# A Review of Deepfake Video Generation Techniques: Architectures, Applications, and Limitations

**İsmail İLHAN**

Adıyaman University, TBMYO Computer Technologies Department, Adıyaman, Turkey
Email: *ismaililhan[at]hotmail.com*

**Abstract:** ***In recent years, deepfake video generation has evolved rapidly due to advancements in deep neural networks and artificial intelligence. This article reviews the key technologies and methods used to produce synthetic videos, including neural network models such as GANs, CNNs, and RNNs. These synthetic videos can be used as a means of animation and entertainment in movies and stories, which serve constructive purposes such as animation and storytelling, but can also be used for malicious purposes. It examines the architectures, datasets, and animation or replacement techniques employed in popular applications, while also highlighting their performance and inherent limitations. The study organizes these techniques into structured tables and diagrams for clarity and offers a reference point for researchers exploring deepfake creation and detection. It is evident that the misuse of such technologies poses ethical and societal challenges, making this a timely and critical area of inquiry.***

**Keywords:** Deepfake videos, face manipulation, GAN, video synthesis, neural networks

## 1. Introduction

Creating face synthesis used to require a number of applications and professional skills. Deep learning technology is now used in face synthesis. Thanks to the applications shared with open-source codes, synthesized videos can be easily created without the need for professional skills. Deep fake applications, which first came to the agenda in 2017, are progressing rapidly day by day [1]. Various methods are emerging [2]. Videos produced using such applications have begun to spread rapidly on social media. The development of advanced deep networks and the availability of vast amounts of data have made it increasingly difficult for humans and even advanced algorithms to distinguish them from authentic content.

Nguyen et al. reviewed the algorithms used to generate deep fakes and the deep fake detection methods in the literature. They outlined the principles behind these algorithms, the role of deep learning, and key challenges in video forensics [1]. Mirsky et al. analyzed in detail the methods used to create and detect deep fakes. They presented block diagrams of the methods and tables of their architectures [2]. Tolosana et al. comprehensively analyzed the face images in deep forged videos, processing techniques (face synthesis, expression change, identity change, attribute manipulation), and methods for forgery generation and detection. They discussed the problems and future trends of deep fakes [3]. In his work, Verdoliva described an analysis of visual media integrity verification methods, i.e. methods for detecting manipulated images and videos, emphasizing modern data-driven forensic methods. He presented the limits, issues and challenges of existing forensic tools [4]. Kietzmann et al. provided information about deep forgery and the technologies used and categorized them and presented their risks and opportunities. They proposed a new framework for combating deepfakes by making recommendations [5].

In this study, information about deep fake videos is given; the benefits and harms that it may cause according to the purposes of use are mentioned. In the third section, the methods investigated are classified and the technologies used are explained in block diagrams. In the fourth section, the proposed methods for creating deep fake videos are analyzed. The characteristics of the methods according to the animation/replacement classes and the components used in the models are presented in a table and the prominent limitations of the methods are given. The last section provides a general evaluation of the subject.

This article aims to provide a structured analysis of the key technologies and neural network architectures used in the creation of deepfake videos, offering a summarized reference for researchers and practitioners in the field. Understanding the mechanics and evolution of deepfake creation methods is crucial for developing robust detection systems and informing policy frameworks, especially given the increasing misuse of synthetic media in politics, cybersecurity, and digital communication.

## 2. Deep Fake Videos

Deepfake is a combination of the words 'deep learning' and 'fake' and is primarily a neural network-generated product. In its most common form, it involves the production and manipulation of human images. Deep fakes are artificial intelligence applications that can process audio and images separately and create new videos or merged videos with the fake audio and images they produce, ultimately producing fakes that are very similar to the original videos [1]-[4]. The algorithms used in deep fake videos improve themselves to be more realistic by imitating the facial expressions, features and voice of the individual and use deep learning technology to do so. Deep learning is formed by the use of advanced technology artificial neural networks, especially in systems that use big data and can detect features using feature extraction methods from training data [6]. Deep fake manipulation allows the face

Paper ID: SR25709023231 DOI: https://dx.doi.org/10.21275/SR25709023231

of an actor in a video to be replaced with the face of another actor. Figure 1 shows example images of famous people's faces being copied and recreated. Of course, both actors must have sufficient and usable footage.

