Artificial intelligence art

Ioana Cheres
Department of Computer Science,
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
cheresioana@gmail.com

***Abstract*—Deepfake is creating realistic forgeries that make people unable to distinguish reality from fiction, while filtering algorithms distort the perception by amplifying the interests of the user. From the way we perceive society to the way we perceive ourselves, the virtual environment is changing fundamental mechanisms of life. Due to the rapid growth of the virtual world and its hazy moral context, our society needs to build the groundwork of an ethical code for the online environment. Powerful technologies are used to produce art that captivates and impresses the general public. The artworks highlight ethical problems that our society is facing, creating a thoughtful space of analysis, experience and debate around the moral concerns in the technological era. This paper contains a short analysis on some of the main threats of artificial intelligence along with the proposal of art as a way to approach technology ethics on a large scale.**

***Index Terms*—AI, ethics, art with AI**

I. Introduction

A complete and valid perspective, gathered from a multitude of different subjective information sources, is the perfect context in which a person can make decisions. The growth of social platforms made them slowly become the main source of information for a large percentage of people [1]. The information users receive is being filtered by artificial intelligence algorithms that rank each post based on likes, comments, and time spent watching in order to ensure comfort and deliver "only" desirable content. This creates well-separated and polarized communities where the users perspective is narrowed to its beliefs, amplifying the factionalism of the modern world [2].

The pleasurable posts the viewers see are not only filtered but they can also be forgeries. Deepfake has become a popular technology which allows the creation of fake media on a targeted person. Creating deepfake content is so easy that anyone, with no technical skills, can use it on public websites [3], [4]. This technology has already started to be weaponized to harass persons by creating deepfake pornograpy [5]. Being a victim of deepfake can be a serious problem and due to the noncentralised nature of the internet the fake content may never disappear, irreparably damaging the public image and mental health of the victim.

The distortion of information does not only affect the perception about society but also change the opinion and standards about the self. The way humans measure their success is by comparing with others [6]. Seeing society through the veil of social media affects self-esteem because people compare themselves with virtual profiles [4]. Artificial intelligence algorithms can remove imperfections, lighten the color of the skin, and modify body proportions, features that humans do use to display their perfect image. There is no surprise that photo editing behaviour is very popular [7] making people feel like everyone around has the perfect look. This is the cause of increased dissatisfaction of people with their own body and is affecting persons from all ages [2].

The artificial intelligence advancements impact society as a whole, not only specialists who understand and develop these technologies. Our society needs a way of debating meaningful ethical problems on a large scale. We argue that the latest forms of art created using AI technology and promoted by organizations like ArsElectronica or NestaItalia, are a solution to involve the population in the analysis of complex ethical issues raised by technology advancements. My work is an interactive art installation called "The profile" and a presentation of it can be found online [8].

Through my project users with different backgrounds can experience the power of some of the most advanced technologies. Once the user enters the installation he or she experiences ethical concerns regarding the impact of the latest technologies in a metaphoric approach. The project is structured in three parts that match the steps of creating an online identity.

First the user explores a virtual environment where he projects ideas in the physical world using his own body. He is becoming an active cog in the mechanisms of spreading fake, aggressive and sexualised content, being faced with some of the major social media trends that influence society. The content is structured in three circles of hell inspired by Dante's Divine Comedy: limbo, lust and anger [9].

The second part is agreeing the terms and conditions, allowing the software to use private data and narrow the users perspective, which is the modern Faustian bargain. The Faustian pact is essentially the renunciation of any form of self-determination, decision, free will and the transfer of decision-making authority to an external entity. It is a comfortable position, in which the truths are clear and come ready, nullifying the need for discernment. Today we witness the modern Faustian bargain: the information people receive is almost always in concordance with their beliefs, everything else being filtered by algorithms.

The third part is exploring the profile in which the viewer

is confronted with a deepfake of his own body. The user can see itself like in a mirror but the software takes control over the image at random intervals, increasing the confusion between what is real and what is generated. Instead of an emotionally distant written article, the user experiences first hand the impact of private data manipulation being, for a short period of time, a victim of deepfake.

II. Related work

Generating artificial images with neural networks is a challenging task. There are three main directions based on the input of artificial network: a random vector (noise) [10], a text description [11] or another image [12]. The base architecture for neural network image generation are conditional adversarial neural networks (GAN) [10]. GAN architecture contains two neural networks called the generator and the discriminator that play a minmax game. The generator takes as input a noise vector and has to output an image, while the Discriminator takes as input an image and has to output the probability of that image being a forgery (generated by the Generator).

