DOI: https://doi.org/10.33103/uot.ijccce.23.2.12

# A survey on Deep Learning Face Synthesis and Animation Techniques Used in Deepfake

Abdulamir A. Karim[1], Suha Mohammed Saleh[2]
[1,2]Computer Science Department, University of Technology, Baghdad, Iraq
[1]abdulamir.a.karim@uotechnology.edu.iq, [2]cs.19.83@grad.uotechnology.edu.iq

***Abstract***— *From big data analytics to computer vision and human-level control, deep learning has been effectively applied to a wide range of complicated challenges. However, these same deep learning advancements have also been used to develop malicious software that threatens individuals' personal data, democratic processes, and even national security. Apps backed by deep learning have lately appeared, with deepfake being one of the most notable. Deepfake algorithms can create fake images and videos that humans cannot distinguish them from authentic ones. One of the fields that deep learning accomplished major success is face synthesis and animation generation. On the other hand, it can create unethical software called deepfake that presents a severe privacy threat or even a huge security risk that can affect innocent people. This work introduces the most recent algorithms and methods used in deepfake. In addition, it provides a brief explanation of the principles that underpin these technologies and facilitates the development of this field by identifying the challenges and scopes that require further investigation in the future.*

***Index Terms***— *deep learning; face synthesis; deepfake; face animation; motion transfer.*

## I. INTRODUCTION

A deepfake (mixing "deep learning" and "fake" as one word) refers to generation by animating the target person's face onto video footage of another source individual. Using deep learning DL techniques makes it possible to make a person saying and doing something he/she did not do. Face-swap, puppet-master, and lip-sync are some categories of using DL for deepfake purposes. Applications for this technology are both innovative and productive. For example, natural video dubbing of foreign films, historical teaching through the reanimation of historical characters, and digitally putting on clothing while shopping are all possible applications [1-,3]. Deepfakes, on the other hand, are known for their unethical and malevolent characteristics, even though they have many good uses.

Many DL models can be implemented for deepfake, such as autoencoders and generative adversarial networks GANs [4 – 6]. A person's facial expressions and emotions are studied to produce face images of another person who exhibits comparable expressions and activities [7]. The main target for deepfake is celebrities and politicians since there are many images online, which is sufficient to train DL models to generate videos, images, and fake voices or speech. When deepfake technologies can be used to generate films of international leaders giving fictitious remarks for deception, it poses a threat to global security, according to experts [8, 9].

Deepfakes became more popular due to these occurrences, and technology has advanced rapidly in recent years as a result. Since 2017, more than 250 publications have been published, an increase from three papers in 2017 [1].

Received 15/November/2022; Accepted 07/December/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.12

This article intends to present them with the following information: This study includes an overview of several methodes for producing them. First, section 2 delves into the fundamentals of deepfake methods and how deep learning has made such revolutionary technologies a reality. Next, section 3 explains some notable works and methods used. Finally, in addition to discussing their flaws, this paper highlights the present limits of deepfakes. Thus, to equip the readers with a better knowledge of how current deepfakes are generated.

## II. DEEPFAKE GENERATION

The majority of deepfakes are generated by generative or autoencoder neural networks, as explained below. A high-level overview of these networks and their training will be provided here. In addition, this work will fox on two primary techniques are used in this field autoencoder and GANs.

### A. Autoencoder

In this method, the autoencoder consists of two parts, the encoder, which extracts latent space from face images, and the decoder, which reconstructs them from their latent space. The latent space is generated in the middle of the autoencoder, representing the bottleneck layer in this structure, as represented in *Fig. 1*. Deep fake usually uses many encoders and decoders then modifies the codings producing high results. When the encoder and decoder are identical, the structure is referred to as autoencoder. When the encoder learns the decoder's posterior distribution via a variant, on the other hand, then the structure is called variational autoencoder VAE. In the case of VAE, the encoder responds more effectively to resampling and adjustment, and the outputs are better overall.