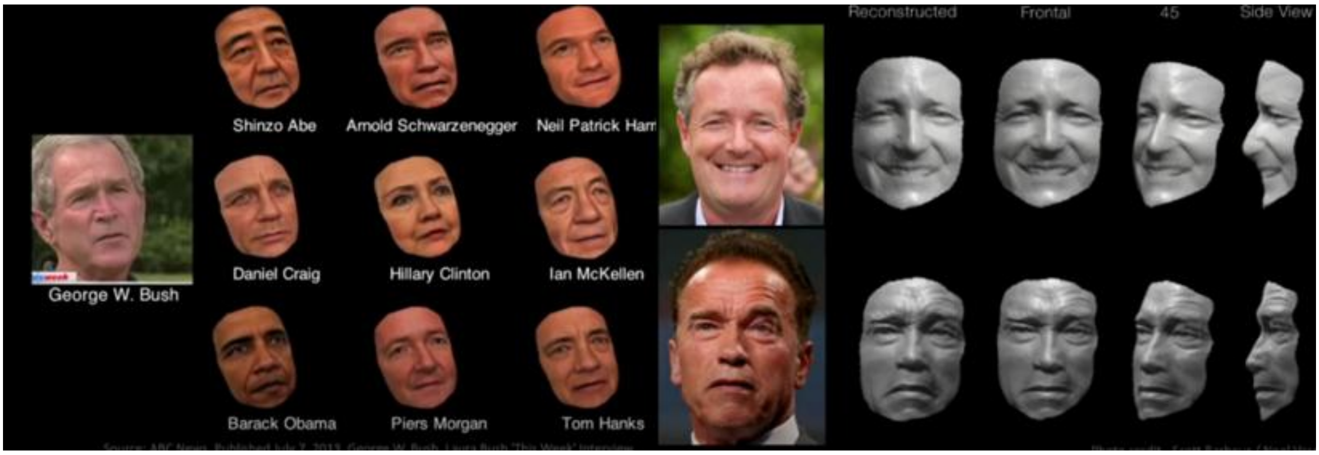

**Figure 1:** Example demonstration [7] of deep fake video creation

Fake video synthesis is an animation technique. The social impact of the malicious use of animation is huge. Deep forgery algorithms require large amounts of image or video data. This is why celebrities, actresses and politicians have become targets for deep fakes because of the large number of videos available. As a result, a lot of turmoil can occur. Examples include deceiving the public, creating political and religious tensions, defaming election campaigns, influencing financial markets, sabotaging military and security inspections, committing identity fraud, creating fake news, making pornographic or unwanted videos of people for revenge or blackmail, etc. In contrast to these, deepfakes in a good sense are used for realistic video dubbing of movies, re-enactment of historical figures, shopping dress-up and entertainment. Nowadays, it is important to distinguish deep fake videos used for many purposes from real videos and it is necessary to build systems that can do this. Today, deep fake video detection applications and systems continue to be created. DARPA (United States Defense Advanced Research Projects Agency), Facebook, Microsoft and The Partnership on AI (The Partnership on AI) are the most important organizations in the Deepfake Detection Challenge [1].

## 3. Methods Used in the Creation of Deep Fake Videos

Deepfaking techniques can be divided into 2 main groups; learning-based methods and approaches inspired by computer graphics. Examples of learning-based methods are Deepfakes Faceswap, Faceswap-GAN. Those that use graphical techniques are Face2Face and FaceSwap. Although there are various neural networks, a structure consisting of variations and combinations of adversarial networks and encoder-decoder networks is used to create deep fake video. In general, these are Convolutional Autoencoders, Generative Adversarial Network, Convolutional Neural Network, Recurrent Neural Networks. In deep faking, large data sets with some constraints are used. These constraints include resolution, angle, fixed background, proximity, portrait shot. If there is not enough data needed, the generated video will not be realistic as there will be deficiencies. Therefore, the data set is important both in the training phase and in the production phase. Deep fakes created with many different methods can be accepted as real by users.

Manipulations are made temporally, spatially and both temporally and spatially in videos. Spatial manipulation includes block and pixel manipulation and object insertion, deletion or replacement in the xy-plane. Temporal modification is the insertion, deletion, copy and paste operations in frames; it is the modification at time t of the xy axis. Temporal-spatial modification is a change in the sequence of frames at the scene level. Deep fake video creation can be categorized as animation, modification, editing and synthesis [2]. An example of the operations performed on deep fake images created by animation or face replacement methods is shown in Figure 2 [2].

### 3.1 Revitalization

Deep fake animation is the process of using the source image (xs) to create a mouth, pose, gaze or body in the target image (xt). Mouth animation: The manipulation of the xt mouth by the xs mouth or by using voice or text as the xs source. Gaze animation: The direction of xt's eyes and the position of the eyelids are manipulated by xs. Pose animation: xt's closing position is controlled by xs. Expression animation: is the generation of xt's expression from xs. It directs the target's mouth and pose and allows a wide flexibility. Body animation: is human pose synthesis, i.e. xt's posing of the body similar to facial animation.