The input data shape and content influences drastically the performance of the neural network. The input data represented by a text can be changed into a multidimensional space using a story encoder [11] or it can be another observable image [12].

The structure of the GAN can also be modeled according to the needs. In [11] are used two discriminators, one specialised on the detection of story inconsistency and one on the images realism. Another approach is to use a markovian discriminator that classifies each $N \times N$ patch of the image and outputs the final result as an average of all the patches [12] [13].

A popular motion transfer approach is the latest framework called "Everybody Dance Now" [14] in which a fake video of the subject is generated where he or she is dancing like a professional dancer starting from the OpenPose [15] skeleton representation. For the realism of the images this project uses separate GANs for face and body in order to maximize the human aspect of recognizing identity: increasing the realism of the face. Recent work in video to video synthesis [16] offer image translation based on a general abstraction of human images obtained from sketch filters on the original subject.

To create videos that automatically generate content of a person mouthing words using existing footage has been done with Video Rewrite [17] by mapping facial keypoints between the subjects. The project focuses on facial expressions and mouth position and not on full body position transfer. MoCoGAN [18] uses MUG Facial Expression Database [19] to map expressions on target faces and extended this capability to full body image retargeting using a noise vector as input.

Form an artistic perspective, the artificial intelligence is compared with the invention of applied pigments [20]. Algorithmic art, which is a term used to describe any art that can not be created without the use of programming, captures world wide headlines and sells at auction at considerable prices [21]. Important art exhibition like Athens Biennial host artificial intelligence artworks like "Seamless" by Theo Triantafyllidis, and there are even dedicated art museum and exhibitions only for artificial intelligence [21]. The AICAN project studies the artistic creative process and the evolution from perception to cognitive opinions in the context of art generated by programs [22]. It uses a creative adversarial network (CAN) that is a variation of GAN with stylistic ambiguity used for improving the complexity and novelty of the resulting images [23]. More than that, Google has launched Deep Dream Generator, a set of AI tools for creative visual content generation.

My work is made possible by recent rapid advances in the realistic image-to-image translation, pose estimators and art experiments. Modern pose detection using the library OpenPose [15] and image-to-image translation using pix2pix neural networks [12] are the technical building blocks on top of which I extend my project. The rising of artificial intelligence in visual art has open a new world of possibilities in exploring and creating meaningful and complex artworks. The decomposition of the frame along with the image simplification, real time rendering and the ethical dimension of the project are the most important components that extend the current state of the art and offer a new perspective in this field.

III. System overview

The art installation has four important components: physical devices and their relative location, the distributed architecture, the local component for user interaction and the remote component for heavy computation.

A. The physical components

The location of the physical devices is relevant for the quality of the images taken. The system has a limited amount of time to generate a realistic image, so the position of the elements ensure a simplified image with no background and shadows. The physical components of the installation are: the webcam, the screen, the background and two lights. In Figure

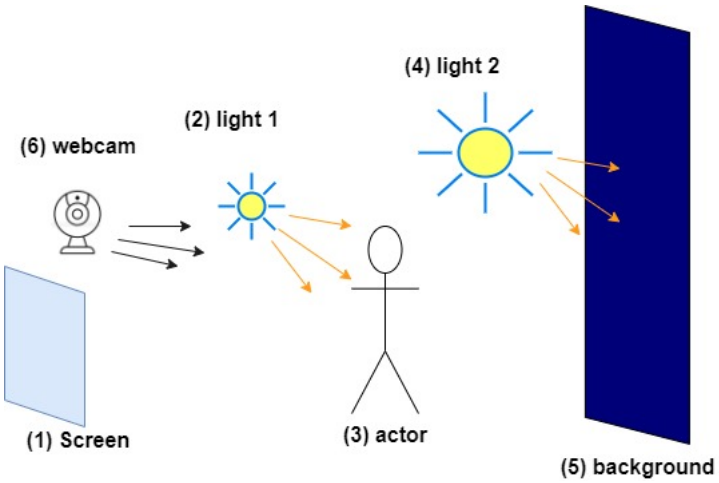

Fig. 1. Physical components of the installation

1 can be seen the position of the physical components of the installation. The webcam is the main source of information for the system that captures frames of the subject. The background

needs to be in one color for the program to crop easy the person out of the frames. There is an important relation between the two sources of light: light 2, that is between the user and the background has to be three times stronger than light 1 which is between the user and the screen. This localisation of the lights makes the user not to cast any shadow on the background that could complicate the structure of the frames by adding a shadow element. It also adds a frontal light on the subject for diminishing the shadows on the body while also allowing details like the texture of the clothes and skin to be captured accurately on the frames.

