | Real Images | Fake Images |
|---|---|---|
| 100K-Generated-Images (2019) [40] | - | 100,000 (StyleGAN) |
| 100K-Faces (2019) [46] | - | 100,000 (StyleGAN) |
| DFFD (2019) [17] | - | 100,000 (StyleGAN)<br>200,000 (ProGAN) |
| iFakeFaceDB (2019) [16] | - | 250,000 (StyleGAN)<br>80,000 (ProGAN) |

## III. Entire Face Synthesis

### A. Manipulation Techniques and Public Databases

This manipulation creates entire non-existent face images. Table I summarises the main publicly available databases for research focused on the entire face synthesis. Four different databases of fake images are of relevance here, all of them based on the same GAN architectures: ProGAN [47] and StyleGAN [40]. It is interesting to remark that each fake image may be characterised by a specific GAN fingerprint just like natural images are identified by a device-based fingerprint (i.e., PRNU). In fact, these fingerprints seem to be dependent not only of the GAN architecture, but also of the different instances of it [48], [49].

In addition, as indicated in Table I, it is important to note that the four mentioned databases only contain fake images generated using the GAN architectures discussed. In order to perform fake detection experiments on this manipulation group, researchers need to obtain real face images from other public databases such as CelebA [50], FFHQ [40], CASIA-WebFace [51], and VGGFace2 [52], among others.

We provide next a description of each public database. In [40], Karras *et al.* released a set of 100,000 synthetic face images, named 100K-Generated-Images[10]. This database was generated using their proposed StyleGAN architecture, which was trained using the FFHQ dataset [40]. StyleGAN is an improved version of their previous popular approach ProGAN, 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 learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis.

Another public database is 100K-Faces [46]. This database contains 100,000 synthetic images generated using StyleGAN. 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.

[10]https://github.com/NVlabs/stylegan

(a) Fake (b) Fake after GANprintR

Fig. 2. Examples of a fake image created using StyleGAN and its improved version after removing the GAN-fingerprint information with GANprintR [16].

Recently, Dang *et al.* introduced in [17] a new database named Diverse Fake Face Dataset (DFFD). Regarding the entire face synthesis manipulation, the authors created 100,000 and 200,000 fake images through the pre-trained ProGAN and StyleGAN models, respectively.

Finally, Neves *et al.* presented in [16] the iFakeFaceDB database. This database comprises 250,000 and 80,000 synthetic face images created with StyleGAN and ProGAN, respectively. As an additional feature in comparison to previous databases, and in order to hinder fake detectors, in this database the fingerprints produced by the GAN architectures were removed through an approach named GANprintR (GAN fingerprint Removal), while keeping very realistic appearance. Fig. 2 shows an example of a fake image directly generated with StyleGAN and its improved version after removing the GAN-fingerprint information. As a result of the GANprintR step, iFakeFaceDB presents a higher challenge for advanced fake detectors compared with the other databases.

### *B. Manipulation Detection*

Different studies have recently evaluated the difficulty of detecting whether faces are real of artificially generated. Table II shows a comparison of the most relevant approaches in this area. For each study, we include information related to the method, classifiers, best performance, and databases considered. We highlight in **bold** the best results achieved for each public database. It is important to remark that in some cases, different evaluation metrics are considered, e.g., Area Under the Curve (AUC) or Equal Error Rate (EER), which complicates the comparison among the studies.

Some authors propose to analyse the internal GAN pipeline in order to detect different artifacts between real and fake images. In [53], the authors hypothesised that the colour is markedly different between real camera images and fake synthesis images. They proposed a detection system based on colour features and a linear Support Vector Machine (SVM) for the final classification, achieving a final 70.0% AUC for the best performance when evaluating with the NIST MFC2018 dataset [58].

Another interesting approach in this line was proposed in [54]. Wang *et al.* conjectured that monitoring neuron behavior could also serve as an asset in detecting fake faces since layer-by-layer neuron activation patterns may capture more subtle features that are important for the facial manipulation detection system. Their proposed approach, named FakeSpoter, extracted as features neuron coverage behaviors of real and fake faces from deep face recognition systems (i.e., VGG-Face [59], OpenFace [60], and FaceNet [61]), and then trained a SVM for the final classification. The authors tested their proposed approach using real faces from CelebA-HQ [47] and FFHQ [40] databases and synthetic faces created through InterFaceGAN [62] and StyleGAN [40], achieving for the best performance a final 84.7% fake detection accuracy using the FaceNet model.

Fake detection systems inspired in steganalysis have also been studied. Nataraj *et al.* proposed in [55] a detection system based on a combination of pixel co-occurrence matrices and Convolutional Neural Networks (CNN). Their proposed approach was initially tested through a database of various objects and scenes created through CycleGAN [63]. Besides, the authors performed an interesting analysis to see the robustness of the proposed approach against fake images created through different GAN architectures (CycleGAN vs. StarGAN), with good generalisation results. This detection approach was implemented later on in [16] considering images from the 100K-Faces database, achieving an EER of 12.3% for the best fake detection performance. This result is remarked in *italics* in Table II to indicate that it was not provided in the original paper.

Many studies have also focused on the detection of the special fingerprints inserted by GAN architectures using pure deep learning methods. Yu *et al.* proposed in [56] an attribution network architecture to map an input image to its corresponding fingerprint image. Therefore, they learned a model fingerprint for each source (each GAN instance plus the real world), such that the correlation index between one image fingerprint and each model fingerprint serves as softmax logit for classification. Their proposed approach was tested using real faces from CelebA database [50] and synthetic faces created through different GAN approaches (ProGAN [47], SNGAN [64], CramerGAN [65], and MMDGAN [66]), achieving a final 99.5% fake detection accuracy for the best performance. However, this approach seemed not to be very robust against unseen simple image perturbation attacks such as noise, blur, cropping or compression, unless the models were re-trained again.

Related to the unseen conditions just commented, Marra *et al.* performed in [57] an interesting study in order to detect unseen types of fake generated data. Concretely, they proposed a multi-task incremental learning detection method in order to detect and classify new types of GAN generated images, without worsening the performance on the previous ones. Two different solutions regarding the position of the classifier were proposed based on the successful algorithm iCaRL for incremental learning [67]: *i)* Multi-Task MultiClassifier (MT-MC), and *ii)* Multi-Task Single Classifier (MT-SC). Regarding the experimental framework, five different GAN approaches were considered in the study, CycleGAN [63], ProGAN [47], Glow [68], StarGAN [42], and StyleGAN [40]. Their proposed detection approach, based on the XceptionNet model, achieved promising results being able to correctly detect new GAN generated images.

Attention mechanisms have also been applied to further improve the training process of the detection systems. Dang

TABLE II
**ENTIRE FACE SYNTHESIS:** COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE NOT PROVIDED IN THE ORIGINAL WORK. AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

| Study | Method | Classifiers | Best Performance | Databases (Generation) |
|---|---|---|---|---|
| McCloskey and Albright (2018) [53] | GAN-Pipeline Features | SVM | AUC = 70.0% | NIST MFC2018 |
| Wang *et al.* (2019) [54] | GAN-Pipeline Features | SVM | Acc. = 84.7% | Own (InterFaceGAN, StyleGAN) |
| Nataraj *et al.* (2019) [55] | Steganalysis Features | CNN | *EER = 12.3% [16]* | *100K-Faces (StyleGAN)* |
| Yu *et al.* (2019) [56] | Deep Learning Features | CNN | Acc. = 99.5% | Own (ProGAN, SNGAN, CramerGAN, MMDGAN) |
| Marra *et al.* (2019) [57] | Deep Learning Features | CNN + Incremental Learning | Acc. = 99.3% | Own (CycleGAN, ProGAN, Glow, StarGAN, StyleGAN) |
| Dang *et al.* (2019) [17] | Deep Learning Features | CNN + Attention Mechanism | **AUC = 100%** **EER = 0.1%** | **DFFD (ProGAN, StyleGAN)** |
| Neves *et al.* (2019) [16] | Deep Learning Features | CNN | **EER = 0.3%** | **100K-Faces (StyleGAN)** |
| | | | **EER = 4.5%** | **iFakeFaceDB** |

*et al.* carried out in [17] a complete analysis of different types of facial manipulations. They proposed to use attention mechanisms and popular CNN models such as Xception-Net and VGG16. For the entire face synthesis manipulation, the authors achieved a final 100% AUC and around 0.1% EER considering real faces from CelebA [50], FFHQ [40], and FaceForensics++ [12] databases and fake images created through ProGAN [47] and StyleGAN [40] approaches. The impressive results achieved show the importance of novel attention mechanisms [69].

Neves *et al.* performed in [16] an in-depth experimental assessment of this type of facial manipulation considering different state-of-the-art detection systems and experimental conditions, i.e., controlled and in-the-wild scenarios. Four different fake databases were considered: *i)* 150,000 fake faces collected online[11] and based on StyleGAN architecture, *ii)* the 100K-faces public database, *iii)* 80,000 synthetic faces generated using ProGAN, and *iv)* the iFakeFaceDB database, an improved version of previous fake databases in which the GAN-fingerprint information has been removed using the GANprintR approach. In controlled scenarios, they achieved similar results as the best previous studies (EER = 0.02%). However, in more challenging scenarios in which images (real and fake) come from different sources (mismatch of datasets), a high degradation of the fake detection performance is observed. Finally, the results achieved over their public iFakeFaceDB database with an EER = 4.5% for the best fake detectors remark how challenging is iFakeFaceDB even for the most advanced manipulation detection methods. Related to this enhanced fake content, Cozzolino *et al.* proposed in [70] a similar approach based on GAN to inject camera traces into synthetic images to spoof state-of-the-art fake detectors.

Finally, we also include for completeness some important references to other recent studies focused on the detection of general GAN-based image manipulations, not facial ones. In particular, we refer the reader to [71], [72].

[11]https://thispersondoesnotexist.com

## IV. IDENTITY SWAP

### A. Manipulation Techniques and Public Databases

This is one of the most popular face manipulation research lines nowadays due to the great public concerns around DeepFakes [2], [3]. It consists of replacing the face of one person in a video with the face of another person. Unlike the entire face synthesis manipulation, where manipulations are carried out at image level, in identity swap the goal is to generate realistic fake videos.

Since publicly available fake databases such as the UADFV database [73], up to the recent Celeb-DF and DFDC databases [26], [74], many visual improvements have been carried out, increasing the realism of fake videos. As a result, identity swap databases can be divided into two different generations. Table III summarises the main details of each public database, grouped in each generation. As can be seen, in this type of facial manipulation both real and fake videos are usually included in the databases.

In this section, we first provide the main details of each database, to finally summarise at a higher level the key differences among the two generations.

Three different databases are grouped in the first generation. UADFV was one of the first public databases [73]. This database comprises 49 real videos from Youtube, which were used to create 49 fake videos through the FakeApp mobile application[12], swapping in all of them the original face with the face of Nicolas Cage. Therefore, only one identity is considered in all fake videos. Each video represents one individual, with a typical resolution of 294×500 pixels, and 11.14 seconds on average.

Korshunov and Marcel introduced in [1] the Deepfake-TIMIT database. This database comprises 620 fake videos of 32 subjects from the VidTIMIT database [75]. Fake videos were created using the public GAN-based face-swapping algorithm[13]. In that approach, the generative network is adopted

[12]https://www.malavida.com/en/soft/fakeapp/
[13]https://github.com/shaoanlu/faceswap-GAN

TABLE III
**IDENTITY SWAP:** PUBLICLY AVAILABLE DATABASES.