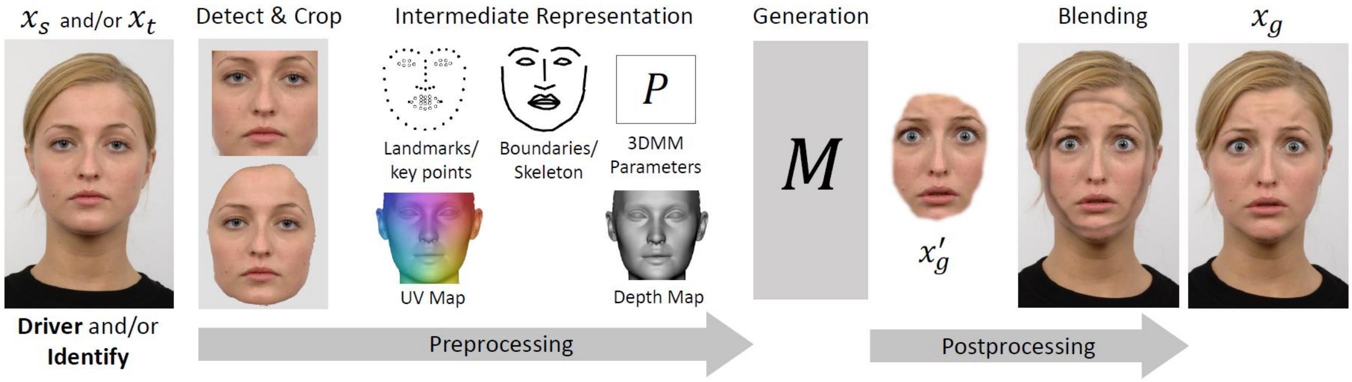

**Figure 2:** Stages of creation of animation or face replacement deep fake images [2].

### 3.2 Replacement

Deep pseudo-replacement is the process by which the content of xt is replaced by the content of xs and the identity of xs is preserved. Transfer is the replacement of xt content with xs content. Swap is the redirection by xt of the content transferred from xs to xt.

### 3.3 Editing

Deep fake editing is when xt features are added, changed or removed. For example, changing the target's clothes, beard, age, hair, hair, weight, beauty and ethnicity. Although mainly used for entertainment, ages and genders can be changed to create dynamic profiles online.

### 3.4 Synthesis

Deep fake synthesis is the creation of a new image without any targeting. Human face and body synthesis techniques can be used to create new characters for movies and games or to create fake people.

In general, the following steps are applied to create deep fake images [2].
Step 1) Allow a network to work directly on the image and perform the mapping for itself.
Step 2) Train an Encoder-Decoder network to separate the facial expression and then edit/modify the encodings of the target before passing through the decoder.
Step 3) Add an additional encoding (e.g. Action Units or Embedding) before transmitting to the decoder.
Step 4) Transform the interface/body representation into the desired identity/expression before rendering.
(e.g. transform the boundaries with a secondary mesh or create a 3D model of the target with the desired wording).
Step 5) Use the optical flow field from subsequent frames in a source video to run the generator.
Step 6) Create a combination of 3D rendering, distorted image or created content and original content (hair, skin, scene, etc.). Pass the product through another network to improve realism and quality.

Although there are various types of neural networks, a structure consisting of variations and combinations of adversarial networks and encoder-decoder networks is used to create deep fake video. Their basic technologies are described below.

### 3.5 Autoencoders (Convolutional Autoencoders)

One of the well-known deep learning techniques. Deep fake video was first developed by Reddit user FakeApp using autoencoder-decoder matching structure [2], [8]. In this method, the autoencoder extracts the features of the facial images and the decoder uses them to reconstruct the facial images. In this method, shown in Figure 3, there are two encoder-decoders where the encoder's parameters are shared between two network pairs.

While the encoders are connected, the decoders are disparate and there is no connection between them. The fact that the encoders are connected allows it to find and learn the similarity between the images of the two faces. This approach has been used in DeepFaceLab, DFaker and DeeFake-tf.

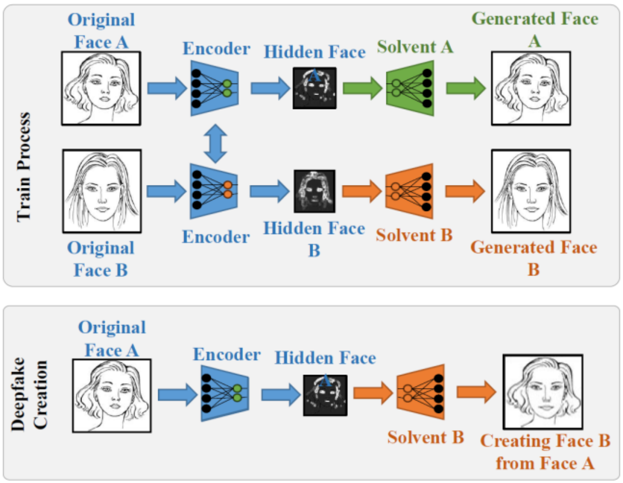

**Figure 3:** Block representation of the structure of autoencoders [1], [8]

### 3.6 GAN (Generative Adversarial Network)

It consists of 2 neural networks working in opposition to each other. As shown in Figure 4, the generator tries to create a realistic image while the parser tries to decide whether the image is fake or not. If the generator fools the parser network, the parser improves its decision-making power by referring to the information. Similarly, if the parser realizes that the image created by the generator is fake, the generator updates itself to make more realistic images. This never-ending cycle continues until eventually an image, video or sound is created that cannot be recognized as fake by the human eye. After training, the parser network is removed and only the generative network is used to create content. The most prominent GAN types are VanillaGAN, DCGAN, InfoGAN, DiscoGAN, WasssersteinGAN.