B. The local software component

The local component parses each frame and computes the image displayed on the screen. In the first two circles the user explores the platform and interacts with the environment using its skeleton which represents his virtual ego.

Each frame taken by the webcam is decomposed in three separate images like showed in Figure 2: skeleton, background and person.

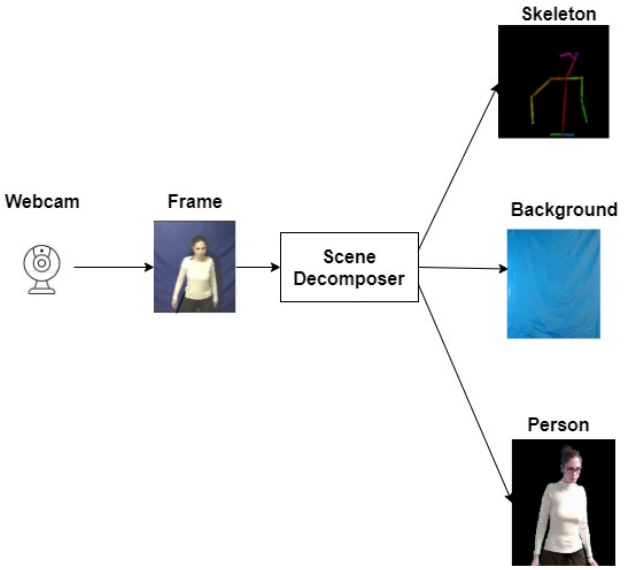

Fig. 2. Scene decomposer

The skeleton is generated using OpenPose library [15] and has as input the current frame. The background is known from before the user enters the installation and, due to the position and intensity of the lights, it can be extracted from the main frame to obtain the person only. Once all the secondary images are computed, the skeleton and person image are saved for further training the generative adversarial neural network in the third part.

Scene Composer component detailed in Figure 3 takes all the partial images computed. It creates the frame that will be displayed on the screen. From the person image a mask is computed, that acts as a display area for the scene components, giving the illusion of projecting through the body. The part of the body that is not over any components is replaced by the skeleton, and the ring of hell is added

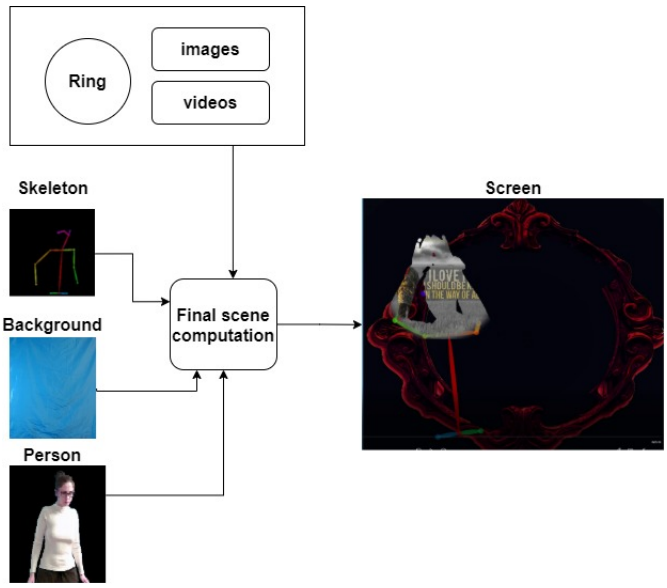

Fig. 3. Scene composer

behind the visible components and the skeleton. After the main elements are computed, the background image is added. All the elements are treated as layers and the operations on final frame composition are done using OpenCv library [24].

When the subject has to pass from one level to another he has to place his hands inside two circles. The Euclidian distance between the center of the circle and hands joints is computed and, if it is smaller than a value ϵ, the user passes in the next level.

While exploring the first two parts the system is saving tuples of images of the person and the matching skeleton. Once enough images are taken, it sends them to the remote component for the neural network computation. When the users finishes exploring the first two parts of the installation, the system downloads the weights of the remote trained generative adversarial neural network.