**1st Generation**

| Database | Real Videos | Fake Videos |
|---|---|---|
| UADFV (2018) [73] | 49 (Youtube) | 49 (FakeApp) |
| DeepfakeTIMIT (2018) [1] | - | 620 (faceswap-GAN) |
| FaceForensics++ (2019) [12] | 1,000 (Youtube) | 1,000 (FaceSwap)<br>1,000 (DeepFake) |

**2nd Generation**

| Database | Real Videos | Fake Videos |
|---|---|---|
| DeepFakeDetection (2019) [77] | 363 (Actors) | 3,068 (DeepFake) |
| Celeb-DF (2019) [26] | 890 (Youtube) | 5,639 (DeepFake) |
| DFDC Preview (2019) [74] | 1,131 (Actors) | 4,119 (Unknown) |

from CycleGAN [63], using the weights of FaceNet [61]. The method Multi-Task Cascaded Convolution Networks is used for more stable detections and reliable face alignment [76]. Besides, the Kalman filter is also considered to smooth the bounding box positions over frames and eliminate jitter on the swapped face. Regarding the scenarios considered in DeepfakeTIMIT, two different qualities are considered: *i)* low quality (LQ) with images of 64×64 pixels, and *ii)* high quality (HQ) with images of 128×128 pixels. Additionally, different blending techniques were applied to the fake videos regarding the quality level.

One of the most popular databases in this type of facial manipulation is FaceForensics++ [12]. This database was introduced early 2019 as an extension of the original FaceForensics database [78], which was focused only on expression swap. FaceForensics++ contains 1000 real videos extracted from Youtube. Regarding the identity swap fake videos, they were generated using both computer graphics and DeepFake approaches (i.e., learning approach). For the computer graphics approach, the authors considered the publicly available FaceSwap algorithm[14] whereas for the DeepFake approach, fake videos were created through the DeepFake FaceSwap GitHub implementation[15]. The FaceSwap approach consists of face alignment, Gauss Newton optimization and image blending to swap the face of the source person to the target person. The DeepFake approach, as indicated in [12], is based on two autoencoders with a shared encoder that are trained to reconstruct training images of the source and the target face, respectively. A face detector is used to crop and to align the images. To create a fake image, the trained encoder and decoder of the source face are applied to the target face. The autoencoder output is then blended with the rest of the image using Poisson image editing [79]. Regarding the figures of the FaceForensics++ database, 1000 fake videos were generated for each approach. Later on, a new dataset named DeepFakeDetection, grouped inside the 2nd generation due to its higher realism, was included in the FaceForensics++ framework with the support of Google [77]. This dataset comprises 363 real videos from 28 paid actors in 16 different scenes. Additionally, 3068 fake videos are included in the dataset based on DeepFake FaceSwap GitHub implementation. It is important to remark that for both FaceForensics++ and DeepFakeDetection databases different levels of video quality are considered, in particular: *i)* RAW (original quality), *ii)* HQ (constant rate quantization parameter equal to 23), and *iii)* LQ (constant rate quantization parameter equal to 40). This aspect simulates the video processing techniques usually applied in social networks.

Regarding the databases included in the 2nd generation, we highlight the recent Celeb-DF and DFDC databases released at the end of 2019. Li *et al.* presented in [26] the Celeb-DF database. This database aims to provide fake videos of better visual qualities, similar to the popular videos that are shared on the Internet[16], in comparison to previous databases that exhibit low visual quality with many visible artifacts. Celeb-DF consists of 890 real videos extracted from Youtube, and 5,639 fake videos, which were created through a refined version of a public DeepFake generation algorithm, improving aspects such as the low resolution of the synthesised faces and colour inconsistencies.

Facebook in collaboration with other companies and academic institutions such as Microsoft, Amazon, and the MIT launched at the end of 2019 a new challenge named the Deepfake Detection Challenge (DFDC) [74]. They first released a preview dataset consisting of 1,131 real videos from 66 paid actors, and 4,119 fake videos. Fake videos were generated using two different unknown approaches. The complete DFDC dataset was released later and comprises over 470 GB of content (real and fake)[17].

Finally, to conclude this section, we discuss at a higher level the key differences among fake databases from the 1st and 2nd generations. In general, fake videos from the 1st generation are characterised by: *i)* low-quality synthesised faces, *ii)* different colour contrast among the synthesised fake mask and the skin of the original face, *iii)* visible boundaries of the fake mask, *iv)* visible facial elements from the original video, *v)* low pose variations, and *vi)* strange artifacts among sequential frames. Also, they usually consider controlled scenarios in terms of camera position and light conditions. Many of these aspects have been successfully improved in databases of the 2nd generation, not only at visual level, but also in terms of variability (in-the-wild scenarios). For example, the recent 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. Fig. 3 graphically summarises the weaknesses present in identity swap databases of the 1st generation and the improvements carried out in the 2nd generation. Finally, it also interesting to remark the larger number of fake videos included in the databases of the 2nd generation.

[14]https://github.com/MarekKowalski/FaceSwap
[15]https://github.com/deepfakes/faceswap
[16]https://www.youtube.com/channel/UCKpH0CKltc73e4wh0_pgL3g
[17]https://www.kaggle.com/c/deepfake-detection-challenge


Fig. 3. Graphical representation of the weaknesses present in identity swap databases of the 1st generation and the improvements carried out in the 2nd generation, not only at visual level, but also in terms of variability (in-the-wild scenarios). Fake images are extracted from: UADFV and FaceForensics++ (1st generation) [12], [73]; Celeb-DF and DFDC (2nd generation) [26], [74].

### B. Manipulation Detection

The development of novel methods to detect identity swap manipulations is continuously evolving. Table IV provides a comparison of the most relevant detection approaches in this area. For each study we include information related to the method, classifiers, best performance, and databases for research. We highlight **in bold** the best results achieved for each public database. It is important to remark that in some cases, different evaluation metrics are considered (e.g., AUC and EER), which complicates the comparison among studies. Finally, the results highlighted in *italics* indicate the generalisation capacity of the detection systems against different unseen databases, i.e., those databases were not considered for training. These results have been extracted from [26] and were not included in the original publications.

The first studies in this area focused on the audio-visual artifacts existed in the 1st generation of fake videos. Korshunov and Marcel evaluated in [1] baseline approaches based on the inconsistencies between lip movements and audio speech, as well as several variations of image-based systems often used in biometrics. For the first case, they considered Mel-Frequency Cepstral Coefficients (MFCCs) as audio features and distances between mouth landmarks as visual features. Principal Component Analysis (PCA) was then used to reduce the dimensionality of the blocks of features, and finally Recurrent Neural Networks (RNNs) based on Long Short-Term Memory (LSTM) to detect real of fake videos (based on [91]). For the second case, they evaluated detection approaches based on: *i)* raw faces as features, and *ii)* image quality measures (IQM) [92]. In particular, they used a set of 129 features related to measures like signal to noise ratio, specularity, blurriness, etc. PCA with Linear Discriminant Analysis (LDA), or SVM were considered for the final classification. Their proposed detection approach based on IQM+SVM provided the best results, with a final 3.3% and 8.9% EER for the LQ and HQ scenarios of the DeepfakeTIMIT database, respectively.

In this line, Matern *et al.* proposed in [80] fake detection systems based on relatively simple visual aspects such as eye colour, missing reflections, and missing details in the eye and teeth areas. Two different classifiers were considered in this analysis: *i)* a logistic regression model, and *ii)* a Multilayer Perceptron (MLP) [93]. Their proposed approach was tested using a private database, achieving a final 85.1% AUC for the MLP system.

Fake detection systems based on facial expressions and head movements have also been proposed in the literature. Yang *et al.* observes in [81] that some DeepFakes are created by splicing synthesised face regions into the original image, and in doing so, introducing errors that can be revealed when 3D head poses are estimated from the face images. Thus, they performed an study based on the differences between head poses estimated using a full set of facial landmarks (68 extracted from DLib [94]) and those in the central face regions to differentiate DeepFakes from real videos. Once these features are extracted and normalised (mean and standard deviation), a SVM is considered for the final classification. Their proposed approach was originally evaluated with the UADFV database, achieving a final 89.0% AUC. However, this pre-trained model (using UADFV database) seems not to generalise very well to other databases as depicted in Table IV.

Another interesting approach in this line was proposed by Agarwal and Farid in [82]. They proposed a detection system based on both facial expressions and head movements. For the feature extraction, the OpenFace2 toolkit was considered [95], obtaining an intensity and occurrence for 18 different facial action units related to movements of facial muscles such as cheek raiser, nose wrinkle, mouth stretch, etc. Additionally, four features related to head movements were considered. As a result, each 10-second video clip is reduced to a feature vector of dimension 190 using the Pearson correlation to measure the linearity between features. Finally, the authors considered a SVM for the final classification. Regarding the experimental framework, the authors built their own database based on videos downloaded from YouTube of persons of interest talking in a formal setting, for example, weekly address, news interview, and public speech. In most videos the person is primarily facing towards the camera. Regarding the DeepFake videos, the authors trained one GAN per person based on faceswap-GAN[18]. Their proposed approach achieved a final 96.3% AUC as the best fake detection performance, being robust against new contexts and manipulation techniques.

Another interesting research line is based on the detection of the artifacts included by the face manipulation pipeline. In [83], Li and Lyu hypothesised that some DeepFake algorithms can only create images of limited resolution, which need to be further warped to match the original faces in the source video. Such transforms leave distinctive artifacts in the resulting DeepFake videos. Thus, the authors proposed a detection system based on CNNs in order to detect the presence of such artifacts from the detected face regions and the surrounding areas. Four different CNN models were trained from scratch: VGG16 [96], ResNet50, ResNet101, and ResNet152 [97]. Their proposed detection approach was tested using the UADFV and DeepfakeTIMIT databases, outperforming the state of the art for those databases.

Li *et al.* proposed later on in [26] an improved version of the work presented in [83]. In this case, the authors included a new spatial pyramid pooling module to better handle the variations in the resolution [98]. This detection approach was evaluated using different databases, achieving state-of-the-art results in some of them.

Approaches based on mesoscopic and steganalysis features have also been proposed in the literature. Afchar *et al.* proposed in [84] two different networks composed of few layers in order to focus on the mesoscopic properties of the images: *i)* a CNN network comprised of 4 convolutional layers followed by a fully-connected layer (Meso-4), and *ii)* a modification of Meso-4 consisted of a variant of the Inception module introduced in [99], named MesoInception-4. Their proposed approach was originally tested against DeepFakes using a private database, achieving a 98.4% of fake detection accuracy for the best performance. That pre-trained detection model was tested against unseen databases in [26], proving to be a robust