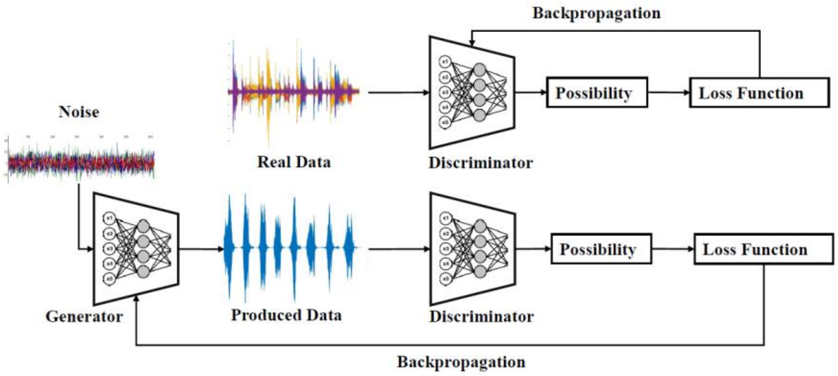

**Figure 4:** Block representation [9] of GAN structure

CNN (Convolutional Neural Networks): CNN learns the pattern hierarchy in the data and is more efficient in processing images. It is used to detect artifacts in deep fake images. CNN processes the image with various layers as shown in Figure 5. These are:

**Convolutional Layer:** It is responsible for detecting the features of the image. This layer applies some filters to the image to extract low and high-level features from the image.

**Non-Linearity Layer (RELU):** Introducing non-linearity into the system

**Pooling (Down sampling) Layer:** Reduces the number of weights and checks compliance

**Flattening Layer (flattening):** Prepares data for classical Neural Network

**Fully-Connected Layer:** Standard Neural Network used in classification

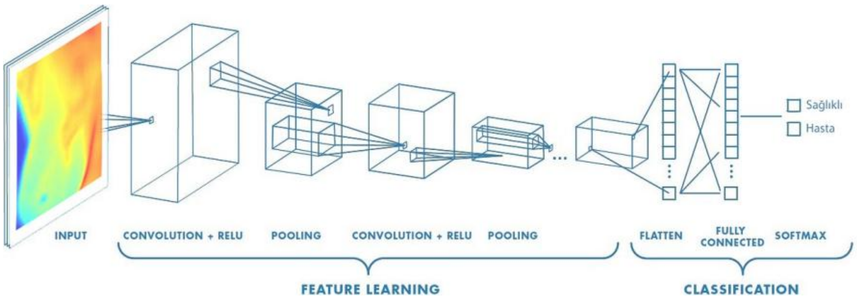

**Figure 5:** Block representation [10] of CNN structure

### 3.7 RNN (Recurrent Neural Networks)

RNN is a type of neural network that can process sequential and variable length data. As shown in Figure 6, the network remembers its previous state and can use it for its current state. Recurrent structures differ from feedforward structures in that they use their output as input for the next process. We can say that recurrent networks have a memory. The reason for adding memory to a network is that the input set, which comes in a certain order, has a meaning for the output.

In deep forging, RNNs are often used for audio and video processing. More advanced versions of RNNs include LSTM (Long Short-Term Memory), GRU (Gate Recurrent Units) neural networks.

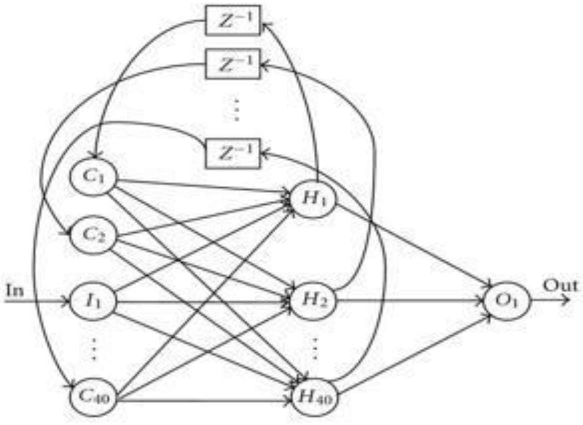

**Figure 6:** Illustration of the RNN structure [11]

## 4. Analysis of the Methods Used in the Generation of Deep Forged Videos

Many methods have been used in different models to create deep fake videos. The list given in Table 1 shows the applications created in different structures with the animation model [2]. In the majority of applications, portraits are used for source and target images. There are also applications that use voice, text and body images as sources. There are different numbers of encoders, decoders, parsers and other networks in the structures of the applications. The applications presented different animation methods such as mouth, expression, gaze, pose and body. All of the apps used a matching technique. The outputs of the applications are images or videos, ranging from low quality to high quality.