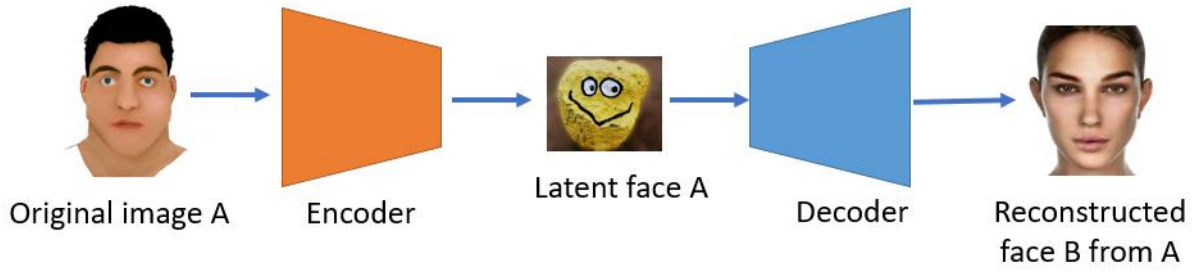


FIG. 1. AUTOENCODER STRUCTURE.

### B. Gans

Generally speaking, a GAN comprises two neural networks that compete: the generator and the discriminator denoted as G and D, respectively, *Fig. 2* illustrates GAN structure. To deceive D, G generates false samples $x_G$, and D trains to distinguish between actual images ($x \in X$) and fake images ($x_G = G(z)$ where $z \sim N$, $N$ represents random noise to start generating images). Thus, Separate adversarial loss functions are utilized for training D and G. Eqs. (1,2) define the loss function used by the generator and the discriminator, respectively:

$$L(G) = min\, log(1 - D(G(z))) \quad (1)$$

$$L(D) = max\, log\, D(x) + log(1 - D(G(z))) \quad (2)$$

G learns how to create samples indistinguishable from the original distribution due to playing this zero-sum game over time. Once D has been trained, it is removed from the system, and G produces results. When performed on images, photorealistic in appearance images are produced.

Received 15/November/2022; Accepted 07/December/2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.12

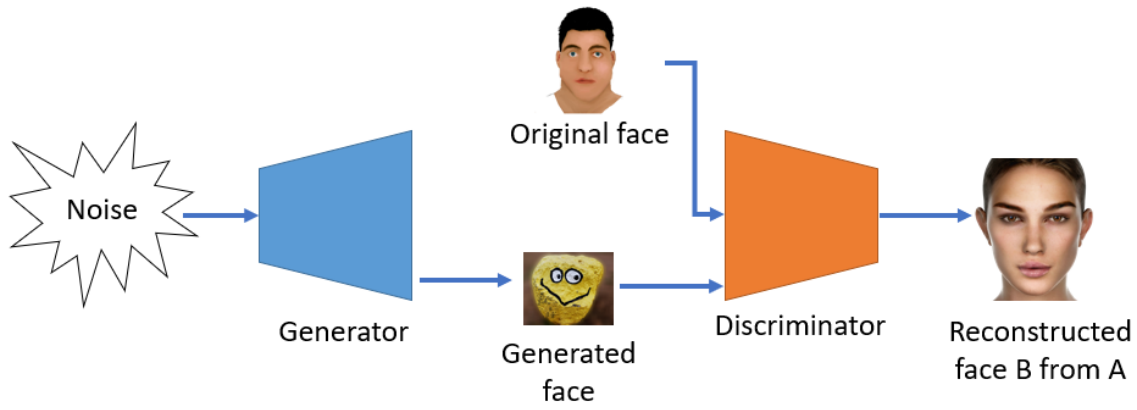


FIG. 2. GAN STRUCTURE.

## III. LITERATURE REVIEW

This section explains some researches related to deepfake methods and the methods used on them.

**Liu, M. Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., and Kautz**, J. (2019). Suggested the Few-shot UNsupervised Image-to-image Translation (FUNIT). Used a combination of GAN and a face recognition pre-trained network, they made an unsupervised model with image-to-image transformation ability [6].

**Park, T., Liu, M. Y., Wang, T. C., and Zhu, J. Y**. (2019). Instead of applying the affine transformation in the normalization layers, regularized spatially-adaptive normalization uses the spatial input. The first image semantic synthesis model's proposed normalization findings allow for realistic interior, outdoor, landscape, and city settings [10].

**Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., & Zafeiriou, S.** (2020). have developed a novel method for reconstructing faces from random pictures. They used 3D face reconstruction and trained image translation networks to assess high quality [11].

**Ha, S., Kersner, M., Kim, B., Seo, S., & Kim, D**. (2020, April). Their methodology, MarioNETte, uses an attention block and target component aligning, allowed for produced images to be injected directly into the target by the features retrieved from the target. At the same time, to further minimize identity preservation, they implemented a landmark transformer that corrected the identity mismatch without the involvement of a supervisor. It comprises a conditional generator and a discriminator, which are the building blocks of this design [12].

**Deng, Y., Yang, J., Chen, D., Wen, F., & Tong, X**. (2020). This work adopted StyleGAN [13]. In addition, they trained the generator and discriminator with novel loss functions, tweaked the original network's latent variable layer, and controlled the generator's results—Mixing VAEs of three hidden multilayer perceptron layers with GANs [14].

**Tewari, A., Elgharib, M., Bharaj, G., Bernard, F., Seidel, H. P., Perez, P., ... & Theobalt, C.** (2020). Suggested a new solution for the face rigs problem in StyleGAN [13]. They used the weights of the mentioned pre-trained model to avoid the necessity of extra training data. They achieved high-quality results by integrating a face restoration model with a discrete renderer [15].

**Li, L., Bao, J., Yang, H., Chen, D., & Wen, F**. (2019). In the first stage, facial features are extracted with the help of an encoder and an attentional denormalization layer generator that has been specially

Received 15/November/2022; Accepted 07/December/2022

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.12

designed for this purpose. Then, in the second stage, a novel Heuristic Error Acknowledging Refinement Network improves the results of face occlusions [16].

**Nirkin, Y., Keller, Y., & Hassner, T.** (2019). Face swapping is accomplished by the usage of GANs in this method. First, two different generators are used for the reenacted face and its segmentation, one for the reenacted face and its segmentation, and another for the target image face and hair segmentation. Then, another generator paints over the missing sections, and the entire reenactment and target face are blended by a blending generator [17].

**Olszewski, K., Tulyakov, S., Woodford, O., Li, H., & Luo, L**. (2019). They used symmetrical autoencoder architecture with 2D and 3D convolutional layers. In addition, a bottleneck layer connecting the encoder and the decoder, trilinear interpolation, is used to link the resampling layer to the encoded bottleneck layer, which converts the encoded bottleneck image to the target image. Thus, this network was called transformable bottleneck network TBN [18].

**Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019).** Two models create posture representations from video frames of the target subject. Using a pose detector P in the training model, It first learns the mapping G with an adversarial discriminator D to distinguish between actual and artificial sequences. Next, pose detector P to source person joints, normalization method Norm the destination person bones, pose representations created. Finally, trained mapping G is applied [19].

**Thies, J., Elgharib, M., Tewari, A., Theobalt, C., & Nießner, M**. (2020, August). They employed a 3D face model to represent facial motions to accomplish realistic facial reproduction utilizing audio inputs. In this procedure, estimating voice emotion is crucial. Person-specific expression spaces are established for each target sequence, and a latent audio-expression space is created. The audio-expression space can be mapped to the person-specific expression space to reproduce. The final image is created using a novel light neural rendering method using the projected expression and recovered audio characteristics [20].

**Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy**. 2018. Introduced ReenactGAN, a facial motion and expression transfer system that uses DL to transfer facial motion and expressions from a random individual to a target individual. Instead of transferring directly in pixel space, they first map the source face onto a latent boundary space, which would cause structural artifacts. The source face's border is then fitted to the destination face's border using a transformer. Lastly, a decoder creates the target face [21].

**Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz.** 2018. The Motion and Content decomposed GAN (MoCoGAN) model was introduced. This method generates videos by mapping random vectors to video frames and then displaying the results. The content and motion of any video are two distinct elements. Stochastic motion is used to describe motion in which the content remains unchanging. They also presented an unsupervised adversarial learning technique that would use image and video discriminators to learn motion and content deconstruction [22].

**Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro**. 2018. This paper proposes a new method for video-to-video conversion that couples a Spatio-temporal adversarial objective with the framework for generative adversarial learning by using generators and discriminators that have been adequately developed [23].

**Seyed Ali Jalalifar, Hosein Hasani, and Hamid Aghajan**. 2018. Used face synthesis to generate lip sync with input audio. In this work, the recurrent neural network has been used. To create realistic face

Received 15/November/2022; Accepted 07/December/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.12

images, conditional GANs have been used. By combining both networks, a sequence of realistic faces with audio tracks could be created [24]

**Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Perez, Christian Richardt, Michael Zollhofer, and Christian Theobalt**. 2018. A method based on a rendering-to-video translation network was developed, which turns a series of introductory animation renderings into a realistic and spatially consistent video using a network of translating nodes. This map was obtained by the use of a space-time conditioning volume formulation [25].

Table I summarizes the most recent researches that handled deepfake technologies.

TABLE I. SOME RECENT RESEARCHES ABOUT DEEPFAKE

| Author(s) | Dataset used | Method |
|---|---|---|
| [6] | Animal Faces, Birds, Flowers, Foods | Combination of GANs and face recognition model. |
| [10] | COCO-Stuff, ADE20K, ADE20K-outdoor, Cityscapes, Flickr Landscapes. | pix2pix using residual network and generator model. |
| [11] | Collection of 200 individuals | GANs |
| [12] | CelebV | GANs, Autoencoder |
| [14] | FFHQ | Mixing GANs and VAE |
| [15] | trained using the 3DMM's parameters and StyleGAN's input | GANs |
| [16] | EgoHands, GTEA Hand2K, ShapeNet | GANs |
| [17] | IJB-C | GANs |
| [18] | ShapeNet | Autoencoder with transformable bottleneck |
| [19] | Self-collected dataset | GANs |
| [20] | Videos downloaded from the internet | DeepSpeech for voice generation with light neural rendering network |
| [21] | CelebV | Autoencoder with UNet |
| [22] | TaiChi | GANs |
| [23] | Cityscapes, Apolloscape, Face video, Dance video | Conditional GANs |
| [24] | President Obama's weekly videos | Conditional GANs |
| [25] | Collected videos | Conditional GANs |

## IV. DEEPFAKE LIMITATIONS

One of the deepfake limitations is the need for a massive amount of data for training. Without training the model on enough data, the results will be unpleasant. However, at the same time, more data means more training time and more need for high computational power. Moreover, beyond quality, existing deepfake systems have flaws. For example, the content is constantly driven and generated anteriorly. So, the reenactment is static. Thus, a suitable match cannot always be made, and this method is limited in its adaptability. It is reasonable to expect that the next deepfakes would use target videos to stylize the generated data with frequent emotions and gestures, similar to current deepfakes. It will make it possible to automate the method of making realistic deepfakes to a significant degree.

A new invention has just emerged: real-time deepfakes, which have been produced and are currently in use. On the other hand, the results were not pleasant. The target's hair, teeth, tongue,

Received 15/November/2022; Accepted 07/December/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.12

shadows, and the ability to display the target's fingers while touching the face are among the other restrictions. Deepfakes, on the other hand, are already quite convincing and growing at a rapid rate.

## V. DISCUSSION

Deepfakes are not necessarily malicious in their intentions and can be used in many practical ways, such as video editing and entertainment. Nevertheless, because modern technology makes it easy to create live materials, attackers use this fact to launch attacks against innocent individuals. Individuals who are the targets of these assaults can be psychological, political, financial, and physical harm. As time progresses, we expect to see these malicious deepfakes spread across a broader range of various modalities and sectors. In this work, we investigated human reenactment and replacement deepfakes specifically. In addition, we talked through the functions of various technologies and the distinctions between their design elements in the expectation that the community will be able to use this knowledge to understand deepfakes better.

## REFERENCES

[1] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes," ACM Computing Surveys, vol. 54, no. 1. 2021. doi: 10.1145/3425780.

[2] Y. Mirsky and W. Lee, "Creation and Detection of Deepfakes: A Survey," ACM Comput. Surv. 1, 1, Article, vol. 1, no. 1, 2020.