In the third part the system captures the images of the person and computes the skeleton, and then it generates the fake body. The user can see himself on the screen, in the exact position that he is standing, but the body he has is generated by a neural network. At random time intervals the system takes control over the image of the user. It alters the joint position of the skeleton and feeds forward the neural network with the computed skeleton, resulting in an image of the subject in a different position than it currently has. The program has specific algorithms that modify the joints in order for the image from the screen to wave to the subject, dance and move from side to side. In order to have a big psychological impact, the system computes the sequence of independent moves to start and end in the current position of the user.

C. The remote software component

The remote software component is designed for heavy computation. It has the objective to train as fast as possible

the pix2pix [12] generative adversarial neural network.

The neural network is composed of two multilayer perceptron called the Generator and the Discriminator. The Generator has to learn to compute realistic images of the body having as input the skeleton from Openpose and the Discriminator has to learn to predict which images are real and which are computed by the Generator as showed in Figure 4.

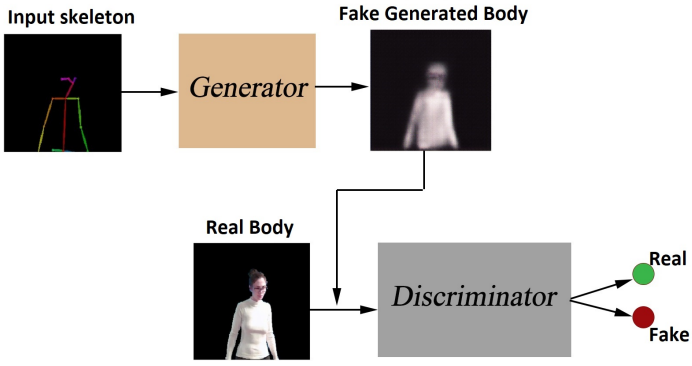

Fig. 4. Generative Adversarial Neural Network

The structure of the GAN is following pix2pix architecture presented in [12]. The Generator has eight layers of encoding, seven layers of decoding and skip connections. The Discriminator is a PatchGAN that classifies a 70x70 patch of the input image.

The data is resized to match the size of the input layer, random flipped in order to increase the generalization capabilities and then normalized.

D. The distributed architecture

In the first two parts the user is interacting with the platform and the system captures images of him. In the third part the user sees on the screen a deepfake of his own body. Due to the complex capabilities needed, the system has to be distributed from a hardware perspective. The systems needs to be moved easily between art exhibition and it may work in deficient environments like dust and moisture but it also has to have a high computing power in order to calculate the weights of the generator in less than twelve minutes.

The main component is at the location of the user and is a part of the installation. It has the purpose of interacting with the subject in all three parts and capturing the data. The local hardware has reduced capabilities and performance, but can be moved easily and is more robust and enduring to external factors.

When enough images of the subject are taken, they are sent to the remote component which is located in a fixed place and has a Linux based server. This component, once it receives the data, trains the generative adversarial neural network needed for the fake body generation and then sends the weights back to the main component. The remote component's hardware has an increased performance but can not be moved or held in deficient environment conditions. Because of this limitations, the remote server can not be physically at the location of the installation, but its capabilities can be exploit by using a distributed architecture of the software.

The main component uses a library called Paramiko [25] that implements SSH protocol to securely send and receive data. It sends UNIX commands over SSH and starts on the remote component an Nvidia docker container [26] which has the python code that trains the neural network. The architecture of the system as a whole is a Master-Slave architecture.The local component is the master, knowing the state and location of the remote component and giving it the commands it needs. The remote component is the slave and it is unaware of the existence of the master, having the sole purpose to compute the weights of the neural network.

E. The artistic and psychological dimension

Form an artistic perspective the project is a unique combination of the latest technology with elements of classic European literature. The quote from Faust [27] and the reinterpreted rings of hell from The Divine Comedy [9] are the guiding elements of the user in his experience, marking the passing from one level to another. They also represent principles and ideas that the subject is already familiar with and can associate with the new trends from social media presented in the installation.

