International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

An Approach to Education: Improvements using Image Animation for Deepfakes

Shreeya Kokate*[1], Hriday Hegde[1], Nishad Lokhande[1], Nileema Pathak[2]

[1]*Student, Department of Information Technology, University of Mumbai, India*
[2]*Asst. Professor, Department of Information Technology, University of Mumbai, India*

Abstract - *Modernization of education does not just pertain to changes in teaching techniques but it also means to have a concrete plan to revolutionize how students learn as well as how the message goes across from the teacher to the student. It has been observed as well as studied that humans understand visual information better than textual. Educators can use deepfakes to engage and deliver lessons in a far more innovative format inplace of traditional media formats. This paper explores artificial intelligence as a tool for education. The paper introduces an interface where professors can upload their recorded videos and become an AI presenter. Once a particular student is logged in, the AI presenter can then engage with students during lessons. This AI presenter is based on the concept of Deepfakes which generates hyper realistic videos using First Order Motion Model that can boost engagement and memory retention in the virtual environment. Deepfakes is derived from the terms - "deep learning" and "fake" which uses advanced technologies to modify videos in new ways in which people appear to speak words or perform actions that didn't actually happen.*

Keywords: DeepFakes, digital education, generative adversarial networks, first order motion model, artificial intelligence technology, neural networks

1. INTRODUCTION

The need for more innovative approaches comes from the growing demand of learning in a fun way. Implementation of AI in education via the usage of deepfakes is a shot at fulfilling this growing demand. The professor and other colleagues had to opt for alternative techniques in order to create valuable virtual lessons during the coronavirus pandemic. This involves splitting longer lectures into shorter components with increased presence of media and other properties. One of the major problems faced by many institutions is to cater to good quality and affordable training. Development of more interactive and more approachable content allows professors to compete with a variety of innovation also vying for a student's undivided attention.

Historical personalities are ingrained into our education system where discoveries and inventions come into play. May it be their effect on various technologies and sections of science or their effect on different stages of human development. An approach to making education better could be that these historical personalities themselves could impart their research in a video format. In retrospect, this is a visual representation and not the actual personality, so the proposal for the title of DeepFakes in Education.

DeepFakes in its essence is the creation of synthetic media while keeping the scope of the original media. Deepfakes are the product of artificial intelligence applications that merge, combine, replace, and superimpose images and video clips to create fake videos that appear authentic (Maras & Alexandrou, 2018)[26]. JFK's resolution to end the cold war speech, which was never delivered, was recreated using his voice and speech style that might get students to learn about the issue in an innovative manner.

With the advent of AI, deepfakes can be seen penetrating to the nook and corner of the world. The in-depth concept understanding by bringing art to life provides a significant opportunity for the utilization of AI for delivering educational content. Udacity in an epigram to Soul Machines who recently developed the first digital teacher in the world, has taught more than 250,000 children about energy and has been looking at ways to automatically generate Artificial Intelligence for them - something that will change the game in academia. Professional level lecture clips require not only a veritable studio's worth of equipment, but significant resources to transfer, edit, and upload footage of each lesson, so that's why research scientists at Udacity came up with an online learning platform with over 100,000 courses and are investigating a new machine learning framework that automatically generates lecture videos from audio narration alone. [1]

This paper introduces Artificial Intelligence technology into the education system to improve its quality and reach among the students. With the help of AI, imagination capabilities can reach new heights which in turn helps in better understanding. The interface can be deployed on user-friendly platforms that can be accessed and managed easily 'by students and professors.

2. RELATED WORK

A certain person's face image can be animated to follow the facial expressions of another individual. In order to use special areas such as faces, human silhouettes, and gestures, conventional image animation and visual re-objection approaches require a strong animated object prior to them [2]. Previous research has explored how neural space-temporal networks can render video images from noise vectors [3]. More recently, the issue of conditional video output has been discussed in various approaches. X2 Face [4] uses a dense movement field to produce video output by image warping. The animation of images can also be treated as a problem of translation from one area to another. In the context of the Image-to-image Translation Process of Isolaetal, Wang et al. [5] transmitted human motion. Tulyakov et. al. [7] provided MoCoGAN in an even broader range of applications to concoct videos from noise, categorical labels or static pictures.