[18]https://github.com/shaoanlu/faceswap-GAN

TABLE IV

**IDENTITY SWAP:** COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. RESULTS IN *italics* INDICATE THAT THEY WERE PUBLISHED IN [26], BUT NOT IN THE ORIGINAL WORK. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

| Study | Method | Classifiers | Best Performance | Databases |
|---|---|---|---|---|
| Korshunov and Marcel (2018) [1] | Audio-Visual Features | PCA+RNN | **EER = 3.3%** | **DeepfakeTIMIT (LQ)** |
| | | PCA+LDA, SVM | **EER = 8.9%** | **DeepfakeTIMIT (HQ)** |
| Matern *et al.* (2019) [80] | Visual Features | Logistic Regression MLP | AUC = 85.1% | Own |
| | | | *AUC = 70.2%* | *UADFV* |
| | | | *AUC = 77.0%* | *DeepfakeTIMIT (LQ)* |
| | | | *AUC = 77.3%* | *DeepfakeTIMIT (HQ)* |
| | | | *AUC = 78.0%* | *FF++ / DFD* |
| | | | *AUC = 66.2%* | *DFDC Preview* |
| | | | *AUC = 55.1%* | *Celeb-DF* |
| Yang *et al.* (2019) [81] | Head Pose Features | SVM | AUC = 89.0% | UADFV |
| | | | *AUC = 55.1%* | *DeepfakeTIMIT (LQ)* |
| | | | *AUC = 53.2%* | *DeepfakeTIMIT (HQ)* |
| | | | *AUC = 47.3%* | *FF++ / DFD* |
| | | | *AUC = 55.9%* | *DFDC Preview* |
| | | | *AUC = 54.6%* | *Celeb-DF* |
| Agarwal and Farid (2019) [82] | Head Pose and Facial Features | SVM | AUC = 96.3% | Own (FaceSwap, HQ) |
| Li *et al.* (2019) [26], [83] | Face Warping Features | CNN | AUC = 97.7% | UADFV |
| | | | **AUC = 99.9%** | **DeepfakeTIMIT (LQ)** |
| | | | **AUC = 99.7%** | **DeepfakeTIMIT (HQ)** |
| | | | *AUC = 93.0%* | *FF++ / DFD* |
| | | | ***AUC = 75.5%*** | *DFDC Preview* |
| | | | *AUC = 64.6%* | *Celeb-DF* |
| Afchar *et al.* (2018) [84] | Mesoscopic Features | CNN | Acc. = 98.4% | Own |
| | | | *AUC = 84.3%* | *UADFV* |
| | | | *AUC = 87.8%* | *DeepfakeTIMIT (LQ)* |
| | | | *AUC = 68.4%* | *DeepfakeTIMIT (HQ)* |
| | | | Acc. $\simeq$ 90.0% | FF++ (DeepFake, LQ) |
| | | | Acc. $\simeq$ 94.0% | FF++ (DeepFake, HQ) |
| | | | Acc. $\simeq$ 98.0% | FF++ (DeepFake, RAW) |
| | | | Acc. $\simeq$ 83.0% | FF++ (FaceSwap, LQ) |
| | | | Acc. $\simeq$ 93.0% | FF++ (FaceSwap, HQ) |
| | | | Acc. $\simeq$ 96.0% | FF++ (FaceSwap, RAW) |
| | | | *AUC = 75.3%* | *DFDC Preview* |
| | | | *AUC = 54.8%* | *Celeb-DF* |
| Zhou *et al.* (2018) [85] | Steganalysis Features + Deep Learning Features | CNN SVM | *AUC = 85.1%* | *UADFV* |
| | | | *AUC = 83.5%* | *DeepfakeTIMIT (LQ)* |
| | | | *AUC = 73.5%* | *DeepfakeTIMIT (HQ)* |
| | | | *AUC = 70.1%* | *FF++ / DFD* |
| | | | *AUC = 61.4%* | *DFDC Preview* |
| | | | *AUC = 53.8%* | *Celeb-DF* |
| Rössler *et al.* (2019) [12] | Mesoscopic Features Steganalysis Features Deep Learning Features | CNN | Acc. $\simeq$ 94.0% | FF++ (DeepFake, LQ) |
| | | | **Acc. $\simeq$ 98.0%** | **FF++ (DeepFake, HQ)** |
| | | | **Acc. $\simeq$ 100.0%** | **FF++ (DeepFake, RAW)** |
| | | | Acc. $\simeq$ 93.0% | FF++ (FaceSwap, LQ) |
| | | | **Acc. $\simeq$ 97.0%** | **FF++ (FaceSwap, HQ)** |
| | | | **Acc. $\simeq$ 99.0%** | **FF++ (FaceSwap, RAW)** |
| Nguyen *et al.* (2019) [86] | Deep Learning Features | AE + Multi-Task Learning | *AUC = 65.8%* | *UADFV* |
| | | | *AUC = 62.2%* | *DeepfakeTIMIT (LQ)* |
| | | | *AUC = 55.3%* | *DeepfakeTIMIT (HQ)* |
| | | | *AUC = 76.3%* | *FF++ / DFD* |
| | | | EER = 15.1% | FF++ (FaceSwap, HQ) |
| | | | *AUC = 53.6%* | *DFDC Preview* |
| | | | *AUC = 54.3%* | *Celeb-DF* |
| Nguyen *et al.* (2019) [87] | Deep Learning Features | Capsule Networks | *AUC = 61.3%* | *UADFV* |
| | | | *AUC = 78.4%* | *DeepfakeTIMIT (LQ)* |
| | | | *AUC = 74.4%* | *DeepfakeTIMIT (HQ)* |
| | | | *AUC = 96.6%* | *FF++ / DFD* |
| | | | *AUC = 53.3%* | *DFDC Preview* |
| | | | *AUC = 57.5%* | *Celeb-DF* |
| Dang *et al.* (2019) [17] | Deep Learning Features | CNN + Attention Mechanism | AUC = 99.4% EER = 3.1% | DFFD |
| Dolhansky *et al.* (2019) [74] | Deep Learning Features | CNN | **Precision = 93.0% Recall = 8.4%** | **DFDC Preview** |
| Güera and Delp (2018) [88] | Image + Temporal Features | CNN + RNN | Acc. = 97.1% | Own |
| Sabir *et al.* (2019) [89] | Image + Temporal Features | CNN + RNN | **AUC = 96.9%** | **FF++ (DeepFake, LQ)** |
| | | | **AUC = 96.3%** | **FF++ (FaceSwap, LQ)** |
| Tolosana *et al.* (2020) [90] | Facial Regions Features | CNN | **AUC = 100.0%** | **UADFV** |
| | | | **AUC = 99.4%** | **FF++ (FaceSwap, HQ)** |
| | | | **AUC = 91.0%** | **DFDC Preview** |
| | | | **AUC = 83.6%** | **Celeb-DF** |

approach in some cases such as with the FaceForensics++ database.

Zhou *et al.* proposed a two-stream network for face manipulation detection. In particular, the authors considered a fusion of two streams: *i)* a face classification stream based on the CNN GoogLeNet [99] to detect whether a face image is fake or not, and *ii)* a path triplet stream that is trained using steganalysis features of images patches with a triplet loss, and a SVM for the classification. The initial system was trained to detect expression swap manipulations. Nevertheless, Li *et al.* evaluated in [26] the generalisation capacity of the pre-trained model (trained using SwapMe app) to detect identity swap manipulations, resulting to be one the most robust approaches against the recent Celeb-DF database [26].

An exhaustive analysis of different fake detection methods was carried out by Rössler *et al.* using FaceForensics++ database [12]. Five different detection systems were evaluated: *i)* a CNN-based system trained through handcrafted steganalysis features [100], *ii)* a CNN-based system whose convolution layers are specifically designed to suppress the high-level content of the image [101], *iii)* a CNN-based system with a global pooling layer that computes four statistics (mean, variance, maximum, and minimum) [102], *iv)* the CNN MesoInception-4 detection system described in [84], and finally *v)* the CNN-based system XceptionNet [103] pre-trained using ImageNet database [104] and re-trained for the face manipulation detection task. In general, the detection system based on XceptionNet architecture provided the best results in both types of manipulation methods, DeepFakes and FaceSwap. In addition, the detection systems were evaluated considering different video quality levels in order to simulate the video processing of many social networks. In this real scenario, the accuracy of all detection systems decreased when lowering the video quality, remarking how challenging is this task in real scenarios.

Recent deep learning methods considered in computer vision have been applied to further improve the detection of identity swap manipulations. In [86], Nguyen *et al.* proposed a CNN system that uses multi-task learning to simultaneously detect fake videos and locate the manipulated regions. They considered a detection system based on an autoencoder. Concretely, they proposed to use a Y-shaped decoder in order to share valuable information between the classification, segmentation, and reconstruction tasks, improving the overall performance by reducing the loss. Their proposed approach was evaluated with the FaceSwap manipulation method for the FaceForensics++ database [78], achieving a best performance of 15.07% EER, far from other detection approaches. In addition, this model seems not to generalise very well for other databases, with results below 80% AUC.

Later on, the same authors presented in [87] a new fake detection system based on the recent Capsule Networks. This approach uses fewer parameters than traditional CNN with similar performance [105]–[107]. The proposed detection system was originally evaluated using FaceForensics++ database with accuracies higher than 90%. The same pre-trained detection model was tested against unseen databases in [26], showing poor generalisation results, as it happens in most fake detection systems.

Attention mechanisms have also been applied to further improve the training process of the detection systems. Dang *et al.* performed in [17] a thorough analysis of different face manipulations. They proposed a detection system based on CNN and attention mechanisms to process and improve the feature maps of the classifier model. Their proposed attention map can be implemented easily and inserted into existing backbone networks, through the inclusion of a single convolution layer, its associated loss functions, and masking the subsequent high-dimensional features. Their proposed detection approach was tested with the DFFD database (based on a combination of the previous FaceForensics++ databases and a collection of videos from the Internet). In particular, for identity swap detection, their proposed approach achieved an AUC of 99.43% and EER of 3.1%. Despite of the fact that it is difficult to provide a fair comparison among studies as different experimental protocols are considered, it is clear that their detection approach provides state-of-the-art results.

In [74], in addition to the description of the DFDC database, the authors provided baseline results using three simple detection systems: *i)* a small CNN model composed of 6 convolution layers and 1 fully-connected layer to detect low-level image manipulations, *ii)* an XceptionNet model trained using only face images, and *iii)* an XceptionNet model trained using the full image. The detection system based on XceptionNet, considering only the face image (not the full image), provided the best results with 93.0% precision and 8.4% recall.

Detection systems based not only on features at image level, but also at temporal level, along the frames of the video, have also been studied in the literature. Güera and Delp proposed in [88] a temporal-aware pipeline to automatically detect fake videos. They considered a combination of CNNs and RNNs. For the CNN, the authors used InceptionV3 [108] pre-trained using ImageNet database [104]. For the RNN system, they considered a LSTM model composed of one hidden layer with 2048 memory blocks. Finally, two fully-connected layers were included, providing the probabilities of the frame sequence being either real or fake. Their proposed approach was evaluated using a proprietary database with a final 97.1% accuracy.

In this line, Sabir *et al.* proposed a method to detect fake videos based on using the temporal information present in the stream [89]. The intuition behind this model is to exploit temporal discrepancies across frames. Thus, they considered a recurrent convolutional network similar to [88], trained in this study end-to-end instead of using a pre-trained model. Their proposed detection approach was tested through FaceForensics++ database, achieving AUC results of 96.9% and 96.3% for the DeepFake and FaceSwap methods, respectively. Only the low-quality videos were considered in the analysis.

Finally, the discriminative power of each facial region for the detection of fake videos was studied in [90]. The authors considered a fake detection system based on XceptionNet. Databases from both 1$^{st}$ and 2$^{nd}$ generations were considered in the experimental framework, concluding that poor fake detection results are achieved in the latest DeepFake video databases of the 2$^{nd}$ generation compared with the 1$^{st}$ gener-

ation, with results of 91.0% and 83.6% AUC for the DFDC Preview and Celeb-DF databases, respectively. It is important to highlight that, contrary to [26], a separate fake detection system was specifically trained for each database.

In conclusion, although many different approaches have been proposed in the literature, they all show poor generalisation results to unseen databases, as indicated in Table IV. In addition, we also highlight the poor detection results achieved by most approaches on the DeepFake databases of the 2nd generation with results below 60% AUC.

## V. Attribute Manipulation

### A. Manipulation Techniques and Public Databases

This face manipulation consists of modifying in an image some attributes of the face such as the colour of the hair or the skin, the gender, the age, adding glasses, etc. Despite the success of GAN-based frameworks for general image translations and manipulations [42], [63], [109]–[113], and in particular for face attribute manipulations [42], [43], [114]–[119], few databases are publicly available for research in this area, to the best of our knowledge. The main reason is that the code of most GAN approaches are publicly available, so researchers can easily generate their own fake databases as they like. Therefore, this section aims to highlight the latest GAN approaches in the field, from older to closer in time, providing also the link to their corresponding codes.