Almost all of the replacement models used portraits for the source and target images. Few of the replacement models used the transfer method, most used the swap method. Model structures and matching techniques are similar to animation models. When making fake videos, different loss functions are applied to create more pleasing visual videos. Additional algorithms can be applied to avoid temporal inconsistencies in the videos. In this way, smoother deep fake videos are created.

Deep fake video generation models are briefly described in the network.

FT-GAN is a method that creates an image from text. It synthesizes high-quality images with a fully trained GAN using a text encoder and image decoder. It gives acceptable results compared to previous proposed text-image synthesis methods. It has some disadvantages in the training phase [9].

Recycle-GAN; using temporal information in the synthesis process, combining the effect of spatial and temporal constraints to create videos that preserve the style of identity in a given space [12].

Synthesizing Obama; a method of facial animation with voice synthesizing a video of President Obama [13]. This method uses an RNN technique trained with millions of video frames to synthesize the mouth shape from audio. In the method, new mouth textures are created by synthesizing the mouth region with audio data. Then, new video images are created by taking appropriate head and torso images from the trained datasets. It synthesizes the movement and appearance of the mouth and head for a realistic speech. It applies several intermediate operations for pixel alignment and quality image. These include bending to match the jawline and synthesizing facial texture and teeth.

ReenactGAN performs animation synthesis by applying an efficient and reliable boundary-based transfer. It uses a transformer to fit the boundaries while transferring the facial movements and expressions of the source image to the target image, and uses a target-specific decoder to generate the target image. Since it has a feed-forward structure, it can work in real time [14].

**Table 1:** Deep fake video creation models (P: portrait, B: body, A: audio, V: video)

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2017 | FT-GAN | LSTM,CNN,GAN | Face animation | + | + | + | * | | 2 | 2 | 3 | 0 | P | P | SCUText2face |
| 2018 | Recycle-GAN | GAN | Image identity synthesis | + | + | + | | | 4 | 4 | 2 | 0 | P | - | Viper |
| 2018 | DeepFaceLab | GAN | Face switching | + | + | + | | | 1 | 2 | 1 | 1 | PV | - | FaceForensics++ |
| 2017 | Syth. Obama | LSTM | Mouth and Pose resuscitation | + | | + | | | 0 | 0 | 0 | 1 | A | PV | Obama video |
| 2018 | ReenactGAN | GAN | Face animation | + | + | + | | | 1 | 1 | - | 1 | P | P | Celebrity Video/ Boundary Estimation Dataset, DISFA |
| 2018 | Vid2vid | Autoencoders | Synthesis of human poses | + | + | + | * | + | 3 | 3 | 2 | 1 | PV | - | YouTube dancing videos, Street-scene videos, Face videos |
| 2019 | Everybody D. N. | GAN | Synthesis of human poses | | + | + | | + | 0 | 2 | 4 | 2 | B | - | YouTube short videos |
| 2019 | Few-shot Vid2Vid | - | Synthesis of human poses | + | + | + | * | + | 3 | 3 | 2 | 4 | PB | PB | YouTube dancing videos, Street-scene videos, Face videos |
| 2018 | paGAN | GAN | Facial animation (Real-time) | + | + | + | + | | 1 | 1 | 1 | 1 | P | P | Chicago Face Dataset, e compound facial expressions (CFE), Radbound Faces |
| 2018 | X2Face | U-Net, pix2pix | Facial revitalization | + | + | + | | | 2 | 2 | 0 | 1 | P | P | VoxCeleb video |
| 2018 | FaceID-GAN | GAN | Facial revitalization | + | + | + | | | 1 | 1 | 2 | 1 | P | P | CASIA-WebFace, CelebA, IJB-A, LFW |
| 2019 | wg-GAN | GAN | Facial animation (Real-time) | + | + | | | | 2 | 2 | 3 | 0 | P | P | MMI Facial Expression, MUG |
| 2019 | FSGAN | GAN, CNN, U-Net, Pix2pixHD | Face swapping and animation | + | + | + | | | 1 | 1 | 1 | 1 | P | P | IJB-C |
| 2019 | FaceSwapNet | pix2pix | Facial revitalization | + | + | | | | 4 | 2 | 1 | 0 | P | P | RaFD |
| 2019 | FusionNet | U-Net | Facial revitalization | + | + | + | • | | 1 | 2 | 3 | 3 | P | P | EOTT, CelebA, RAF-DB, FFHQ |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2019 | Speech2Vid | CNN | Facial revitalization | + | | | | | 3 | 1 | 0 | 2 | A | PV | VGG Face, VoxCeleb2, LRS2 |
| 2020 | MarioNETte | Autoencoders | Facial revitalization | + | + | + | | | 2 | 2 | 1 | 3 | P | P | VoxCeleb1, CelebV |
| 2016 | Face2Face | Graphics-Based | Face Swap (Real-time) | | | | | | - | - | - | 3 | P | P | Youtube |
| | FaceSwap | Graphics-Based | Face Swap (Real-time) | | | | | | - | - | - | 3 | P | P | Youtube |
| 2018 | FaceSwap GAN | GAN | Face Swap | | | | | | 1 | 2 | 2 | 1 | - | P | Youtube |
| 2018 | DeepFaceLab | GAN, TrueFace | Face Swap | | | | | | 1 | 2 | 0-1 | 0 | - | P | FaceForensics++ |
| 2017 | Fast Face Swap | | Face Swap | | | | | | 0 | 0 | 0 | 2 | P | P | |
| 2018 | RSGAN | GAN, Separator networks | Face Swap | | | | | | 4 | 3 | 2 | 1 | P | P | CelebA |
| 2019 | FS Face Trans. | GAN | Face Swap | | | | | | 1 | 1 | 2 | 2 | P | P | CelebA |
| 2019 | FaceShifter | GAN | Face Swap | | | | | | 3 | 3 | 3 | 0 | P | P | FaceForensics++ |