3. MODEL

These applications are mostly developed on principles which are based on deep learning techniques. The ability to represent complex, highly dimensional data is renowned for its deep learning. Deep auto coders, used commonly for reducing dimensionality and image compression, are one of the variants of deep networks with these capabilities. The autoencoder extracts latent functions of facial images and reconstructs facial images with the decoder. There is a requirement for two encoder-decoder pairs in which each pair is used to train image sets, in order to swap faces between source and target images, and the parameters of the encoder are split between two network pairs[22]. In other words, two pairs have the same encoder network[22]. This technique helps the common coder, which typically has the same characteristics as eyes, nose and mouth, to recognise and learn similarities between two sets of facial images, which is relatively challenging[22].

This approach is applied in several works such as DeepFaceLab, DFaker, DeepFaketf (tensorflow-based deepfakes)[22]. In reference to Figure 1 two Networks use the same encoder but different decoders for the training process (top)[22]. An image of face A is encoded with the common encoder and decoder with decoder B to create a deepfake (bottom)[22]. By adding adversarial loss and perceptual loss implemented in VGGFace to the encoder-decoder architecture, an improved version of deepfakes based on the generative adver-sarial network (GAN), was proposed in[22]. The perceptual loss of VGGFace enhances eye movements and lets you smooth out artifacts in segmentation masks leading to better quality performance images[22]. The facial recognition can be made more stable and facial alignment more accurate by using the multi-task convolutional neural network (CNN) from the FaceNet implementation[22].

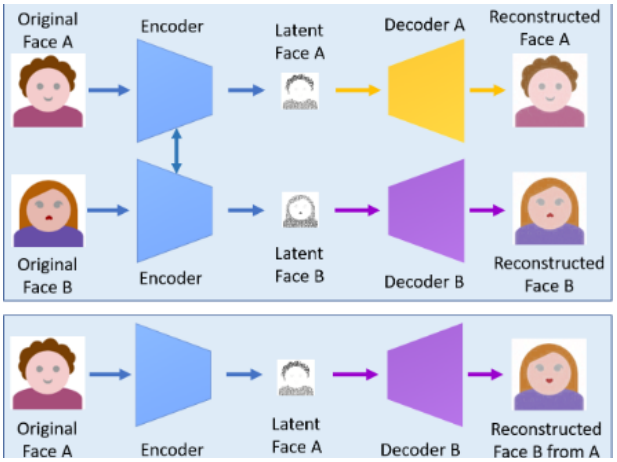


Figure 1: Creation of DeepFakes [22]

3.1 FIRST ORDER MOTION MODEL

Here, figures are animated in a source image S based on motion in a video having similar figures. When the figures are in motion, two videos cannot be used, thus, a self-supervised method inspired by Monkey-Net was used. We use video datasets of the same figure category. This data set is used to create new videos which are a combination of single frames and learned feature motion representations from the original data. It encodes motion by observing the frames, a mixture of motion-specific keypoint displacements and local affine transformations that take account of collinearity and distance ratio. Two modules, namely the motion evaluation module and the model image generation form this structure. The motion measurement module plays the part of estimating the dense motion field between the video frame to the source frame. The motion field is modeled by a function in backward optical flow in which the driving video's each pixel location is mapped with its corresponding source frame location.

For comparison, the paper assumes that the abstract frame estimates two transformations independently: from a reference frame to the source and the driving video. During the testing phase, the model receives the source image pairs and the driving frames, which can be very different visually, from another film. The model that estimates motion works in two stages instead of predicting directly. In the initial step, both transformations are approximated from sets of sparse trajectories, accomplished using self-supervised key points. The key points of a driving video and source frame are independently predicted by an encoder-decoder network. The keyboard display results in a portable motion display. The movement in the close proximity of each keypoint is modeled on local affinity transformations. In comparison with only key point displacements, the local affine transformation models an entire series of

transformations. A number of keypoint positions and affine transformations are used for Taylor expansion representation. Now the network key detector outputs key point locations together with the parameters of every refined change.