In [114], the authors introduced the Invertible Conditional GAN (IcGAN)[19] for complex image editing as the union of an encoder used jointly with a conditional GAN (cGAN) [120]. This approach provides accurate results in terms of attribute manipulation. However, it seriously changes the face identity of the person.

Lample *et al.* proposed in [117] an encoder-decoder architecture that is trained to reconstruct images by disentangling the salient information of the image and the attribute values directly in the latent space[20]. However, as it happens with the IcGAN approach, the generated images may lack some details or present unexpected distortions.

An enhanced approach named StarGAN[21] was proposed in [42]. Before the StarGAN approach, many studies had shown promising results in image-to-image translations for two domains in general. However, few studies had focused on handling more than two domains. In that case a direct approach would be to build different models independently for every pair of image domains. StarGAN proposed a novel approach able to perform image-to-image translations for multiple domains using only a single model. The authors trained a conditional attribute transfer network via attribute classification loss and cycle consistency loss. Good visual results were achieved compared with previous approaches. However, it sometimes includes undesired modifications from the input face image such as the colour of the skin.

Almost at the same time He *et al.* proposed in [119] attGAN[22], a novel approach that removes the strict attribute-independent constraint from the latent representation, and just applies the attribute-classification constraint to the generated image to guarantee the correct change of the attributes. AttGAN provides state-of-the-art results on realistic attribute manipulation with other facial details well preserved.

One of the latest approaches proposed in the literature is STGAN[23] [43]. In general, attribute manipulation can be tackled by incorporating an encoder-decoder or GAN. However, as commented Liu *et al.* [43], the bottleneck layer in the encoder-decoder usually provides blurry and low quality manipulation results. To improve this, the authors presented and incorporated selective transfer units with an encoder-decoder for simultaneously improving the attribute manipulation ability and the image quality. As a result, STGAN has recently outperformed the state of the art in attribute manipulation.

Despite of the fact that the code of most attribute manipulation approaches are publicly available, the lack of public databases and experimental protocols results crucial when comparing among different manipulation detection approaches. Otherwise, it is not possible to perform a fair comparison among studies. Up to now, to the best of our knowledge, the DFFD database [17] seems to be the only public database that considers this type of facial manipulations. This database comprises 18,416 and 79,960 fake images generated through FaceApp and StarGAN approaches, respectively.

### B. Manipulation Detection

Attribute manipulations have been originally studied in the field of face recognition in order to see how robust biometric systems are against physical factors such as plastic surgery, cosmetics, makeup or occlusions [127]–[131]. However, it has been the recent success of mobile applications such as FaceApp that has motivated the research community to detect digital face attribute manipulations. Table V provides a comparison of the most relevant approaches in this area. We include for each study information related to the method, classifiers, best performance, and databases for research.

Some authors propose to analyse the internal GAN pipeline to detect different artifacts between real and manipulated images. Similar to the entire face synthesis manipulations, Wang *et al.* conjectured in [54] that monitoring neuron behavior could also serve as an asset in detecting fake faces since layer-by-layer neuron activation patterns may capture more subtle features that are important for the facial manipulation detection system. Their proposed approach, named FakeSpoter, extracted as features neuron coverage behaviors of real and fake faces from deep face recognition systems (VGG-Face [59], OpenFace [60], and FaceNet [61]), and then trained a SVM for the final classification. The authors tested their proposed approach using real faces from CelebA-HQ [47] and FFHQ [40] databases and synthetic faces created through InterFaceGAN [62] and StyleGAN [40], achieving for the best

[19]https://github.com/Guim3/IcGAN
[20]https://github.com/facebookresearch/FaderNetworks
[21]https://github.com/yunjey/stargan/blob/master/README.md
[22]https://github.com/LynnHo/AttGAN-Tensorflow
[23]https://github.com/csmliu/STGAN

TABLE V
**ATTRIBUTE MANIPULATION:** COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

| Study | Method | Classifiers | Best Performance | Databases (Generation) |
|---|---|---|---|---|
| Wang *et al.* (2019) [54] | GAN-Pipeline Features | SVM | Acc. = 84.7% | Own (InterFaceGAN/StyleGAN) |
| Nataraj *et al.* (2019) [55] | Steganalysis Features | CNN | Acc. = 99.4% | Own (StarGAN/CycleGAN) |
| Bharati *et al.* (2016) [121] | Deep Learning Features (Face Patches) | RBM | Overall Acc. = 96.2%<br>Overall Acc. = 87.1% | Own (Celebrity Retouching, ND-IIITD Retouching) |
| Jain *et al.* (2019) [122] | Deep Learning Features (Face Patches) | CNN + SVM | Overall Acc. = 99.6%<br>Overall Acc. = 99.7% | Own (ND-IIITD Retouching, StarGAN) |
| Tariq *et al.* (2018) [123] | Deep Learning Features | CNN | AUC = 99.9%<br>AUC = 74.9% | Own (ProGAN, Adobe Photoshop) |
| Dang *et al.* (2019) [17] | Deep Learning Features | CNN + Attention Mechanism | **AUC = 99.9%**<br>**EER = 1.0%** | **DFFD (FaceApp/StarGAN)** |
| Wang *et al.* (2019) [124] | Deep Learning Features | DRN | AP = 99.8% | Own (Adobe Photoshop) |
| Marra *et al.* (2019) [57] | Deep Learning Features | CNN + Incremental Learning | Acc. = 99.3% | Own (Glow/StarGAN ) |
| Zhang *et al.* (2019) [125] | Spectrum Domain Features | GAN Discriminator | Acc. = 100% | Own (StarGAN/CycleGAN) |
| Rathgeb *et al.* (2020) [126] | PRNU Features | Score-Level Fusion | EER = 13.7% | Own (5 Public Apps) |

performance a final 84.7% manipulation detection accuracy using the FaceNet model.

Fake detection systems inspired in steganalysis have also been studied. As described in Sec. III-B for the entire face synthesis, Nataraj *et al.* proposed in [55] a detection system based on the combination of pixel co-occurrence matrices and CNN. They created a new fake dataset based on attribute manipulations using the StarGAN approach [42] trained through the CelebA database [50], achieving a final 99.4% accuracy for the best result.

Many studies have also focused on pure deep learning methods, either feeding the networks with face patches or with the complete face. In [121], Bharati *et al.* proposed a deep learning approach based on a Restricted Boltzmann Machine (RBM) in order to detect digital retouching of face images. The input of the detection system consisted of face patches in order to learn discriminative features to classify each image as original or retouched. Regarding the databases, the authors generated two fake databases from the original ND-IIITD database (collection B [132]) and a set of celebrity facial images downloaded from the Internet. Fake images were generated using the professional software PortraitPro Studio Max[24], considering aspects such as skin texture, shape of eyes, nose, lips and overall face, prominence of smile, lip shape, and eye colour. Their proposed approach achieved overall accuracies for manipulation detection of 96.2% and 87.1% for the celebrity and ND-IIITD retouching databases, respectively.

A similar approach based on non-overlapping face patches was presented in [122]. Jain *et al.* proposed a CNN feature extractor composed of 6 convolutional layers and 2 fully-connected layers. Additionally, residual connections were considered inspired by a ResNet architecture [97]. Finally, a SVM was used for the final classification. Regarding the experimental framework, the ND-IIITD retouched database presented in [121] was considered. Additionally, the authors considered fake images created through the StarGAN approach [42], trained using the CelebA database [50]. In general, good detection results were achieved in both manipulation approaches, achieving almost 100% manipulation detection accuracy.

Deep learning methods based on the complete face have been further studied in the literature, achieving in general very good results. Tariq *et al.* evaluated in [123] the use of different CNN architectures such as VGG16 [59], VGG19 [59], ResNet [97], or XceptionNet [103], among others. For the real face images, the CelebA database [50] was used. Regarding the fake images, two different approaches were considered: *i)* machine approaches based on GAN, in particular ProGAN [47], and *ii)* manual approach based on Adobe Photoshop CS6, including manipulations such as makeup, glasses, sunglasses, hair, and hats. For the experimental evaluation, different sizes of the images were considered (from 32×32 to 256×256 pixels). A final 99.99% AUC was obtained for the machine-created scenario whereas for the human-created scenario this value decreased to a final 74.9% AUC for the best CNN model. Thus, a high degradation of the manipulation detection performance was observed between machine- and human-created fake images.

Attention mechanisms have also been applied to further improve the training process of the detection systems. As

[24]https://www.anthropics.com/portraitpro/

described in previous sections, Dang *et al.* developed in [17] a system able to detect different types of fakes. They used attention mechanisms to process and improve the feature maps of CNN models. Regarding the attribute manipulations, two different approaches were considered: *i)* fake images created through the public FaceApp software, with up to 28 different available filters considering aspects such as hair, age, glasses, beard, and skin colour, among others; and *ii)* fake images created through the StarGAN approach [42], with up to 40 different filters. Their proposed approach was tested using their novel database DFFD, achieving very good results close to 1.0% EER (and 99.9% of AUC).

Wang *et al.* carried out in [124] an interesting research using publicly available commercial software from Adobe Photoshop (Face-Aware Liquify tool [133]) in order to synthesise new faces, and also a professional artist in order to manipulate 50 real photographs. The authors began running a human study through Amazon Mechanical Turk (AMT), showing real and fake images to the participants and asking them to classify each image into one of the classes. The results achieved remark how challenging the task is for humans, with a final 53.5% of accuracy, close to chance (50%). After the human study, the authors proposed two different automatic models: *i)* a global classification model based on Dilated Residual Networks (DRN) to predict whether the face has been warped or not, and *ii)* a local warp predictor based on the optical flow field in order to identify where manipulation occurs, and reverse them. The PWC-Net approach proposed in [134] was considered to compute the flow from original to manipulated and vice versa. Performances of 99.8% and 97.4% for automatic and manual face synthesis manipulation were achieved.

The work [57] by Marra *et al.* also described in Sec. III-B was able to correctly perform discrimination when new GANs were presented to the network and achieved a 99.3% accuracy for their proposed manipulation detection approach, based on the XceptionNet model.

A detection system based on features extracted from the spectrum domain, rather than the raw image pixels, was presented by Zhang *et al.* in [125]. Given an image as input, they applied a 2D DFT to each of the RGB channels, getting one frequency image per channel. Regarding the classifier, they proposed AutoGAN, which is a GAN simulator that can synthesise GAN artifacts in any image without needing to access any pre-trained GAN model. The generalisation capacity of their proposed approach was tested using unseen GAN models. In particular, StarGAN [42] and GauGAN [110] were considered in the evaluation. For the StarGAN approach, good detection results were achieved using the frequency domain (100%). However, for the GauGAN approach, a high degradation of the system performance, 50% accuracy, was observed. The authors claimed that this was produced due to the generator of the GauGAN is drastically different from the CycleGAN (used in training).

Finally, Rathgeb *et al.* proposed in [126] a detection system based on Photo Response Non-Uniformity (PRNU). Specifically, scores obtained from the analysis of spatial and spectral features extracted from PRNU patterns across image cells were fused. Their proposed approach was evaluated over a private database created using 5 different mobile applications, achieving an average 13.7% EER in manipulation detection.

To summarise this section, we can see that the core of most attribute manipulation detection systems are based on deep learning technology, providing in general very good results close to 100% accuracy, as indicated in Table V. This is mainly produced due to the GAN-fingerprint information present in fake images. However, as indicated in the entire face synthesis manipulation, recent studies have been proposed in the literature to remove such GAN fingerprints from the fake images while keeping very realistic appearance [16], [70], which represent a challenge even for the most advanced manipulation detectors.

