8-June-21

By:

Arpajit Adhikary & Tushar Prakash (ECE, Delhi Technological University)

Deep-Fake Generation and Detection

Report

Contents

[Deep-Fakes Generation and Detection 2](#_Toc73995219)

[Introduction 2](#_Toc73995220)

[Generation of DeepFakes 2](#_Toc73995221)

[Entire Face Synthesis 2](#_Toc73995222)

[Attribute Manipulation 2](#_Toc73995223)

[Identity Swap 3](#_Toc73995224)

[Expression Swap 3](#_Toc73995225)

[Other Generation Methods 3](#_Toc73995226)

[Detection of DeepFakes 4](#_Toc73995227)

[Spatial based Detection 4](#_Toc73995228)

[Frequency based Detection 4](#_Toc73995229)

[Biological Signal based Detection 5](#_Toc73995230)

[Conclusion 5](#_Toc73995231)

[References 5](#_Toc73995232)

# Deep-Fakes Generation and Detection

## Introduction

The creation and the manipulation of facial appearance via deep generative approaches is known as DeepFake. With the advances of deep learning and high-fidelity image generation capability utilizing variants of generative adversarial networks (GAN), anyone today can produce a realistically looking face whose identity does not exist in the world. Facial manipulations, such as identity swap, in a video with a high level of realism are also majorly performed. (Juefei-Xu, et al., 2021)

## Generation of DeepFakes

For DeepFake generation methods, there are mainly four categories based on their function:

* Entire face synthesis
* Attribute manipulation
* Identity swap
* Expression swap.

### Entire Face Synthesis

Entire face synthesis aims to generate non-existent face images. The inputs of these networks are random vectors, while the outputs are high-quality fake face images.

The first work involves combining CNN and GAN is a deep convolutional generative adversarial network (DCGAN). It focuses on unsupervised learning and has comparable performance in image classification tasks with the pre-trained discriminator. The generator can easily manipulate lots of the semantic properties (i.e., manipulate attribute of a human face) of generated images profile from its interesting vector arithmetic properties.

### Attribute Manipulation

Attribute manipulation is also known as face editing, which can modify simple face attributes such as hair colour, bald, smile, and retouch complex characteristics like gender, age, etc.

IcGAN is the earliest attempt in GAN-based facial attribute manipulation. Based on an extension of the idea of cGAN, they have evaluated encoders to map an actual image into a latent space and a conditional representation, which allows the reconstruction and modification of arbitrary attributes of authentic human face images. The expression generative adversarial network (ExprGAN) has added an expression controller module that can learn an expressive and compact expression code to the encoder-decoder network. The expression controller module enables it to edit photo-realistic facial expressions with controllable expression intensity.

Although StarGAN is effective, due to the limitation of the content of the datasets, it can only generate a discrete number of expressions. GANimation has introduced a novel GAN conditioning method based on action units (AU) annotations to address this limitation. It defines the human expression with a continuous manifold of the anatomical facial movements. The magnitude of activation of each AU can be controlled independently. Different AUs can also be combined with this method.

### Identity Swap

This function can replace the face in the target image with the face in the source image. The research in this area is not as enthusiastic as in entire face synthesis and attribute manipulation.

The very first work is CycleGAN, which was proposed in 2017. The framework of CycleGAN can be used for identity swap easily. Faceswap-GAN is the implementation of CycleGAN, which provides an identity swap functionality. It simply adds the adversarial loss and perceptual loss to auto-encoder architecture.

The most famous identity swap is FaceSwap which is used for generating DeepFake dataset FaceForensics++. FaceSwap is written in Python and uses face alignment, Gauss-Newton optimization, and image blending to swap a person's face with the face of another person in a provided image.

### Expression Swap

Expression swap is similar to identity swap. It can replace the facial expression in the target image with the facial expression in the source image. It is also known as face re-enactment.

Face2Face has proposed a three-step procedure. It first uses a global non-rigid model-based bundling approach to reconstruct the shape identity of the target human based on a pre-recorded training sequence. Then it uses a transfer function to efficiently exploits deformation transfer in the low-dimensional semantic space. At last, the image-based mouth synthesis approach exploits the best matching mouth shapes offline sample sequence to generate a realistic mouth.

### Other Generation Methods

* **Style transfer**: The previous GAN generator for style transfer is only compatible with one style. To improve the diversity, GatedGAN uses gated networks to transfer multiple styles in a single model. They have added a gated transformer into the encoder-decoder.
* **Inpainting**: The new architecture proposed by them can synthesize novel image structures as well as explicitly utilize surrounding image features as references to make better predictions.
* **Rendering**: CRN has proposed a rendering network to produce a photographic image with a two-dimensional semantic specification of the scene. However, the model is only a single feed-forward network, trained end-to-end with a direct regression objective, which is pretty incredible.
* **Super resolution**: To solve the single image super-resolution (SISR) task, SAN has proposed a second-order attention network for more powerful feature expression and feature correlation learning.
* **Detection evasive**: Recent works on the detection of DeepFake images have pointed out that they are actually easily distinguishable by artifacts in their frequency spectra. Thus, some generation methods attempt to repair the flaw in the generation procedure.
* **De-identification**: They mainly obfuscate identities in photos by the head replacement for data privacy. We consider this as a direction for future research of DeepFake.