During the second step, a dense motion network combines the local approximations to obtain the resulting dense motion field [2]. Furthermore, in addition to the dense motion field, this network outputs an occlusion mask that indicates which image parts of D can be reconstructed by warping of the source image and which parts should be inpainted, i.e. inferred from the context[2]. The generation module finally makes a picture of the original object to move as seen in the driving video. Here, we are using a network of generators which tweaks the source image and paints the image area in the source image.

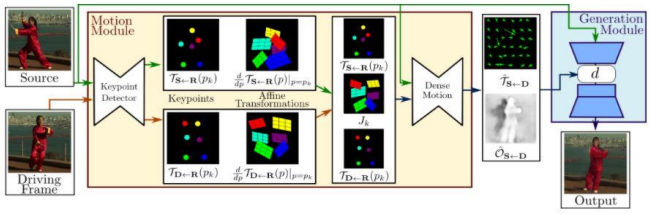


Figure 2: First Order Motion Model [23]

In Local Affine Transformation, the backward optical flow from a driving frame to the source frame is estimated by the motion estimation module. It approximates Taylor's first expansion in a keypoint neighborhood. Assuming an abstract frame of reference, it estimates each transformation of the learned keypoints in the neighborhood with a given frame. Formally, consider Taylor's first-order expansions in some key points as a result of a transformation. The keypoint predictor network also generates four more channels for each keypoint after measuring a first order of the Taylor extension for both the source and driving frames.The matrix coefficients come from these channels and the corresponding confidence map is calculated by means of the weighted average spatial value.

While creating combinations of local motions, a convolution network estimates in the main points and the original source image from a variety of Taylor approximations. The local patterns are pixel-to-pixel, linked with drive video and not the source image, since each video's pixel position is mapped to the appropriate location in the source image. It is difficult for the network to predict due to the misalignment problem. To provide inputs already, the source frame is twisted based on estimated local transformations. We therefore get transformed images that are aligned in a keypoint neighbourhood.

For occlusion-aware image generation, the source image is not pixel-to-pixel aligned with the image to be generated[2]. A warping technique is used to deal with this misalignment. The occluded areas cannot be retrieved by image-warping[2]. The occlusion map is used to mask the areas to be painted in the feature map. The mask reduces the impact of the characteristics that fit the occlusive pieces. By adding a channel to the final layer of the dense motion network, the occlusion mask of a sparse keypoint representation is calculated[2]. Finally, the transformed feature map is sent to the next generation module's network layers to make the image sought.

With many losses, the system is end-to-end trained. The reconstruction loss involves the channel feature extracted from the particular layer of VGG-19 alongside the input drive frame and the corresponding reconstructed frame, and the number of feature channels in this layer are denoted by I. In a few resolutions that are used, a pyramid is built from a down-sample.

To impose equivariance constraint, there is no key point annotation during training needed in the keypoint predictor. This may lead to unstable performance. Equivariance constraint is one of the most important factors driving the discovery of unsupervised keypoints[2]. It forces the model to predict consistent keypoints with respect to known geometric transformations[2]. Since the motion estimator not only predicts the keypoints, but also the Jacobians, it is calculated that the well-known equivalence loss involves constraints on Jacobians. Then losses are identical to the constraints of the keypoint position. In all experiments, equal loss weights are used. The aim is animation of the figure in the source frame by use of the driving video. Per frame is processed independently. The relative motion between driving videos to the source frame is transferred, rather than transfer the motion encoded to the source frame.

3.2 GENERATIVE ADVERSARIAL NETWORKS

GAN is made up of two neural network modules, namely, the generator(G) and the discriminator(D). They both have an adverse role in which the generator tries by generating identical data to deceive the discriminator. The aim of discriminator is to ensure that the false information from real data is not misled. They both learn and train complex information such as audio, video or image files simultaneously.