In the Vid2Vid model, there are two separators: image and video separators. The first one is a discriminator that distinguishes between (real image, corresponding semantic image) or (generated image, corresponding semantic image) and evaluates whether the image generated from the semantic image is reasonable or not. Image discriminators are introduced with multiple scales. The latter distinguishes between (the real image, the corresponding optical flow at the previous time) or (the generated image, the corresponding optical flow at the previous time) and determines whether the motion of the video is natural. The training is done in a spatio-temporally progressive manner. Simply put, it is a learning method that starts learning with a small number of frames and coarse resolution and gradually increases the number of frames and resolution, alternately [15].

In the Everybody Dance Now model, the source person's movements in the video are transferred to the target person's movements. To perform the process, the pipeline is divided into three stages. These are pose detection, overall pose normalization and mapping from the normalized pose bar figures to the target. In the pose detection stage, a pre-trained state-of-the-art pose detector is used to generate pose bar figures given frames from the source video. The overall pose normalization phase takes into account the differences between the source and target body shapes and positions within the frame. Finally, a system is used to learn to match from pose bar figures to images of the target person using an adversarial training network [16].

Although the method does not use temporal information during training, temporally stable video sequences are generated that can be fully controlled. After training, it generates dynamic image-based avatars that can be controlled in real time on mobile devices. To do so, it computes a set of output images from textures in UV space. Using the expressive blend shapes of a subject at runtime, it linearly blends these base textures to achieve the desired appearance [17].

X2Face is a method that synthesizes the target image using image, audio and pose sources [18].

FaceID-GAN generates identity-preserving videos by distinguishing between real and synthesized identities using a face identity classifier in a GAN structure. The identity classifier is used to extract identity features from both the input (real) and output (synthesized) face images of the generator, greatly reducing the training difficulty of GAN [19].

wg-GAN; in this method, a real-time synthesis is performed on a single image. The method factorizes the nonlinear geometric transformations exhibited in facial expressions with lightweight 2D curves and leaves the appearance detail synthesis to conditional generative neural networks for high-quality facial animation generation [20].

FSGAN is a method for face replacement and animation that does not require a prior training process. It is an RNN-based implementation that uses both poses and expressions and performs image synthesis using Poisson optimization in its architecture [21].

FaceSwapNet uses two encoders and a decoder to adapt anyone's face to target individuals. Using the neutral expression of the target person as a reference image, the second module uses geometry information from the swapped landmark to generate photo-realistic and emotion-like images. In addition, a novel triple perceptual module is used to learn geometry and appearance information simultaneously from the generator [22].

FusionNet is a method with superior transfer accuracy and identity preservation. It performs face reconstruction using a single target face source. A method is used to further improve the synthesis quality in the mustache and hair regions [23].

Speech2Vid; An encoder-decoder convolutional neural network generates a video of a speaking face. The method takes still images of the target face and an audio speech segment as input and creates a video of the target face lip synchronized with the audio. The method works in real-time and can be applied to faces and audio not seen during training [24].

MarioNETte uses image attention block, target feature alignment and landmark transformer components to address the mismatch between target identity and source identity. By making adjustments to the relevant features, high quality reconstruction of invisible identities in synthesized images is performed [25].

Face2Face generates a real-time face animation video. It is a facial animation system that transfers the expressions of the source video to the target video while preserving the identity of the target person. The original implementation is based on two video input streams with manual keyframe selection. In the method, these extracted frames are used to resynthesize the face under different lighting and expressions [26], [27].