## VI. Expression Swap

### A. Manipulation Techniques and Public Databases

This manipulation, also known as face reenactment, consists of modifying the facial expression of the person. We focus on the most popular techniques Face2Face and NeuralTextures, which replace the facial expression of one person in a video with the facial expression of another person (also in a video). To the best of our knowledge, the only available database for research in this area is FaceForensics++ [12], an extension of FaceForensics [78].

Initially, the FaceForensics database was focused on the Face2Face approach [44]. This is a computer graphics approach that transfers the expression of a source video to a target video while maintaining the identity of the target person. This was carried out through manual keyframe selection. Concretely, 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. Then, fake videos were generated by transferring the source expression parameters of each frame (i.e., 76 Blendshape coefficients) to the target video. Later on, the same authors presented in FaceForensics++ a new learning approach based on NeuralTextures [45]. This is a rendering approach that uses the original video data to learn a neural texture of the target person, including a rendering network. In particular, the authors considered in their implementation a patch-based GAN-loss as used in Pix2Pix [110]. Only the facial expression corresponding to the mouth was modified. It is important to remark that all data is available on the FaceForensics++ GitHub[25]. In total, there are 1,000 real videos extracted from Youtube. Regarding the manipulated videos, 2,000 fake videos are available (1,000 videos for each considered fake approach). In addition, it is important to highlight that different video quality levels are considered, in particular: *i)* RAW (original quality), *ii)* HQ (constant rate quantization parameter equal to 23), and *iii)* LQ (constant rate quantization parameter equal to 40). This aspect simulates the video processing techniques usually applied in social networks.

In addition to the Face2Face and NeuralTexture techniques considered in expression swap manipulations at video level,

[25]https://github.com/ondyari/FaceForensics

TABLE VI
**EXPRESSION SWAP:** COMPARISON OF DIFFERENT STATE-OF-THE-ART DETECTION APPROACHES. THE BEST RESULTS ACHIEVED FOR EACH PUBLIC DATABASE ARE REMARKED IN **BOLD**. FF++ = FACEFORENSICS++, AUC = AREA UNDER THE CURVE, ACC. = ACCURACY, EER = EQUAL ERROR RATE.

| Study | Method | Classifiers | Best Performance | Databases (Generation) |
|---|---|---|---|---|
| Matern *et al.* (2019) [80] | Visual Features | Logistic Regression, MLP | AUC = 86.6% | FF++ (Face2Face, RAW) |
| Afchar *et al.* (2018) [84] | Mesoscopic Features | CNN | Acc. = 83.2% | FF++ (Face2Face, LQ) |
| | | | Acc. = 93.4% | FF++ (Face2Face, HQ) |
| | | | Acc. = 96.8% | FF++ (Face2Face, RAW) |
| | | | Acc. $\simeq$ 75% | FF++ (NeuralTextures, LQ) |
| | | | Acc. $\simeq$ 85% | FF++ (NeuralTextures, HQ) |
| | | | Acc. $\simeq$ 95% | FF++ (NeuralTextures, RAW) |
| Rössler *et al.* (2019) [12] | Mesoscopic Features<br>Steganalysis Features<br>Deep Learning Features | CNN | Acc. $\simeq$ 91% | FF++ (Face2Face, LQ) |
| | | | **Acc. $\simeq$ 98%** | **FF++ (Face2Face, HQ)** |
| | | | **Acc. $\simeq$ 100%** | **FF++ (Face2Face, RAW)** |
| | | | Acc. $\simeq$ 81% | FF++ (NeuralTextures, LQ) |
| | | | **Acc. $\simeq$ 93%** | **FF++ (NeuralTextures, HQ)** |
| | | | **Acc. $\simeq$ 99%** | **FF++ (NeuralTextures, RAW)** |
| Nguyen *et al.* (2019) [86] | Deep Learning Features | Autoencoder | EER = 7.1% | FF++ (Face2Face, HQ) |
| | | | EER = 7.8% | FF++ (NeuralTextures, HQ) |
| Dang *et al.* (2019) [17] | Deep Learning Features | CNN + Attention Mechanism | **AUC = 99.4%**<br>**EER = 3.4%** | **FF++ (Face2Face, -)** |
| Sabir *et al.* (2019) [89] | Image + Temporal Features | CNN + RNN | **Acc. = 94.3** | **FF++ (Face2Face, LQ)** |
| Amerini *et al.* (2019) [135] | Image + Temporal Features | CNN + Optical Flow | Acc. = 81.6% | FF++ (Face2Face, -) |

different approaches have been recently proposed to change the facial expression in both images and videos. A very popular approach was presented in [136]. Averbuch-Elor *et al.* proposed a technique to automatically animate a still portrait using a video of a different subject, transferring the expressiveness of the subject of the video to the target portrait. Unlike Face2Face and NeuralTexture approaches that require videos from both input and target faces, in [136] just an image of the target is needed. In this line, a recent approach was recently presented in [137], providing very good results in both one-shot and few-shot learning.

Finally, we also highlight other popular approaches at image level. For example, mobile applications such as FaceApp[26] allow to easily change the level of smiling, from happier to angrier. These approaches are based on current GAN architectures. For example, Choi *et al.* showed in [42] the potential of StarGAN to change an input image to different expression levels such as angry, happy, neutral, sad, surprised, and fearful. Other recent GAN approaches that improve both the image quality of the fake images and the control editing of the parameters are InterFaceGAN [62], UGAN [138], STGAN [43], and AttGAN [119].

### *B. Manipulation Detection*

This section aims to provide an overview of the expression swap detectors at video level using the FaceForensics++ database, as this is the only publicly available database for research in this area, to the best of our knowledge. Manipulations at image level (not video) can be detected using the same approaches described in Sec. III-B and V-B.

[26]https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341

Table VI provides a comparison of the most relevant approaches in the area of expression swap detection. For each study we include information related to the method, classifiers, best performance, and databases. We highlight in **bold** the best results achieved for the only public database, FaceForensics++. It is important to remark that in some cases, different evaluation metrics are considered (e.g., AUC and EER), which makes it difficult to perform a fair comparison among the studies.

Some of the following methods were already discussed in Sect. IV-B for identity swap detection. Here we summarise the results achieved by them in detecting expression swap manipulations.

Preliminary studies have focused on the visual features existed in fake videos such as the eye colour, missing reflections, etc. In [80] by Matern *et al.*, the proposed approach was tested using the FaceForensics++ database, but only the Face2Face manipulation technique, achieving a final 86.6% AUC for the best performance.

Approaches based on mesoscopic and steganalysis features have also been studied in the literature. In [84], the proposed approach was tested using the Face2Face fake videos from the FaceForensics++ database [12], achieving in general good results, especially for RAW-quality videos. The same approach was later on tested in [12] against NeuralTextures fake videos, obtaining lower accuracy results compared with the Face2Face scenario.

Recent deep learning methods have also been applied with good results. In [12], the detection system based on XceptionNet provided the best results in both Face2Face and NeuralTextures manipulations, close to 100% on RAW quality.

In addition, the detection systems were evaluated considering different video quality levels in order to simulate the video processing of many social networks. In this real scenario, the accuracy of all detection systems was degraded with the video quality, as it happens in identity swap manipulations.

In [86], the proposed approach based on multi-task learning was evaluated with the FaceForensics++ database. For the Face2Face method, a 7.1% EER was achieved on HQ videos whereas for the NeuralTexture method, the EER increased a bit more to a final 7.8% EER in manipulation detection.

Attention mechanisms have been recently proposed in [17] to further improve the training process. The proposed detection approach was tested using the DFFD database, which for the expression swap manipulation is based only on data from FaceForensics++ database. The proposed approach achieved an AUC = 99.4% and EER = 3.4%.

Another interesting line is based on the analysis of both image and temporal information. In [89], the proposed approach based on recurrent convolutional networks was tested using the FaceForensics++ database, achieving AUC results of 94.3% for the Face2Face technique. Only the low-quality videos were considered in the analysis. Finally, in [135], Amerini *et al.* proposed the adoption of optical flow fields to exploit possible inter-frame dissimilarities, using the PWC-Net approach [134]. The optical flow is a vector field computed among consecutive frames to extract apparent motion in the scene. The use of this approach is motivated as fake videos should have unnatural optical flow due to the unusual movement of lips, eyes, etc. Preliminary results were obtained using both VGG16 and ResNet50 networks, obtaining an Acc. = 81.6% for the best performance in manipulation detection.

Finally, as stated previously, most of the approaches reported here for expression swap detection have also been used for identity swap detection as reviewed in Sec. IV-B. In general, it seems that similar features can be learnt by the fake detectors to distinguish between real and fake content, achieving good results in both types of manipulations. We highlight the potential of novel techniques such as attention mechanisms to better guide the networks during the training process, as shown in [17], achieving AUC results of 99.4% for detecting both identity swap and expression swap manipulations.

## VII. Other Face Manipulation Directions

The four classes of face manipulation techniques described before are the ones that are receiving most attention in the last few years, but they do not perfectly represent all possible face manipulations. This section discusses some other challenging and dangerous approaches in face manipulation: face morphing, face de-identification, and face synthesis based on audio or text (i.e., audio-to-video and text-to-video).

### A. Face Morphing

Face morphing is a type of face manipulation that can be used to create artificial biometric face samples that resemble the biometric information of two or more individuals [139], [140]. This means that the new morphed face image would be successfully verified against facial samples of these two or more individuals creating a serious threat to face recognition systems [141], [142]. In this sense, face morphing is a different type of facial manipulation compared to the four main types covered in this survey. Also, it is worth noting that face morphing is mainly focused on creating fake samples at image level, not video such as identity swap manipulations.

There has been a large amount of research in the field of face morphing recently [143]–[147]. A very complete review of this field has been published by Scherhag *et al.* [140] in 2019 including both morphing techniques and also morphing attack detection. The authors state that even though the large amount of publications, the research in this field is still in its infancy, with many open issues and challenges such us generating high-quality morphed images, the lack of publicly available large-scale databases or benchmarks, the lack of metrics for reporting the vulnerability of face recognition systems to morphing attacks, etc. Very recently, in [148] a thorough evaluation of differential morphing attack detection has been carried out including the application of deep face representations.

### B. Face De-Identification

The main aim of face de-identification (de-ID) is to remove the identity information present on a face image or video in order to preserve the privacy of the person [149]. This can be achieved in several ways. The simplest way can be just to obfuscate the face by blurring or pixelation (e.g., in Google Maps Street View). More sophisticated methods try to provide face images with different identities but maintaining all other factors (pose, expression, illumination, etc.) unaltered. Therefore, the concept of face de-ID is very general. One possible option to achieve face de-identification could be through face identity swap.

Earlier works in this area were based on applying face de-ID to still images. In [150] Gross *et al.* presented a multi-factor framework for de-ID, which combined linear, bilinear, and quadratic models. They showed their method was able to protect privacy while preserving data utility on an expression-variant face database. More recently, the developments of image synthesis methods based on deep neural networks, in particular GAN, have inspired new face de-ID methods such as [151], [152], which uses synthesised faces to replace the original ones. Also, in [153], the authors proposed the use of Semi-Adversarial Networks (SAN) to confound arbitrary face-based gender classifiers.

Recently, in [154] Gafni *et al.* presented in 2019 a method that provides face de-ID with convincing performance even in unconstrained videos. Their approach is based on an adversarial autoencoder coupled with a trained face classifier. This way they can achieve a rich latent space, embedding both identity and expression information. Also, in [155] a new face de-ID method based on a deep transfer model was presented. This method treats the non-identity related facial attributes as the style of the original faces, and uses a trained facial attribute transfer model to extract and map them to different faces achieving very promising results both in single images and videos.