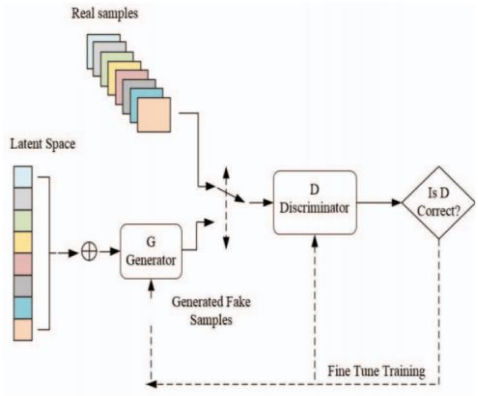


Figure 3: Working of GANs [21]

The generator model produces random noise images and then studies how realistic images are produced. The input random noise is uniformly or usually sampled and fed into an image generator. False images and real images from the trained collection are fed into the discriminator, which learns how to discern false images from actual pictures. The output is the likelihood of actual input. If the input is real, the output is 1 and the output is 0 when generated.

Training data will obtain the probability of the specified sample. If the generator's output value is high then the discriminator assumes that the generator is nothing but training data, so 1-D(G(z)) becomes extremely low, and needs to be reduced to a minimum. For the Discriminator, we want to maximize that, so the optimal state of the discriminator is calculated accordingly. The training of the generator is done in a way that it produces the results for the discriminator so that it cannot discriminate between the image and training data. In this case, an attempt is made by the discriminator to maximize the target while an attempt is made by the generator to minimize it. As a result of maximizing or minimizing, the minimax term is obtained. Together they both learn by alternating descent.

4. PROPOSED WORK

The proposed system aims to accomplish a life-like video replica of historical figures that the professors can easily use to impart certain subjects or topics proposed by those very same historical personalities. These historical personalities don't get to bring their words to the globe today, but due to modern technologies and innovative techniques, we will hear it now. Thus, we aim to make it possible to make new videos of historical figures, where they themselves share their achievements. It is possible to bring historical figures back to life and to create more interactive classes for schools. Some similar techniques already exist, but deepfakes can take it to a subsequent level. Absence of an innovative and interactive way of teaching students online. Students require a visual format of teaching to understand concepts clearly. The project proposes a way for having a visual format of presenting a topic by the scholar/inventor/famous personality themselves which is attained by using a first order motion model for generating these synthetic videos and an interactive interface for the professors to access and use the technology.

4.1 DATASET

The Dataset used for the process is KomNET: Face Image Dataset from Various Media for Face Recognition [24]. It is a precompiled dataset of various angled images which can deeply describe the human face. The different faces used were to bring in diversity to the sample and to have higher accuracy and quality.

All the training images used must be of the same dimensions, for eg. 224x224. The driving video to be used must be of dimensions 256x256. Both training images and driving video should be focused on the face for better output quality. The resizing of driving video is handled internally via the use of imageio-ffmpeg library in Python.

4.2 TRAINING

For training the model, a dataset of face focused images was used. The images were composed of many angles of the face. The quality of the output was dependent on many factors which consisted of original image quality, the number of images used for training the model and also the format of the input source image. When PNG image format was used for the source image, the quality was more composed as compared to when JPG image format was used. This accounted to the fact that PNG image inputs would generate almost lossless outputs. The quality of image animation also improved on custom training of the model. The speed of processing and generation was unaltered and was mainly dependent on the system's hardware used in the system. All training and tests were performed on a standard i7 9th generation processor with Nvidia GeForce GTX 1660Ti laptop variant graphics processor with 16 GB RAM. These tests were also done on a system which consisted of an Nvidia GTX 1650 graphics processor with an i5 9th generation processor with 8 GB RAM. The differences were not significant in terms of quality but more on the aspect of time.

4.3 AUDIO

We use a library named MoviePy to extract audio from the original video which is then merged with the generated video having the appearance of the character required. MoviePy is a Python library that is used for audio- video editing, trimming, concatenations, title insertions, video compositing, video processing, and creation of custom effects[25]. MoviePy on its own can deal with various forms of video and audio formats and can be mildly used as a video editor.