## Detection of DeepFakes

In recent years, studies are continuously working on developing various techniques to identify whether a still image or video is synthesized with AI (especially manipulated with GANs and its variants) or produced naturally with a camera.

### Spatial based Detection

* **Image forensics-based detection**: The traditional forensics-based techniques inspect the disparities in pixel-level, which is investigated by recent studies for DeepFake detection. They provide explainable clues in the detection and introduce the differences between real and fake. However, these works suffer the robustness issues when the images or videos are manipulated by simple transformations.
* **DNN-based detection**: These methods are totally data-driven by utilizing existing or designing new DNN-based models by extracting spatial features to improve the effectiveness and generalization ability of detection. However, these DNN-based detection methods all suffer from the adversarial attacks with additive noises and all the studies failed in evaluating their effectiveness in tackling adversarial noise attacks.
* **Obvious artifacts clues**: Due to the limitation of existing AI techniques, the generated DeepFakes exhibit some obvious artifacts which could be leveraged for detection by using some simple DNN models. Chai et al. (2020) investigated those local patches had redundant artifacts which could be used for differentiating fake faces. A fully convolutional approach is applied for training classifiers to focus on image patches. This approach can be well generalised across different network architectures, image datasets, etc.
* **Detection and localization**: Beyond DeepFake detection, some researchers are working on locating the manipulated regions which provides evidence for forensics and inspires future work to develop more powerful DeepFake detectors by focusing on the manipulated regions. FakeLocator investigates the architecture of existing GANs and observed that the imperfection of up-sampling methods exhibits obvious clues for detection and forgery localization where the manipulated area could be precisely marked. They employ an encoder-decoder network to extract the fake texture with devised grey-scale prediction map for better detection and localization. FakeLocator performs well across different GANs and shows strong generalization capabilities in unknown synthetic techniques.
* **Facial image pre-processing**: Some studies propose pre-processing the facial images before sending them to binary classifiers for discrimination. These works hope that the pre-processed DeepFakes could expose their fake textures to classifiers. FakeSpotter observes that the layer-by-layer neuron behaviours provide more subtle features for capturing the differences between real and fake faces. This work provides a new insight for spotting fake faces by monitoring third-party DNN-based neuron behaviours, which could be extended to other fields like fake speech detection.

### Frequency based Detection

* **GAN-based artifacts**: Instead of examining the visual artifacts, some researchers are working on investigating the imperfection design of existing GANs, which provides obvious signals for differentiating real and fake faces. They are normally working on the frequency domain.
* **Frequency domain**: The differences between real and synthesized fake faces could also be revealed in the frequency domain. Experiments demonstrate that a classifier with a simple linear model and a CNN-based model could both achieve promising results on the entire frequency spectrum.

### Biological Signal based Detection

Real still facial images and videos are produced with cameras, which are natural compared to the synthesized fake faces. The biological signal artifacts in the synthesized fake faces provide obvious clues for fake detection. These biological signals can be classified into the following categories:

* **Visual-audio inconsistency**: For DeepFake video, combining visual and audio to identify the inconsistency in fake faces is a new insight for distinguishing DeepFakes. These methods can well explain why the video is fake. A Siamese network is employed for modelling the visual and audio in videos with a combination of two triplet loss functions for measuring the similarity.
* **Visual inconsistency**: Visual inconsistency indicates that the synthesized faces are not natural, especially the shape, facial features, and landmarks of faces. The fixed size of synthesized faces leaves artifacts in warping to match the source face, which can be employed for DeepFake detection. Then, a CNN model is trained for detecting the artifacts. The lack of eye blinking is another sign for exposing DeepFakes.
* **Biological signal in video**: Biological signals in the video are not easily replicable. In FakeCatcher, six different biological signals are extracted to exploit the spatial and temporal coherence for authenticating real videos taken by the camera. Studies have shown that the heart rate could be used for detecting fake videos.

## Conclusion

This report portrays the taxonomy of various DeepFake generation methods and the categorization of various DeepFake detection methods. The requirement for both the generation and detection can be clearly derived from the stating of all the techniques and counter-techniques.

# References

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2021, 2 27). *Countering Malicious DeepFakes: Survey, Battleground, and Horizon*. Retrieved from Papers With Code: https://paperswithcode.com/paper/countering-malicious-deepfakes-survey