FaceSwap is a graph-based approach to transfer the face region from a source video to a target video. The face region is extracted based on sparsely detected facial landmarks. Using these landmarks, the method builds a 3D template model using blend shapes. This model is projected back to the target image using the textures of the input image, minimizing the difference between the projected shape and the localized landmarks. Finally, the generated model is blended with the image and color correction is applied. These steps are performed for all source and target frame pairs until a video is finished. The application is computationally lightweight and can be run efficiently on processors [26, 1].

Faceswap-GAN incorporates adversarial and perceptual loss to further improve face swapping operations. In Neural Tissue synthesis [26, 1], 3D reconstruction of images is done under imperfect geometry conditions and generated at real-time rates. High-level encoding of surface appearance and 3D environment is captured. Source candidates on target candidates help the network to manipulate them easily.

DeepFaceLab; It is an easy-to-use, open source, very popular face replacement application. Its highly efficient components can be modified by users to create different versions. After pre-processing such as face recognition, face alignment, face segmentation, synthesis is done with encoder-decoders [28].

RSGAN generates face images through attribute-based editing and synthesis of random face parts. It independently handles face and hair appearances in hidden spaces and then performs face replacement by replacing the hidden-gap representations of faces and reconstructing the whole face image with them [29].

FS Face Trans. is a face replacement application that can also handle gaze direction, glasses and hair elements that are consistent with a given face source. It has a GAN structure based on hidden nodes. SPADE and AdaIN modules are included to inject semantic priorities into networks [30].

FaceShifter has a two-stage structure using Adaptive Attentional Denormalization (AAD) and Heuristic Error Acknowledging Refinement Network (HEAR-Net) layers. In the first stage, multi-level target face attributes are extracted and the identity and attributes are adapted for face synthesis. In the second stage, it is used to process facial occlusions. It is able to repair abnormal regions by self-checking [31].

The latest form of deepfakes goes beyond simple face-swapping to whole-head synthesis (head puppetry), joint audiovisual synthesis (talking heads) and even whole-body synthesis.

## 5. Deep Fake Video Creation Challenges

Generative networks are data-driven and their output is a result of training data. Therefore, a large amount of sample data is required for a quality result. Efforts are being made to minimize the required training data and source data. Due to the large data capacity and intensive operations, high-capacity and fast hardware is needed in the training processes of applications. Especially GPU devices suitable for the system are preferred. Some institutions have shared high-performance computers with VPN to make them available to researchers. In addition, social media companies can also use cloud services such as AWS, Google Cloud, FloydHub, Paper Spase online. Open-source libraries PyTorch, TensorFlow and Keras are used to develop applications. Platforms such as Colab Google, Azure Notebooks, Jupyter Notebook, IBM DataPlatform Notebooks, Amazon Sagemaker, Gitup and Kaggle are platforms that can be used for machine learning applications.

Matching is cumbersome and impractical when training on multiple identities and actions. For this, paired networks such as Cycle-GAN are used.

The use of a single identity in the training process or being paired with the same identity in training can leave identity leakage in the product.

When some of the source and target images are obstructed by a hand, hair, glasses or other object, it may result in cropped images or inconsistent facial features in the eye and mouth area. To avoid this, some studies use segmentation and in-picture inpainting in the blocked areas.

Most deep pseudo-networks generate significant artifacts such as flickering because they process each frame separately without the context of previous frames. To overcome this problem, temporal coherence losses are used. RNN structures are preferred for this.

Very realistic deep fake videos are made under certain conditions. Note that the victims of deep fakes usually do not wear glasses, do not have beards and usually use a fixed camera. Most of the data used are such videos.

## 6. Conclusion

The creation of manipulated content is a rapidly evolving problem and identifying it is technically challenging. It is therefore necessary to build better detection tools. Artificial intelligence algorithms for deep fakes are several steps ahead of algorithms for deep fake detection. Deep spoofing technologies pose a critical problem because they are rapidly evolving and increasingly easy to use, and because they are rapidly shared and spread on social media. In response, research on solutions needs to focus on providing more robust, scalable and generalizable methods. The limitations of existing detection methods need to be reduced and further generalized. In the fight against deep fakes, official institutions should also take the necessary measures and make the necessary legal arrangements. Awareness programs should be conducted to prevent the effects of deep fakes on society and possible chaos.

Deep counterfeit technology, which is a new topic and a current problem, will continue to develop with new models of deep learning technology and algorithms. It seems that this problem, which emerged as a natural problem of technology, will always remain up to date.