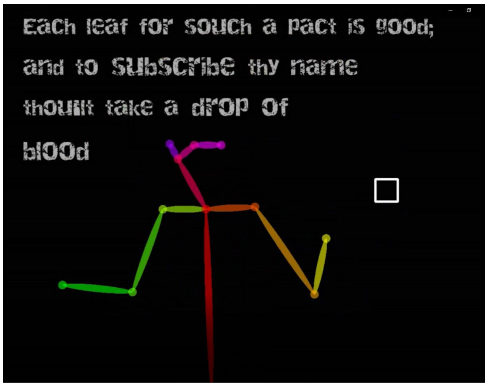

Fig. 5. Faust scene

In the first part the user is discovering the three circles of hell inspired by Dante's work [9]. Each circle has a different chromatic scheme that matches the information presented in order to amplify the psychological impact and create a coherent environment for the experience. The Faustian bargain has to be signed in the modern way, using a checkbox, and this scene is the only one with a complete dark background, in order to suggest the solemnity of the moment and to contrast drastically all the other elements as showed in Figure 5.

The third part is focused on the psychological impact of deepfakes. In an online era psychopathy is positively associated with engagement in a higher frequency of trolling behaviours due to the reduced capacity of experiencing empathy [5]. The subject is becoming, for a short period of time, a victim of deepfake in order to understand the potential impact of this technology and to start questioning the authenticity of the contents of online videos.

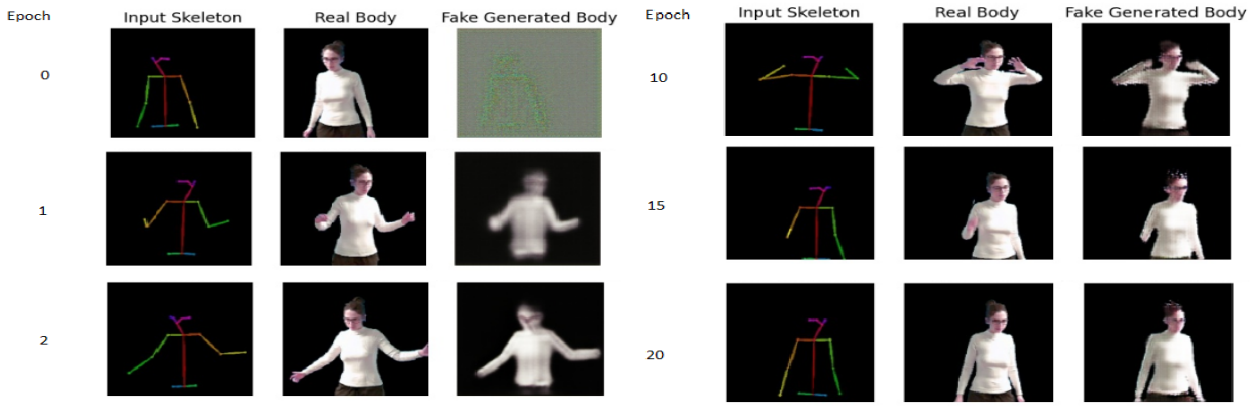

Fig. 6. Evolution of the predicted images on epochs

IV. Experiments and results

A. Setup

The results described are obtained using as a remote component a Linux server with one TeslaV100 GPU and 700GB RAM. The local component runs on Windows and has a GTX 1660, 6GB RAM and i9 processor. The webcam used has a resolution of 1920 * 1080 pixels and 30 frames per second.

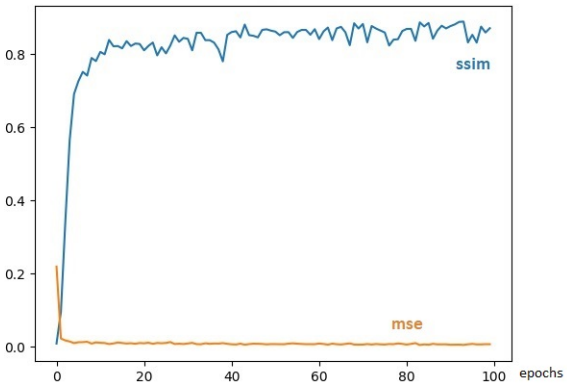

Fig. 7. Generator evaluation

B. Performance evaluation

In order to test the performance of the GAN I used as training data the images captured and processed by the system when exploring the first two parts of the art installation. In total the system captured 745 tuples of skeleton and body image that were used for training the GAN.

For evaluating the GAN I used two metrics: Mean square Error (MSE) and Structure similarity index (SSIM) [12], [14]. In Figure 7 can be seen the graph on how the two metrics evolved in 100 epochs. As it can be seen from the graph the neural network learns very quickly to generate a realistic body of the subject.