Some other related studies in this area work directly over face representations or deep face models by eliminating there undesired or protected information like identity, gender, or facial expressions [156]–[158]. Once that protected information has been disentangled, a face image or video can then be generated based on the new representations originated in which the protected information has been eliminated, reduced, or obfuscated.

### C. Audio-to-Video and Text-to-Video

A related topic to facial expression swap is the synthesis of video from audio or text. These types of video face manipulations are also known as lip-sinc deep fakes [159].

Regarding the synthesis of fake videos from audio (audio-to-video), Suwajanakorn *et al.* presented in [109] an approach to synthesise high quality videos of a person (Obama in this case) speaking with accurate lip sync. For this, they used as input to their approach many hours of previous videos of the person together with a new audio recording. In their approach they employed a recurrent neural network (based on LSTMs) to learn the mapping from raw audio features to mouth shapes. Then, based on the mouth shape at each frame, they synthesised high quality mouth texture, and composited it with 3D pose matching to create the new video to match the input audio track, producing photorealistic results. In [160], Song *et al.* proposed an approach based on a novel conditional recurrent generation network that incorporates both image and audio features in the recurrent unit for temporal dependency, and also a pair of spatial-temporal discriminators for better image/video quality. As a result, their approach can model both lip and mouth together with expression and head pose variations as a whole, achieving much more realistic results. Also, in [161] Song *et al.* presented a dynamic method not assuming a person-specific rendering network like in [109]. In their approach they are able to generate very realistic fake videos by carrying out a 3D face model reconstruction from the input video plus a recurrent network to translate the source audio into expression parameters. Finally, they introduced a novel video rendering network and a dynamic programming method to construct a temporally coherent and photo-realistic video.

Regarding the synthesis of fake videos from text (text-to-video), in [162] Fried *et al.* proposed a method that takes as input a video of a person speaking and the desired text to be spoken, and synthesises a new video in which the persons mouth is synchronised with the new words. In particular, their method automatically annotates an input talking-head video with phonemes, visemes, 3D face pose and geometry, reflectance, expression and scene illumination per frame. Finally, a recurrent video generation network creates a photorealistic video that matches the edited transcript.

In [159], Agarwal *et al.* propose a method to detect face manipulations based on audio-to-video and text-to-video by exploiting the fact that the dynamics of the mouth shape (visemes) are occasionally inconsistent with a spoken phoneme. They focus on some particular visemes in which the mouth must be completely closed and observe that this does not happen in many manipulated videos.

## VIII. Concluding Remarks

Motivated by the ongoing success of digital face manipulations, specially DeepFakes, this survey provides a comprehensive panorama of the field, including details of up-to-date: *i)* types of facial manipulations, *ii)* facial manipulation techniques, *iii)* public databases for research, and *iv)* benchmarks for the detection of each facial manipulation group, including key results achieved by the most representative manipulation detection approaches.

Generally speaking, most current face manipulations seem easy to be detected under controlled scenarios, i.e., when fake detectors are evaluated in the same conditions they are trained for. This fact has been demonstrated in most of the benchmarks included in this survey, achieving very low error rates in manipulation detection. However, this scenario may not be very realistic as fake images and videos are usually shared on social networks, suffering from high variations such as compression level, resizing, noise, etc. Also, facial manipulation techniques are continuously improving. These factors motivate further research on the generalisation ability of the fake detectors against unseen conditions. This aspect has been preliminary studied in different works [16], [56], [57].

Fusion techniques, at a feature or score level, could provide a better adaptation of the fake detectors to the different scenarios [163]–[165]. In fact, different approaches are already based on the combination of different sources of information, e.g., Zhou *et al.* proposed in [85] a detection system based on the combination of steganalysis and pure deep learning features, whereas Rathgeb *et al.* proposed in [126] the combination of spatial and spectral features. Besides, fusion of other sources of information such as the text or the audio that accompanies the videos when uploading them to social networks could be very valuable to improve the detectors [166]–[168].

We highlight next the key aspects to improve and future trends to follow for each facial manipulation group:

- **Face Synthesis:** current manipulations are usually based on GAN architectures such as StyleGAN, providing very realistic images. Nevertheless, most detectors can easily distinguish between real and fake images, achieving accuracies close to 100%. This is produced due to fake images are characterised by specific GAN fingerprints. But, what if we are able to remove those GAN fingerprints while keeping very realistic synthetic images? Recent approaches have focused on this research line, which represents a challenge even for the best manipulation detection systems [16], [70].
- **Identity Swap:** although many different approaches have been proposed in the literature, it is certainly difficult to decide which is the best one. This is produced due to many different factors. First, most approaches are trained for a specific database and compression level, achieving in general very good results. However, they all show poor generalisation results to unseen conditions. In addition, the fact that different metrics (i.e., Acc., AUC, EER, etc.) and experimental protocols are usually considered does not help to achieve fair comparisons among studies. All

these aspects should be further considered to advance in the field.

Furthermore, we want to highlight the detection results achieved in the latest DeepFake databases of the 2nd generation such as DFDC and Celeb-DF [26], [74]. While fake detectors already achieve AUC results close to 100% in databases of the 1st generation such as UADFV and FaceForensics++ [12], [73], they all suffer from a high performance degradation on the latest ones, in particular for the Celeb-DF database with AUC results below 60% in most cases. Therefore, more efforts are needed to further improve current fake detection systems, for example, through large-scale challenges and benchmarks such as the recent DFDC[27].

- **Attribute Manipulation:** the same aspect highlighted for the face synthesis (GAN fingerprint removal) also applies here as most manipulations are based on GAN architectures. In addition, it is also interesting to remark the scarcity of public databases for research (only the DFFD database is publicly available [17]), and the lack of standard experimental protocols to perform fair comparisons among studies.
- **Expression Swap:** contrary to the identity swap, which has rapidly evolved with the release of improved DeepFake databases, the only public database in expression swap is FaceForensics++, to the best of our knowledge. This database is characterised by visual artifacts that are easy to detect, achieving therefore AUC results close to 100% in several fake detection approaches. We encourage researchers to generate and make public more realistic databases based on recent techniques [109], [136], [161].

All these aspects, together with the development of improved GAN approaches and the recent DeepFake Detection Challenge (DFDC) will foster the new generation of realistic fake images/videos [169] together with more advanced techniques for face manipulation detection.

## ACKNOWLEDGMENTS

This work has been supported by projects: PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), BIBECA (MINECO/FEDER RTI2018-101248-B-I00), and Bio-Guard (Ayudas Fundación BBVA a Equipos de Investigación Científica 2017). Ruben Tolosana is supported by Consejería de Educación, Juventud y Deporte de la Comunidad de Madrid y Fondo Social Europeo.

## REFERENCES

[1] P. Korshunov and S. Marcel, "Deepfakes: a New Threat to Face Recognition? Assessment and Detection," *arXiv preprint arXiv:1812.08685*, 2018.

[2] D. Citron, "How DeepFake Undermine Truth and Threaten Democracy," 2019. [Online]. Available: https://www.ted.com

[3] R. Cellan-Jones, "Deepfake Videos Double in Nine Months," 2019. [Online]. Available: https://www.bbc.com/news/technology-49961089

[4] BBC Bitesize, "Deepfakes: What Are They and Why Would I Make One?" 2019. [Online]. Available: https://www.bbc.co.uk/bitesize/articles/zfkwcqt

[27] https://deepfakedetectionchallenge.ai/

[5] A. Swaminathan, M. Wu and K.J.R. Liu, "Digital Image Forensics via Intrinsic Fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 101–117, 2008.

[6] H. Farid, "Image Forgery Detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.

[7] M. Stamm and K. Liu, "Forensic Detection of Image Manipulation Using Statistical Intrinsic Fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 492–506, 2010.

[8] A. Rocha, W. Scheirer, T. Boult, and S. Goldenstein, "Vision of the Unseen: Current Trends and Challenges in Digital Image and Video Forensics," *ACM Computing Surveys*, vol. 43, no. 4, pp. 1–42, 2011.

[9] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An Overview on Video Forensics," *APSIPA Transactions on Signal and Information Processing*, vol. 1, pp. 1–18, 2012.

[10] A. Piva, "An Overview on Image Forensics," *ISRN Signal Processing*, vol. 2013, pp. 1–22, 2013.

[11] P. Korus, "Digital Image Integrity - a Survey of Protection and Verification Techniques," *Digital Signal Processing*, vol. 71, pp. 1–26, 2017.

[12] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. International Conference on Computer Vision*, 2019.

[13] J. Galbally, S. Marcel, and J. Fierrez, "Biometric Anti-Spoofing Methods: A Survey in Face Recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[14] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics Systems Under Spoofing Attack: An Evaluation Methodology and Lessons Learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.

[15] S. Marcel, M. Nixon, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing (2nd Edition)*, 2019.

[16] J. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "GANprintR: Improved Fakes and Evaluation of the State-of-the-Art in Face Manipulation Detection," *arXiv preprint arXiv:1911.05351*, 2019.

[17] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2020.

[18] C. Canton, L. Davis, E. Delp, P. Flynn, S. McCloskey, L. Leal-Taixe, P. Natsev, and C. Bregler, "Applications of Computer Vision and Pattern Recognition to Media Forensics," in *Conference on Computer Vision and Pattern Recognition*, 2019. [Online]. Available: https://sites.google.com/view/mediaforensics2019

[19] B. Biggio, P. Korshunov, T. Mensink, G. Patrini, D. Rao, and A. Sadhu, "Synthetic Realities: Deep Learning for Detecting AudioVisual Fakes," in *International Conference on Machine Learning*, 2019. [Online]. Available: https://sites.google.com/view/audiovisualfakes-icml2019/

[20] L. Verdoliva and P. Bestagini, "Multimedia Forensics," in *ACM Multimedia*, 2019. [Online]. Available: https://acmmm.org/tutorials/#tut3

[21] K. Raja, N. Damer, C. Chen, A. Dantcheva, A. Czajka, H. Han, and R. Ramachandra, "Workshop on Deepfakes and Presentation Attacks in Biometrics," in *Winter Conference on Applications of Computer Vision*, 2020. [Online]. Available: https://sites.google.com/view/wacv2020-deeppab/

[22] M. Barni, S. Battiato, G. Boato, H. Farid, and N. Memon, "MultiMedia Forensics in the Wild," in *International Conference on Pattern Recognition*, 2020. [Online]. Available: https://iplab.dmi.unict.it/mmforwild/

[23] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: Driving Visual Speech with Audio," *Computer Graphics*, vol. 31, no. 2, pp. 353–361, 1997.

[24] D.P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Proc. International Conference on Learning Representations*, 2013.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Proc. Advances in Neural Information Processing Systems*, 2014.

[26] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A New Dataset for DeepFake Forensics," *arXiv preprint arXiv:1909.12962*, 2019.

[27] I. Yerushalmy and H. Hel-Or, "Digital Image Forgery Detection based on Lens and Sensor Aberration," *International Journal of Computer Vision*, vol. 92, no. 1, pp. 71–91, 2011.

[28] A.C. Popescu and H. Farid, "Exposing Digital Forgeries in Color Filter Array Interpolated Images," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3948–3959, 2005.

[29] H. Cao and A.C. Kot, "Accurate Detection of Demosaicing Regularity for Digital Image Forensics," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 899–910, 2009.

[30] Z. Lin, J. He, X. Tang, and C. Tang, "Fast, Automatic and Fine-Grained Tampered JPEG Image Detection via DCT Coefficient Analysis," *Pattern Recognition*, vol. 42, no. 11, pp. 2492–2501, 2009.