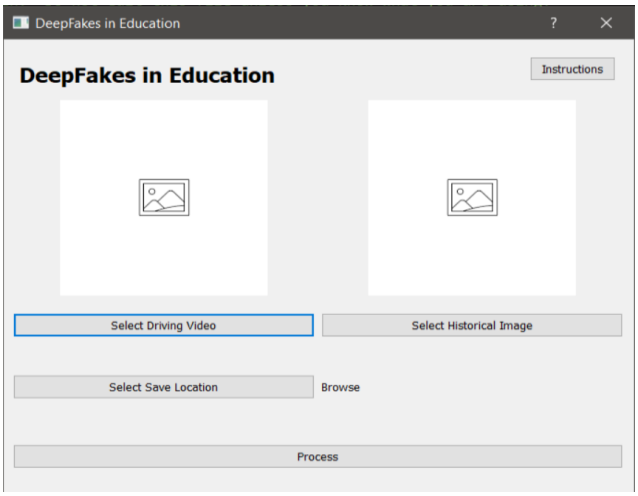
5. INTERFACE DESIGN

The interface is entirely built using the Python framework. PyQt is the programming language GUI module for Python. PyQt is big, native to all the main platforms and probably has the largest community. PyQt

implements Qt library, the most common library. This module provides a unique and easy opportunity to build applications. PyQt has its essential window building blocks known as widgets. A window is a widget, a grid, an image, an icon, a button or a customized widget can also be defined. Buttons are essential aspects called QPushButton in PyQT modules. In addition to adding a widget, the code is the same as building the app or running the loop. If a button is to be created, the text that the button will have should also be specified. A QApplication is initiated followed by a window using the most basic type of QWidget as it will act as a container without any special behavior. Next, the layout is introduced and QPushButtons can be added to it. End with calls to show() and app.exec_() for showing the application. With the help of PyQt we can build interactive graphical user interface (GUI). This GUI will allow the professor to upload their own video along with an image of the figure or any historic personality for generating the new video. The upload and download formats that will be used for this interface will mp4. The user interface will consist of simple blocks which will help the professors to use the system efficiently. Blocks will include buttons to upload images and videos, record videos, and lastly to generate new videos with trained images.

6. RESULTS

Following are the screenshots of the interface and output of the proposed system.