## References

[1] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. (2019). Deep learning for deepfakes creation and detection. arXiv preprint arXiv:1909.11573, 1(2), 2.
[2] Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM computing surveys (CSUR), 54(1), 1-41.
[3] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 64, 131-148.
[4] Verdoliva, L. (2020). Media forensics and deepfakes: an overview. IEEE journal of selected topics in signal processing, 14(5), 910-932..
[5] Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat?. Business Horizons, 63(2), 135-146.
[6] Tümen, V., Söylemez, Ö. F., & Ergen, B. (2017, September). Facial emotion recognition on a dataset using convolutional neural network. In 2017 International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-5). IEEE.
[7] S. Suwajanakorn, “Synthesizing Obama: Learning lip sync from audio,” YouTube, 2017. [Online]. Available: https://www.youtube.com/watch?v=o2DDU4g0PRo. [Accessed: Feb. 10, 2021].
[8] Malavida, “FakeApp 2.2.0,” [Online]. Available: https://www.malavida.com/en/soft/fakeapp/. [Accessed: 2020]
[9] Xu, R., Zhou, Z., Zhang, W., & Yu, Y. (2017). Face transfer with generative adversarial network. arXiv preprint arXiv:1710.06090.
[10] T. Ergin, “Convolutional Neural Network (ConvNet) ya da CNN Nedir? Nasıl Çalışır?,” Medium, 2020. [Online]. Available: https://medium.com/@tuncerergin/convolutional-neural-network-convnet-yada-cnn-nedir-nasil-calisir-97a0f5d34cad. [Accessed: Feb. 10, 2021].
[11] H. Ergüder, “Recurrent Neural Network Nedir?,” Medium, 2019. [Online]. Available: https://medium.com/@hamzaerguder/recurrent-neural-network-nedir-bdd3d0839120. [Accessed: Feb. 10, 2021].
[12] Bansal, A., Ma, S., Ramanan, D., & Sheikh, Y. (2018). Recycle-gan: Unsupervised video retargeting. In Proceedings of the European conference on computer vision (ECCV) (pp. 119-135).
[13] Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG), 36(4), 1-13.
[14] Wu, W., Zhang, Y., Li, C., Qian, C., & Loy, C. C. (2018). Reenactgan: Learning to reenact faces via boundary transfer. In Proceedings of the European conference on computer vision (ECCV) (pp. 603-619).
[15] Wang, T. C., Liu, M. Y., Tao, A., Liu, G., Kautz, J., & Catanzaro, B. (2019). Few-shot video-to-video synthesis. arXiv preprint arXiv:1910.12713.
[16] Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 5933-5942).
[17] Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., ... & Roberts, R. (2018). paGAN: real-time avatars using dynamic textures. ACM Trans. Graph., 37(6), 258.
[18] Wiles, O., Koepke, A., & Zisserman, A. (2018). X2face: A network for controlling face generation using images, audio, and pose codes. In Proceedings of the European conference on computer vision (ECCV) (pp. 670-686).
[19] Shen, Y., Luo, P., Yan, J., Wang, X., & Tang, X. (2018). Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 821-830).
[20] Geng, J., Shao, T., Zheng, Y., Weng, Y., & Zhou, K. (2018). Warp-guided gans for single-photo facial animation. ACM Transactions on Graphics (ToG), 37(6), 1-12.
[21] Nirkin, Y., Keller, Y., & Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 7184-7193).
[22] Zhang, J., Zeng, X., Pan, Y., Liu, Y., Ding, Y., & Fan, C. (2019). Faceswapnet: Landmark guided many-to-many face reenactment. arXiv preprint arXiv:1905.11805, 2, 3.
[23] Zhang, Y., Zhang, S., He, Y., Li, C., Loy, C. C., & Liu, Z. (2019). One-shot face reenactment. arXiv preprint arXiv:1908.03251.
[24] Jamaludin, A., Chung, J. S., & Zisserman, A. (2019). You said that?: Synthesising talking faces from audio. International Journal of Computer Vision, 127, 1767-1779.
[25] Ha, S., Kersner, M., Kim, B., Seo, S., & Kim, D. (2020, April). Marionette: Few-shot face reenactment preserving identity of unseen targets. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 10893-10900).
[26] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1-11).
[27] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2387-2395).
[28] Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., ... & Zhang, W. (2020). Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv. arXiv preprint arXiv:2005.05535.
[29] Natsume, R., Yatagawa, T., & Morishima, S. (2018). Rsgan: face swapping and editing using face and hair

representation in latent spaces. arXiv preprint arXiv:1804.03447.

[30] Korshunova, I., Shi, W., Dambre, J., & Theis, L. (2017). Fast face-swap using convolutional neural networks. In Proceedings of the IEEE international conference on computer vision (pp. 3677-3685).

[31] Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2019). Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457.

## Author Profile

**İsmail İLHAN** received his B.S. and M.S. degrees in Computer Engineering from Fırat University in 2013 and 2025, respectively. Between 2009 and 2024, he worked as a lecturer at Muş Alparslan University. He has been working as a lecturer at Adıyaman University since 2024. He has focused his academic and professional studies mainly on artificial intelligence, deep learning and video processing technologies. He conducts applied projects and research especially in the fields of deepfake detection, real-time image analysis and secure digital media.