[31] Y.L. Chen and C.T. Hsu, "Detecting Recompression of JPEG Images via Periodicity Analysis of Compression Artifacts for Tampering Detection," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 2, pp. 396–406, 2011.

[32] I. Amerini, L. Ballan, R. Caldelli, A. Bimbo, and G. Serra, "A SIFT-Based Forensic Method for Copy-Move Attack Detection and Transformation Recovery," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1099–1110, 2011.

[33] D. Cozzolino, G. Poggi, and L. Verdoliva, "Splicebuster: A new blind image splicing detector," in *Proc. IEEE International Workshop on Information Forensics and Security*, 2015, pp. 1–6.

[34] A. Gironi, M. Fontani, T. Bianchi, A. Piva, and M. Barni, "A Video Forensic Technique for Detecting Frame Deletion and Insertion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 6226–6230.

[35] Y. Wu, X. Jiang, T. Sun, and W. Wang, "Exposing Video Inter-Frame Forgery based on Velocity Field Consistency," in *Proc. IEEE International Conference on Acoustics, Speech and signal Processing*, 2014, pp. 2674–2678.

[36] H. Allcott and M. Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[37] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The Science of Fake News," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[38] I. M. J. Kietzmann, L.W. Lee and T. Kietzmann, "Deepfakes: Trick or treat?" *Business Horizons*, vol. 63, no. 2, pp. 135–146, 2020.

[39] L. Verdoliva, "Media Forensics and DeepFakes: an Overview," *arXiv preprint arXiv:2001.06564*, 2020.

[40] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.

[41] E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, "Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation and COTS Evaluation," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 8, pp. 2001–2014, 2018.

[42] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[43] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.

[44] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-Time Face Capture and Reenactment of RGB Videos," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.

[45] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred Neural Rendering: Image Synthesis using Neural Textures," *ACM Transactions on Graphics*, vol. 38, no. 66, pp. 1–12, 2019.

[46] 100,000 Faces Generated by AI, 2018. [Online]. Available: https://generated.photos/

[47] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," in *Proc. International Conference on Learning Representations*, 2018.

[48] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs Leave Artificial Fingerprints?" in *Proc. IEEE Conference on Multimedia Information Processing and Retrieval*, 2019, pp. 506–511.

[49] M. Albright and S. McCloskey, "Source Generator Attribution via Inversion?" in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[50] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. International Conference on Computer Vision*, 2015.

[51] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning Face Representation From Scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[52] Q. Cao, , L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces Across Pose and Age," in *Proc. International Conference on Automatic Face & Gesture Recognition*, 2018.

[53] S. McCloskey and M. Albright, "Detecting GAN-Generated Imagery Using Color Cues," *arXiv preprint arXiv:1812.08247*, 2018.

[54] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu, "FakeSpotter: A Simple Baseline for Spotting AI-Synthesized Fake Faces," *arXiv preprint arXiv:1909.06122*, 2019.

[55] L. Nataraj, T. Mohammed, B. Manjunath, S. Chandrasekaran, A. Flenner, J. Bappy, and A. Roy-Chowdhury, "Detecting GAN Generated Fake Images Using Co-Occurrence Matrices," *arXiv preprint arXiv:1903.06836*, 2019.

[56] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Analyzing Fingerprints in Generated Images," in *Proc. International Conference on Computer Vision*, 2019.

[57] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental Learning for the Detection and Classification of GAN-Generated Images," in *Proc. International Workshop on Information Forensics and Security*, 2019.

[58] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and J. Fiscus, "MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation," in *Proc. IEEE Winter Applications of Computer Vision Workshops*, 2019, pp. 63–72.

[59] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *Proc. British Machine Vision Conference*, 2015.

[60] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A General-Purpose Face Recognition Library with Mobile Applications," in *CMU School of Computer Science*, 2016.

[61] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2015.

[62] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the Latent Space of GANs for Semantic Face Editing," *arXiv preprint arXiv:1907.10786*, 2019.

[63] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proc. International Conference on Computer Vision*, 2017.

[64] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in *Proc. International Conference on Learning Representations*, 2018.

[65] M. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos, "The Cramer Distance as a Solution to Biased Wasserstein Gradients," *arXiv preprint arXiv:1705.10743*, 2017.

[66] M. Binkowski, D. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. International Conference on Learning Representations*, 2018.

[67] S. Rebuffi, A. Kolesnikov, G. Sperl, and C. Lampert, "iCaRL: Incremental Classifier and Representation Learning," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[68] D. Kingma and P. Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions," in *Proc. Advances in Neural Information Processing Systems*, 2018.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[70] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva, "SpoC: Spoofing Camera Fingerprints," *arXiv preprint arXiv:1911.12069*, 2019.

[71] M. Huh, A. Liu, A. Owens, and A. Efros, "Fighting Fake News: Image Splice Detection Via Learned Self-Consistency," in *Proc. of the European Conference on Computer Vision*, 2018.

[72] P. Zhou, X. Han, V. Morariu, and L. Davis, "Learning Rich Features for Image Manipulation Detection," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[73] Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," in *Proc. International Workshop on Information Forensics and Security*, 2018.

[74] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Ferrer, "The Deepfake Detection Challenge (DFDC) Preview Dataset," *arXiv preprint arXiv:1910.08854*, 2019.

[75] C. Sanderson and B. Lovell, "Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference," in *Proc. International Conference on Biometrics*, 2009.

[76] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[77] Google AI, "Contributing Data to Deepfake Detection Research," 2019. [Online]. Available: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

[78] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces," *arXiv preprint arXiv:1803.09179*, 2018.

[79] P. Pérez, M. Gangnet, and A. Blake, "Poisson Image Editing," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 313–318, 2003.

[80] F. Matern, C. Riess, and M. Stamminger, "Exploiting Visual Artifacts to Expose DeepFakes and Face Manipulations," in *Proc. IEEE Winter Applications of Computer Vision Workshops*, 2019.

[81] X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2019.

[82] S. Agarwal and H. Farid, "Protecting World Leaders Against Deep Fakes," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[83] Y. Li and S. Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[84] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in *Proc. International Workshop on Information Forensics and Security*, 2018.

[85] P. Zhou, X. Han, V. Morariu, and L. Davis, "Two-Stream Neural Networks for Tampered Face Detection," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[86] H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos," *arXiv preprint arXiv:1906.06876*, 2019.

[87] H.H. Nguyen, J. Yamagishi and I. Echizen, "Use of a Capsule Network to Detect Fake Images and Videos," *arXiv preprint arXiv:1910.12467*, 2019.

[88] D. Güera and E. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proc. International Conference on Advanced Video and Signal Based Surveillance*, 2018.

[89] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[90] R. Tolosana, S. Romero-Tapiador, J. Fierrez and R. Vera-Rodriguez, "DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance," *arXiv preprint arXiv:2004.07532*, 2020.

[91] P. Korshunov and S. Marcel, "Speaker Inconsistency Detection in Tampered Video," in *Proc. European Signal Processing Conference*, 2018.

[92] J. Galbally, S. Marcel, and J. Fierrez, "Image Quality Assessment for Fake Biometric Detection: Application to Iris, Fingerprint and Face Recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 710–724, 2014.

[93] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, 2016.

[94] D. King, "DLib-ML: A Machine Learning Toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[95] T. Baltrusaitis, A. Zadeh, Y. Lim, and L. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *Proc. International Conference on Automatic Face & Gesture Recognition*, 2018.

[96] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[97] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[98] ——, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[99] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2015.

[100] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting Residual-Based Local Descriptors as Convolutional Neural Networks: an Application to Image Forgery Detection," in *Proc. ACM Workshop on Information Hiding and Multimedia Security*, 2017.

[101] B. Bayar and M. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *Proc. ACM Workshop on Information Hiding and Multimedia Security*, 2016.

[102] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing Computer Graphics from Natural Images Using Convolution Neural Networks," in *Proc. Workshop on Information Forensics and Security*, 2017.

[103] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[104] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2009.

[105] G.E. Hinton and A. Krizhevsky and S.D. Wang, "Transforming Auto-Encoders," in *International Conference on Artificial Neural Networks*, 2011, pp. 44–51.

[106] S. Sabour, N. Frosst and G.E. Hinton, "Dynamic Routing Between Capsules," in *Proc. Advances in Neural Information Processing Systems*, 2017, pp. 3856–3866.

[107] G.E. Hinton, S. Sabour and N. Frosst, "Matrix Capsules with EM routing," in *Proc. International Conference on Learning Representations Workshop*, 2018.

[108] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.

[109] S. Suwajanakorn, S. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: Learning Lip Sync From Audio," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.

[110] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[111] J. Zhu, R. Zhang, D. Pathak, T. Darrell, A. Efros, O. Wang, and E. Shechtman, "Toward Multimodal Image-to-Image Translation," in *Proc. Advances in Neural Information Processing Systems*, 2017.

[112] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks," in *Proc. International Conference on Machine Learning*, 2017.

[113] D. Bau, J. Zhu, H. Strobelt, B. Zhou, J. Tenenbaum, W. Freeman, and A. Torralba, "GAN Dissection: Visualizing and Understanding Generative Adversarial Networks," *arXiv preprint arXiv:1811.10597*, 2018.

[114] G. Perarnau, J. V. D. Weijer, B. Raducanu, and J. Álvarez, "Invertible Conditional GANs for Image Editing," in *Proc. Advances in Neural Information Processing Systems Workshops*, 2016.

[115] M. Li, W. Zuo, and D. Zhang, "Deep Identity-Aware Transfer of Facial Attributes," *arXiv preprint arXiv:1610.05586*, 2016.

[116] W. Shen and R. Liu, "Learning Residual Images for Face Attribute Manipulation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[117] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato, "Fader Networks: Manipulating Images by Sliding Attributes," in *Proc. Advances in Neural Information Processing Systems*, 2017.

[118] T. Xiao, J. Hong, and J. Ma, "ELEGANT: Exchanging Latent Encodings with GAN for Transferring Multiple Face Attributes," in *Proc. European Conference on Computer Vision*, 2018.

[119] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial Attribute Editing by Only Changing What You Want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[120] M. Mirza and S. S. Osindero, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.

[121] A. Bharati, R. Singh, M. Vatsa, and K. Bowyer, "Detecting Facial Retouching Using Supervised Deep Learning," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 9, pp. 1903–1913, 2016.

[122] A. Jain, R. Singh, and M. Vatsa, "On Detecting GANs and Retouching based Synthetic Alterations," in *Proc. International Conference on Biometrics Theory, Applications and Systems*, 2018.

[123] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. Woo, "Detecting Both Machine and Human Created Fake Face Images in the Wild," in *Proc. International Workshop on Multimedia Privacy and Security*, 2018, pp. 81–87.

[124] S. Wang, O. Wang, A. Owens, R. Zhang, and A. Efros, "Detecting Photoshopped Faces by Scripting Photoshop," *arXiv preprint arXiv:1906.05856*, 2019.

[125] X. Zhang, S. Karaman, and S. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," *arXiv preprint arXiv:1907.06515*, 2019.

[126] C. Rathgeb, A. Botaljov, F. Stockhardt, S. Isadskiy, L. Debiasi, A. Uhl, and C. Busch, "PRNU-based Detection of Facial Retouching," *IET Biometrics*, 2020.

[127] J. Kim, J. Choi, J. Yi, and M. Turk, "Effective Representation Using ICA for Face Recognition Robust to Local Distortion and Partial Occlusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 12, pp. 1977–1981, 2005.

[128] A. Dantcheva, C. Chen, and A. Ross, "Can Facial Cosmetics Affect the Matching Accuracy of Face Recognition Systems?" in *Proc. International Conference on Biometrics: Theory, Applications and Systems*, 2012, pp. 391–398.