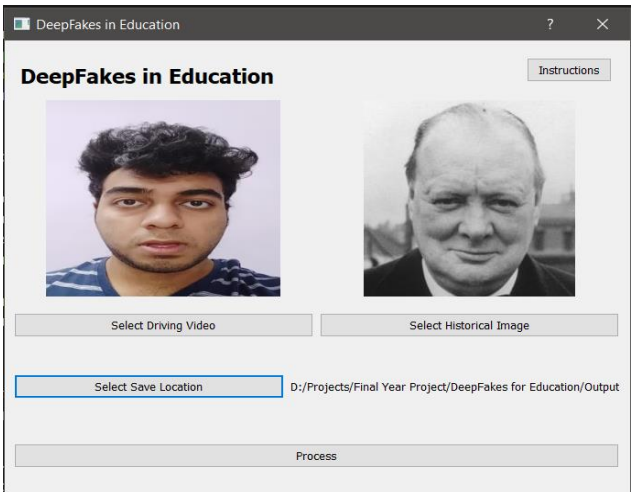


Figure 4: Interface of the system

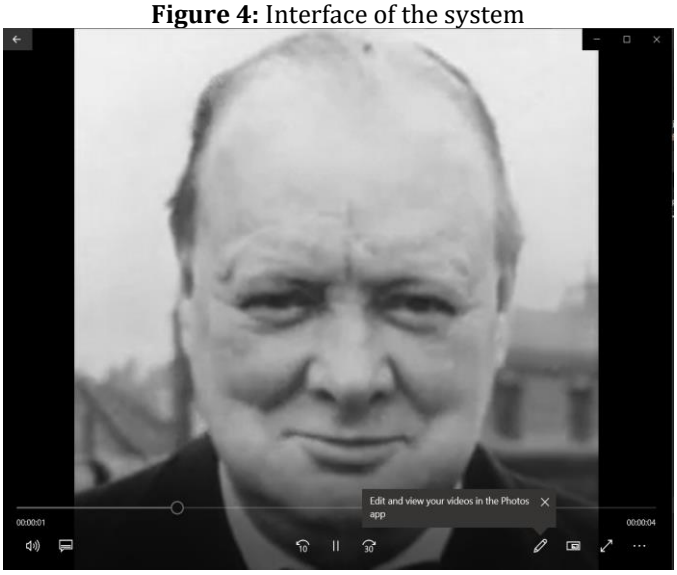


Figure 5: Final video output

7. CONCLUSION

As technology progresses and societies adjust to the concept of deepfakes, most of the use cases of deepfakes will be to improve lives and empower institutions. By increasing the degree of understanding and interest in learning among students due to use of technology, the results produced by this method will show promising outcomes in the domain of education. Students will have an innovative approach to learning using this method, and teachers can have a better way to engage students in online learning. Having historic figures and scholars teach their subjects themselves would help inculcate increased curiosity and fun in learning those subjects online. Subjects like History, Mathematics, Science which take a lot of visualization to understand can be made easy using such generated videos using this system. Such videos can be used during classes as well as online presentations. The easy-to-use UI of this project makes it easy for teachers and online educators to generate video lectures using

faces of scholarly people quite easily. This project can replicate storytelling experience in the actual learning process which will benefit students to a much greater extent.

This paper is based on experimental technology and the contents are to be used only for research or educational purposes. Research on technology that can detect misuse of Deepfakes is being done by many individuals and organisations. A few references to such researches are as follows:

DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection[27], FaceForensics++: Learning to Detect Manipulated Facial Images[28], Detecting Photoshopped Faces by Scripting Photoshop by UC Berkeley and Adobe[29].

8. ACKNOWLEDGEMENT

We have great pleasure in presenting the report on "An Approach to Education: Improvements using Image Animation for Deepfakes". We take this opportunity to express our sincere thanks towards our guide Prof. Nileema Pathak for providing the technical guidelines and the suggestions regarding the line of this work. We would like to express gratitude towards her constant encouragement, support and guidance throughout the development of this project.

9. REFERENCES

[1] https://www.fanaticalfuturist.com/2019/08/edtech-company-udacity-uses-deepfake-tech-to-create-educational-videos-automatically/

[2] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci and Nicu Sebe. First Order Motion Model for Image Animation. In NeurIPS 2019, Vancouver, Canada

[3] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In NIPS, 2016, 26-Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In ICCV, 2017.

[4] Olivia Wiles, A. Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In ECCV, 2018.

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro.Video-to-video synthesis. In NIPS, 2018.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[8] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In CVPR, 2018.

[9] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In ICLR, 2017. Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In CVPR, 2018.

[10] Liang Gonog and Yimin Zhou, "A Review: Generative Adversarial Networks", arXiv, 2020

[11] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In ECCV, 2018.

[12] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In SIGGRAPH, 1999.

[13] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In ICCV, 2017

[14] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. TOG, 2014.

[15] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In CVPR, 2017

[16] Ian J. Goodfellow, Tim Salimans, Alec Radford, Xi Chen, Vicki Cheung, Wojciech Zaremba, "Improved Techniques for training GANs", arXiv, 2016

[17] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In CVPR, 2018.

[18] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In NIPS, 2016

[19] henglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In CVPR, 2019.

[20] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel,Aaron Courville, Yoshua Bengio, "A Recurrent Latent Variable Modelfor Sequential Data", arXiv, 2016

[21] https://blog.statsbot.co/generative-adversarial-networks-gans-engine-and-applications-f96291965b47

[22] Thanh Thi Nguyen, Cuong M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Saeid Nahavandi, "Deep Learning for Deepfakes Creation and Detection: A Survey", arXiv, 2019

[23] https://github.com/AliaksandrSiarohin/first-order-model

[24] https://www.sciencedirect.com/science/article/pii/S2352340920305710

[25] https://pypi.org/project/moviepy/
[26] MikaWesterlund, "The Emergence of Deepfake Technology:A Review", Technology Management Innovation Review, November 2019
[27] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales and Javier Ortega, "DeepFakes and Beyond: A Survey ofFace Manipulation and Fake Detection", arXiv:2001.00179v3 [cs.CV] 18 Jun 2020.
[28] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies and Matthias Nießner, "FaceForensics++:Learning to Detect Manipulated Facial Images", arXiv:1901.08971v3 [cs.CV] 26 Aug 2019.
[29] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, Alexei A. Efros, "Detecting Photoshopped Faces by Scripting Photoshop" , arXiv:1906.05856.