The MSE is decreasing abruptly and then constantly following a straight line, meaning that on pixel to pixel differences the images have a high quality. When looking at the SSIM metrics, the graph shows the partial instability of the generated data. On average after 20 epochs the SSIM is larger than 0.8 which is a good result but the wobble of the graph highlights the partial nondeterministic nature of the neural networks. SSIM is a more accurately metric when dealing with image quality because it takes in consideration the structural similarity of the images, not only pixel differences like the MSE.

In Figure 6 can be seen the results obtained in the first 20 epochs which are the optimum number for training the GAN. The neural network learns fast, and even after one pass through the data it can output the general shape of a human. As early as the 10th epoch shadows start to appear and already in the 20th epoch can be seen a clear distinction between the texture of the clothes, skin and hair.

In total, the system trains on the data in 4 minutes and 8 seconds. The time of sending the images is 2 minutes and the time of receiving the weights is 3 minutes. The system manages to train the remote component in under 12 minutes, which was one of the objectives of this project that make it usable by the general public.

V. Conclusion

The system presented is a complex artificial intelligence art installation that combines powerful technologies in order to give the viewers a meaningful experience. Its performance of creating a deepfake of the users body in under twelve minutes along with the complex scene construction make it appealing for the public. The project offers a new perspective on how to approach technology ethics on a large scale by involving people in the exploration of artificial intelligence.

My work can be extended by artists in the audio direction and enhance, perfect and reinterpret the scene components in the visual direction. Also, on the technical side, the distributed

architecture can be improved for a faster data transfer and the GAN can be perfected to support higher resolution images.

REFERENCES

[1] E. Shearer and A. Mitchell, "News use across social media platforms in 2020," 2021.

[2] P. Van den Berg, S. J. Paxton, H. Keery, M. Wall, J. Guo, and D. Neumark-Sztainer, "Body dissatisfaction and body comparison with media images in males and females," *Body image*, vol. 4, no. 3, pp. 257–268, 2007.

[3] "online deepfake generator." [Online]. Available: https://deepfakesweb.com/

[4] "Online ex2 deepfake generator." [Online]. Available: https://mmasked.com/

[5] D. Fido, J. Rao, and C. A. Harper, "Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography," 2020.

[6] L. Festinger, "A theory of social comparison processes," *Human relations*, vol. 7, no. 2, pp. 117–140, 1954.

[7] M. Lee and H.-H. Lee, "Social media photo activity, internalization, appearance comparison, and body satisfaction: The moderating role of photo-editing behavior," *Computers in Human Behavior*, vol. 114, p. 106579, 2021.

[8] "Presentation link." [Online]. Available: https://storage.rcs-rds.ro/links/c654e523-9ea9-4f62-abcf-884d08bb2ece

[9] D. Alighieri, *Divine Comedy*, 1472, vol. 1.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv preprint arXiv:1406.2661*, 2014.

[11] Y. Li, Z. Gan, Y. Shen, J. Liu, Y. Cheng, Y. Wu, L. Carin, D. Carlson, and J. Gao, "Storygan: A sequential conditional gan for story visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6329–6338.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," 2018.

[14] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5933–5942.

[15] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[17] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 353–360.

[18] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.

[19] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, 2010, pp. 1–4.

[20] B. Agüera y Arcas, "Art in the age of machine intelligence," in *Arts*, vol. 6, no. 4. Multidisciplinary Digital Publishing Institute, 2017, p. 18.

[21] T. Schneider and N. Rea, "Has artificial intelligence given us the next great art movement? experts say slow down, the 'field is in its infancy.'," *Artnet News*, 2018.

[22] M. Mazzone and A. Elgammal, "Art, creativity, and the potential of artificial intelligence," in *Arts*, vol. 8, no. 1. Multidisciplinary Digital Publishing Institute, 2019, p. 26.

[23] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone, "Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms," *arXiv preprint arXiv:1706.07068*, 2017.

[24] A. Mordvintsev and K. Abid, "Opencv-python tutorials documentation," *Obtenido de https://media. readthedocs. org/pdf/opencv-python-tutroals/latest/opencv-python-tutroals. pdf*, 2014.

[25] M. Zadka, "Paramiko," in *DevOps in Python*. Springer, 2019, pp. 111–119.

[26] D. Kang, T. J. Jun, D. Kim, J. Kim, and D. Kim, "Convgpu: Gpu management middleware in container based virtualized environment," in *2017 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2017, pp. 301–309.

[27] J. W. Goethe, *Faust*. De Gruyter, 2021.