[129] N. Kose, L. Apvrille, and J. Dugelay, "Facial Makeup Detection Technique based on Texture and Shape Analysis," in *Proc. International Conference and Workshops on Automatic Face and Gesture Recognition*, 2015.

[130] P. Majumdar, A. Agarwal, R. Singh, and M. Vatsa, "Evading Face Recognition via Partial Tampering of Faces," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[131] C. Rathgeb, A. Dantcheva, and C. Busch, "Impact and Detection of Facial Beautification in Face Recognition: An Overview," *IEEE Access*, vol. 7, pp. 152 667–152 678, 2019.

[132] P. Flynn, K. Bowyer, and P. Phillips, "Assessment of Time Dependency in Face Recognition: An Initial Study," in *Proc. International Conference on Audio-and Video-Based Biometric Person Authentication*, 2003.

[133] Adjust and Exaggerate Facial Features. Adobe Photoshop, 2016. [Online]. Available: https://helpx.adobe.com/photoshop/how-to/face-aware-liquify.html

[134] D. Sun, X. Yang, M.Y. Liu and J. Kautz, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[135] I. Amerini, L. Galteri, R. Caldelli, and A. Bimbo, "Deepfake Video Detection through Optical Flow based CNN," in *Proc. International Conference on Computer Vision*, 2019.

[136] H. Averbuch-Elor, D. Cohen-Or, J. Kopf and M.F. Cohen, "Bringing Portraits to Life," *ACM Transactions on Graphics*, vol. 36, no. 6, p. 196, 2017.

[137] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-Shot Adversarial Learning of Realistic Neural Talking Head Models," in *Proc. International Conference on Computer Vision*, 2019.

[138] D. Zhu, S. Liu, W. Jiang, C. Gao, T. Wu, and G. Guo, "UGAN: Untraceable GAN for Multi-Domain Face Translation," *arXiv preprint arXiv:1907.11418*, 2019.

[139] G. Wolberg, "Image Morphing: a Survey," *The Visual Computer*, vol. 14, no. 8-9, pp. 360–372, 1998.

[140] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face Recognition Systems Under Morphing Attacks: A Survey," *IEEE Access*, vol. 7, pp. 23 012–23 026, 2019.

[141] M. Gomez-Barrero, C. Rathgeb, U. Scherhag, and C. Busch, "Is Your Biometric System Robust to Morphing Attacks?" in *Proc. International Workshop on Biometrics and Forensics*, 2017.

[142] P. Korshunov and S. Marcel, "Vulnerability of Face Recognition to Deep Morphing," *arXiv preprint arXiv:1910.01933*, 2019.

[143] C. Kraetzer, A. Makrushin, T. Neubert, M. Hildebrandt, and J. Dittmann, "Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing," in *Proc. ACM Workshop on Information Hiding and Multimedia Security*, 2017, pp. 21–32.

[144] L.B. Zhang, F. Peng and M. Long, "Face Morphing Detection Using Fourier Spectrum of Sensor Pattern Noise," in *Proc. International Conference on Multimedia and Expo*, 2018.

[145] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch, "Detecting Morphed Face Images Using Facial Landmarks," in *Proc. International Conference on Image and Signal Processing*, 2018.

[146] F. Peng, L.B. Zhang and M. Long, Min, "FD-GAN: Face De-Morphing Generative Adversarial Network for Restoring Accomplices Facial Image," *IEEE Access*, vol. 7, pp. 75 122–75 131, 2019.

[147] M. Ferrara, A. Franco, and D. Maltoni, "Face Morphing Detection in the Presence of Printing/Scanning and Heterogeneous Image Sources," *arXiv preprint arXiv:1901.08811*, 2019.

[148] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch, "Deep Face Representations for Differential Morphing Attack Detection," *arXiv preprint arXiv:2001.01202*, 2020.

[149] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-Based Face De-Identification," in *Proc. Conference on Computer Vision and Pattern Recognition Workshop*, 2006.

[150] R. Gross, L. Sweeney, J. Cohn, F. De la Torre, and S. Baker, "Face De-Identification," in *Protecting Privacy in Video Surveillance*. Springer, 2009, pp. 129–146.

[151] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic, "I Know That Person: Generative Full Body and Face De-identification of People in Images," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[152] Q. Sun, L. Ma, S.O. Joon, L.V. Gool, B. Schiele and M. Fritz, "Natural and Effective Obfuscation by Head Inpainting," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[153] V. Mirjalili, S. Raschka, and A. Ross, "FlowSAN: Privacy-Enhancing Semi-Adversarial Networks to Confound Arbitrary Face-based Gender Classifiers," *IEEE Access*, vol. 7, pp. 99 735–99 745, 2019.

[154] O. Gafni, L. Wolf, and Y. Taigman, "Live Face De-Identification in Video," *arXiv preprint arXiv:1911.08348*, 2019.

[155] Y. Li and S. Lyu, "De-Identification Without Losing Faces," in *Proc. of the ACM Workshop on Information Hiding and Multimedia Security*, 2019.

[156] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings," in *Proc. European Conference on Computer Vision*, 2018.

[157] A. Morales, J. Fierrez, and R. Vera-Rodriguez, "SensitiveNets: Learning Agnostic Representations with Application to Face Recognition," *arXiv preprint arXiv:1902.00334*, 2019.

[158] S. Gong, X. Liu, and A. Jain, "DebFace: De-biasing Face Recognition," *arXiv preprint arXiv:1911.08080*, 2019.

[159] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches," in *Proc. Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[160] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking Face Generation by Conditional Recurrent Adversarial Network," in *Proc. International Joint Conference on Artificial Intelligence*, 2019.

[161] L. Song, W. Wu, C. Qian, R. He, and C. Loy, "Everybody's Talkin': Let Me Talk as You Want," *arXiv preprint arXiv:2001.05201*, 2020.

[162] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala, "Text-Based Editing of Talking-Head Video," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1–14.

[163] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, "Multiple Classifiers in Biometrics. Part 1: Fundamentals and Review," *Information Fusion*, vol. 44, pp. 57–64, 2018.

[164] ——, "Multiple Classifiers in Biometrics. Part 2: Trends and Challenges," *Information Fusion*, vol. 44, pp. 103–112, 2018.

[165] R. S. M. Singh and A. Ross, "A Comprehensive Overview of Biometric Fusion," *Information Fusion*, vol. 52, pp. 187–205, 2019.

[166] T. Agrawal, R. Gupta, and S. Narayanan, "Multimodal Detection of Fake Social Media Use through a Fusion of classification and Pairwise Ranking Systems," in *Proc. European Signal Processing Conference*, 2017, pp. 1045–1049.

[167] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[168] K. Shu, D. Mahudeswaran, and H. Liu, "FakeNewsTracker: a Tool for Fake News Collection, Detection, and Visualization," *Computational and Mathematical Organization Theory*, vol. 25, no. 1, pp. 60–71, 2019.

[169] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," *arXiv preprint arXiv:1912.04958*, 2019.

**Ruben Tolosana** received the M.Sc. degree in Telecommunication Engineering, and his Ph.D. degree in Computer and Telecommunication Engineering, from Universidad Autonoma de Madrid, in 2014 and 2019, respectively. In April 2014, he joined the Biometrics and Data Pattern Analytics - BiDA Lab at the Universidad Autonoma de Madrid, where he is currently collaborating as a PostDoctoral researcher. Since then, Ruben has been granted with several awards such as the FPU research fellowship from Spanish MECD (2015), and the European Biometrics Industry Award (2018). His research interests are mainly focused on signal and image processing, pattern recognition, and machine learning, particularly in the areas of face manipulation, human-computer interaction and biometrics. He is author of several publications and also collaborates as a reviewer in many different high-impact conferences (e.g., ICDAR, IJCB, ICB, BTAS, EUSIPCO, etc.) and journals (e.g., IEEE TPAMI, TCYB, TIFS, TIP, ACM CSUR, etc.). Finally, he has participated in several National and European projects focused on the deployment of biometric security through the world.

**Ruben Vera-Rodriguez** received the M.Sc. degree in telecommunications engineering from Universidad de Sevilla, Spain, in 2006, and the Ph.D. degree in electrical and electronic engineering from Swansea University, U.K., in 2010. Since 2010, he has been affiliated with the Biometric Recognition Group, Universidad Autonoma de Madrid, Spain, where he is currently an Associate Professor since 2018. His research interests include signal and image processing, pattern recognition, and biometrics, with emphasis on signature, face, gait verification and forensic applications of biometrics. He is actively involved in several National and European projects focused on biometrics. Ruben has been Program Chair for the IEEE 51st International Carnahan Conference on Security and Technology (ICCST) in 2017; and the 23rd Iberoamerican Congress on Pattern Recognition (CIARP 2018) in 2018.

**Julian Fierrez** received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Universidad Politecnica de Madrid, Spain, in 2001 and 2006, respectively. Since 2002, he has been with the Biometric Recognition Group, Universidad Politecnica de Madrid. Since 2004, he has been with the Universidad Autonoma de Madrid, where he is currently an Associate Professor. From 2007 to 2009, he was a Visiting Researcher with Michigan State University, USA, under a Marie Curie Fellowship. His research interests include signal and image processing, pattern recognition, and biometrics, with an emphasis on multibiometrics, biometric evaluation, system security, forensics, and mobile applications of biometrics. He has been actively involved in multiple EU projects focused on biometrics (e.g., TABULA RASA and BEAT), and has attracted notable impact for his research. He was a recipient of a number of distinctions, including the EAB European Biometric Industry Award 2006, the EURASIP Best Ph.D. Award 2012, the Miguel Catalan Award to the Best Researcher under 40 in the Community of Madrid in the general area of science and technology, and the 2017 IAPR Young Biometrics Investigator Award. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the IEEE TRANSACTIONS ON IMAGE PROCESSING.

**Aythami Morales** received the M.Sc. degree in telecommunication engineering from the Universidad de Las Palmas de Gran Canaria in 2006 and the Ph.D. degree from La Universidad de Las Palmas de Gran Canaria in 2011. Since 2017, he is Associate Professor with the Universidad Autonoma de Madrid. He has conducted research stays at the Biometric Research Laboratory, Michigan State University, the Biometric Research Center, Hong Kong Polytechnic University, the Biometric System Laboratory, University of Bologna, and the Schepens Eye Research Institute. He has authored over 70 scientific articles published in international journals and conferences. He has participated in national and EU projects in collaboration with other universities and private entities, such as UAM, UPM, EUPMt, Indra, Union Fenosa, Soluziona, or Accenture. His research interests are focused on pattern recognition, computer vision, machine learning, and biometrics signal processing. He has received awards from the ULPGC, La Caja de Canarias, SPEGC, and COIT.

**Javier Ortega-Garcia** received the M.Sc. degree in electrical engineering and the Ph.D. degree (cum laude) in electrical engineering from Universidad Politecnica de Madrid, Spain, in 1989 and 1996, respectively. He is currently a Full Professor at the Signal Processing Chair in Universidad Autonoma de Madrid - Spain, where he holds courses on biometric recognition and digital signal processing. He is a founder and Director of the BiDA-Lab, Biometrics and Data Pattern Analytics Group. He has authored over 300 international contributions, including book chapters, refereed journal, and conference papers. His research interests are focused on biometric pattern recognition (on-line signature verification, speaker recognition, human-device interaction) for security, e-health and user profiling applications. He chaired Odyssey-04, The Speaker Recognition Workshop, ICB-2013, the 6th IAPR International Conference on Biometrics, and ICCST2017, the 51st IEEE International Carnahan Conference on Security Technology.