[3] A. M. Almars, "Deepfakes Detection Techniques Using Deep Learning: A Survey," Journal of Computer and Communications, vol. 09, no. 05, 2021, doi: 10.4236/jcc.2021.95003.

[4] M. Y. Liu, X. Huang, J. Yu, T. C. Wang, and A. Mallya, "Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications," Proceedings of the IEEE, vol. 109, no. 5, 2021, doi: 10.1109/JPROC.2021.3049196.

[5] J. Song, Y. Jin, Y. Li, and C. Lang, "Learning Structural Similarity with Evolutionary-GAN: A New Face De-identification Method," 2019. doi: 10.1109/BESC48373.2019.8962993.

[6] A. Tewari et al., "High-Fidelity Monocular Face Reconstruction Based on an Unsupervised Model-Based Face Autoencoder," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, 2020, doi: 10.1109/TPAMI.2018.2876842.

[7] A. Singh, A. A. George, P. Gupta, and L. Gadhikar, "ShallowFake-Detection of Fake Videos Using Deep Learning," in Lecture Notes in Networks and Systems, 2021, vol. 175. doi: 10.1007/978-3-030-67187-7_19.

[8] T. Hwang, "Deepfakes: A Grounded Threat Assessment," Jul. 2020. doi: 10.51593/20190030.

[9] R. Chesney and D. Citron, "Deepfakes and the new disinformation war," Foreign Affairs, vol. 98, no. 1. 2019.

[10] T. Park, M. Y. Liu, T. C. Wang, and J. Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, vol. 2019-June. doi: 10.1109/CVPR.2019.00244.

[11] A. Lattas et al., "Avatarme: Realistically renderable 3d facial reconstruction 'in-The-wild,'" 2020. doi: 10.1109/CVPR42600.2020.00084.

[12] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "MarioNETte: Few-shot face reenactment preserving identity of unseen targets," 2020. doi: 10.1609/aaai.v34i07.6721.

[13] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, vol. 2019-June. doi: 10.1109/CVPR.2019.00453.

[14] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and Controllable Face Image Generation via 3D Imitative-Contrastive Learning," 2020. doi: 10.1109/CVPR42600.2020.00520.

[15] A. Tewari et al., "Stylerig: Rigging stylegan for 3d control over portrait images," 2020. doi:10.1109/CVPR42600.2020.00618.

[16] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping," Dec. 2019.

[17] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in Proceedings of the IEEE International Conference on Computer Vision, 2019, vol. 2019-October. doi: 10.1109/ICCV.2019.00728.

[18] K. Olszewski, S. Tulyakov, O. Woodford, H. Li, and L. Luo, "Transformable bottleneck networks," in Proceedings of the IEEE International Conference on Computer Vision, 2019, vol. 2019-October. doi: 10.1109/ICCV.2019.00774.

Received 15/November/2022; Accepted 07/December/2022
 ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.2.12

[19] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in Proceedings of the IEEE International Conference on Computer Vision, 2019, vol. 2019-October. doi: 10.1109/ICCV.2019.00603.

[20] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural Voice Puppetry: Audio-Driven Facial Reenactment," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2020, vol. 12361 LNCS. doi: 10.1007/978-3-030-58517-4_42.

[21] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, "ReenactGAN: Learning to Reenact Faces via Boundary Transfer," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2018, vol. 11205 LNCS. doi: 10.1007/978-3-030-01246-5_37.

[22] S. Tulyakov, M. Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing Motion and Content for Video Generation," 2018. doi: 10.1109/CVPR.2018.00165.

[23] T. C. Wang et al., "Video-to-video synthesis," in Advances in Neural Information Processing Systems, 2018, vol. 2018-December.

[24] S. A. Jalalifar, H. Hasani, and H. Aghajan, "Speech-Driven Facial Reenactment Using Conditional Generative Adversarial Networks," Mar. 2018.

[25] H. Kim et al., "Deep video portraits," ACM Transactions on Graphics, vol. 37, no. 4, 2018, doi: 10.1145/3197517.3201283.

*Received 15/November/2022; Accepted 07/December/2022*
 *ISSN (Print) 1811-9212* *ISSN (Online) 2617-3352*