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

# Ruben Tolosana; Ruben Vera-Rodriguez; Julian Fierrez et al.

## 2020

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

# Scholarcy Highlights

* F AKE 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]
* 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
* They considered a detection system based on an autoencoder. 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 [87], achieving a best performance of 15.07% Equal Error Rate (EER), far from other detection approaches
* 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
* While fake detectors already achieve Area Under the Curve (AUC) results close to 100% in databases of the 1st generation such as UADFV and FaceForensics++ [12], [82], 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

# Scholarcy Summary

## INTRODUCTION

F AKE 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].

Most of the features considered in traditional fake detection methods are highly dependent on the specific training scenario, being 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].

This survey provides an in-depth review of digital manip- media forensics, and the latest deep learning techniques.

VIII our concluding Database lation and fake detection approaches considered in general remarks, highlighting open issues and future trends

## TYPES OF FACIAL MANIPULATIONS

Facial manipulations can be categorised in four main different groups regarding the level of manipulation.

This manipulation could benefit many different sectors such as the video game and 3D-modelling industries, but it could be used for harmful applications such as the creation of very realistic fake profiles in social networks in order to generate misinformation.

Attribute Manipulation: this manipulation, 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 [42]

This manipulation process is usually carried out through GAN such as the StarGAN approach proposed in [43].

## ENTIRE FACE SYNTHESIS

This manipulation creates entire non-existent face images. Table I summarises the main publicly available databases for research on detection of image manipulation techniques relying on entire face synthesis.

FakeSpoter, extracted as features neuron coverage behaviors of real and fake faces from deep face recognition systems (i.e., VGG-Face [63], OpenFace [64], and FaceNet [65]), and trained a SVM for the final classification

The authors tested their proposed approach using real faces from CelebA-HQ [48] and FFHQ [41] databases and synthetic faces created through InterFaceGAN [66] and StyleGAN [41], achieving for the best performance a final 84.7% fake detection accuracy using the FaceNet model.

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.

We refer the reader to [80], [81]

## IDENTITY SWAP

This is one of the most popular face manipulation research lines nowadays due to the great public concerns around DeepFakes [2], [3].

They performed an study based on the differences between head poses estimated using a full set of facial landmarks (68 extracted from DLib [104]) and those in the central face regions to differentiate DeepFakes from real videos

Once these features are extracted and normalised, a SVM is considered for the final classification.

Regarding the DeepFake videos, the authors trained one GAN per person based on faceswap-GAN18

Their proposed approach achieved a final 96.3% AUC as the best fake detection performance, being robust against new contexts and manipulation techniques.

We highlight the poor detection results achieved by most approaches on the DeepFake databases of the 2nd generation with results below 60% AUC

## ATTRIBUTE MANIPULATION

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.

Their proposed approach, named FakeSpoter, extracted as features neuron coverage behaviors of real and fake faces from deep face recognition systems (VGG-Face [63], OpenFace [64], and FaceNet [65]), and trained a SVM for the final classification

The authors tested their proposed approach using real faces from CelebA-HQ [48] and FFHQ [41] databases and synthetic faces created through InterFaceGAN [66] and StyleGAN [41], achieving for the best performance a final 84.7% manipulation detection accuracy using the FaceNet model.

III-B for the entire face synthesis, Nataraj et al proposed in [58] 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 [43] trained through the CelebA database [52], achieving a final 99.4% accuracy for the best result.

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], [78], which represent a challenge even for the most advanced manipulation detectors

## EXPRESSION SWAP

This manipulation, known as face reenactment, consists of modifying the facial expression of the person.

It is important to highlight that different video quality levels are considered, in particular: i) RAW, ii) HQ, and iii) LQ

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, different approaches have been recently proposed to change the facial expression in both images and videos.

In [94], 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.

Similar to the identity swap manipulation, the authors proposed fake detectors based on I3D [122] and 3D ResNet [123] approaches, achieving promising results on the low quality videos of the FaceForensics++ database

## Method Visual Features

In [98], the proposed approach based on recurrent convolutional networks was tested using the FaceForensics++ database, achieving AUC results of 94.3% for the Face2Face technique.

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.

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

## 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

## 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 [155], [156]

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 [157], [158].

In this sense, face morphing is a different type of facial manipulation compared with the four main types covered in this survey.

In order to overcome this aspect, Raja et al has recently presented an interesting framework for morphing attack detection [159], including a publicly available database, evaluation platform, and benchmark.

Approaches based on face de-morphing have been studied in order to restore the accomplice’s facial image [166], [167]

## Face De-Identification

The main goal of face de-identification is to remove the identity information present on a face image or video in order to preserve the privacy of the person [168].

This can be achieved in several ways.

The developments of image synthesis methods based on generative deep neural networks, in particular GAN, have inspired new face de-ID methods such as [170]–[175], which use synthesised faces to replace the original ones.

In [177] 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.

Once that protected information has been disentangled, a face image or video can be generated based on the new representations originated in which the protected information has been eliminated, reduced, or obfuscated

## Audio-to-Video and Text-to-Video

A related topic to facial expression swap is the synthesis of video from audio or text.

In [183], 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 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.

In [184] Song et al presented a dynamic method not assuming a person-specific rendering network like in [125]

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.

Their proposed approach achieved good results, specially as the length of the video increases

## Findings

Features based on blinking count and period were extracted to decide whether the video is real or fake.

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 [87], achieving a best performance of 15.07% EER, far from other detection approaches.

Fake images were generated using the professional software PortraitPro Studio Max, 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.

While fake detectors already achieve AUC results close to 100% in databases of the 1st generation such as UADFV and FaceForensics++ [12], [82], 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

## 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.

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.

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.

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 [70] together with more advanced techniques for face manipulation detection