MOTION TRANSFER IN VIDEOS USING DCGAN

Moolchand Sharma[1]*, Prerna Sharma[2], Manish Kumar Jha[2] , Rohan Singh[2]

[1,2]Department of Computer Science & Engineering, MAIT,GGSIPU,DELHI

sharma.cs06@gmail.com[1]*

prernasharma@mait.ac.in[2]

mkjha1998@gmail.com[2]

rohan021195@gmail.com[2]

Abstract— Motion Transfer has a wide variety of applications, such as creating motion synchronized videos in film industries and video making apps. The research paper presents a novel approach for motion transfer from a source video to the target person. This approach focuses on the video to video translation using various poses generated in the frames of video for translation. The approach makes use of Pose Generation Convolutional Neural Network to synthesize arbitrary poses from source videos and train the pix2pix – DCGAN(Deep Convolutional Generative Adversarial Networks), which is a conditional generative adversarial network consisting of multi-scale discriminator and generator for target video frames generation. It uses PatchGAN loss, VGG loss, and Feature Matching Loss function for improving and optimizing models. The presented approach provides compelling results of the generated DCGAN model with the discriminator loss of 0.0003 and a generator loss of 5.8206.

Keywords: Generative Adversarial Networks, Convolutional Neural Network, Pose Detection and Estimation, Video to video translation, Motion transfer

I. INTRODUCTION

Motion transfer between two video subjects is transferring the body posture of one video subject to another. It has a wide variety of applications in film industries. It can be used for synthesizing synchronized dancing videos or action sequences. Another widely used application is generating realistic-looking videos and images for photo editing software. Motion transfer between source and target video subjects is a tricky process that focuses on translating body posture, facial expressions, and source subjects pose to the target subject while generating a constant streamlined motion forming a video.

In this paper, we present a novel approach for motion translation between source and target videos. The goal is to impose poses of the source subject onto the target subject. To achieve this, we propose an end to end approach using Deep Convolutional Generative Adversarial Networks(DCGAN) for generating high-quality motion translated images. We aim to translate the hand, legs, and body posture of the source in the first image to that in the target, as shown in figure 1. Due to the lack of output images of the target in various poses, it is impossible to perform this task of motion transfer through supervised learning techniques. Generative Adversarial Networks are used for generating unsupervised images of the source. Although this technique helps in realistic motion transferring, it is susceptible to the complex background, different object sizes and scales, lighting, clarity, and various noises in the image. The approach is also vulnerable to problems like multiple subjects in the source video performing multiple different actions and position of the target in the generated frames.

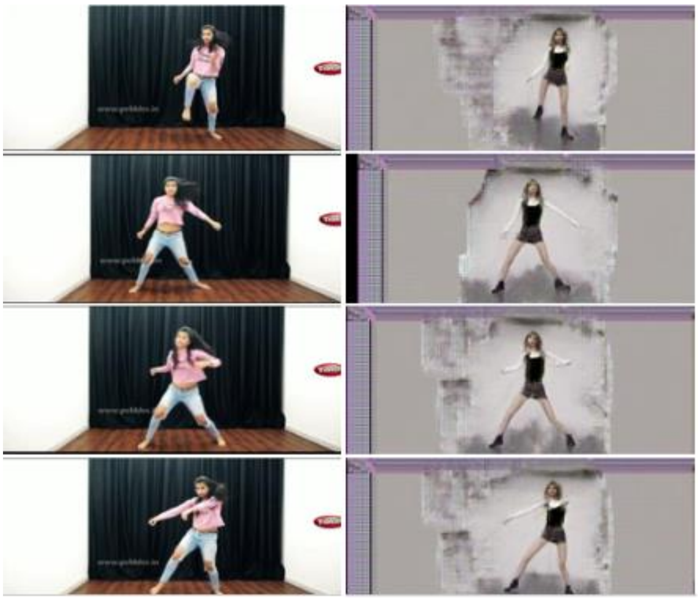

Fig. 1: Motion translation from source to target

The dataset used for training the DCGAN is a video in which the target object is doing body movements, for

Accepted 20th August 2020

each frame of this video poses is generated using open pose[13]. The generated pose is the input for DCGAN, and the original frame is the expected output from the DCGAN. Using this pair of input and output images, we trained our DCGAN (as shown in figure 2).

Apart from the introduction, this paper consists of five other segments: Segment 2 analyses the related work made in the field related to motion transfer learning. The methodology is showcased in segment 3. The results and analyses have been performed in segment 5. In Segment 6, the conclusion has been drawn along with the future scope—the last Segment contains a list of references used in this paper.

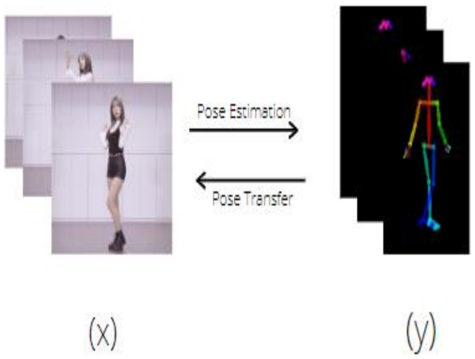

Fig. 2: Our technique makes correspondences by recognizing presents in video outlines (Video to Pose) and afterward figures out how to create pictures of the objective subject from the assessed present (Pose to Video)

II. RELATED WORK

Generative Adversarial Networks belong to the set of generative models; that is, they can produce new content—for example, a human face generator, which generates faces of humans who may not even exist. And speaking in mathematical terms, GAN is basically generating some random data, and we try to model our network in such a way that it produces data from a specific class, which was human faces in a human face generator. Inside GAN, two networks are working, one is a generator, and the other is discriminator; they both are neural networks. What a generator does is, it takes some random input and generates an output. The discriminator takes the output of the generator as input and tells whether it is real or fake. So speaking about training, the generator tries to fool the discriminator by producing an image, which is close to real, and discriminator tries to find out whether the image is real or fake. So there is a fighting kind of going between generator and discriminator, and this fight eventually increases the performance of both generator and discriminator, and eventually our generator is trained.

Pose generative image modeling from DCGAN is a relatively new approach for automatic image generation. Although various techniques have been used for motion transfer video to video translation. The approach in [1] renders subject images in various motions using multi-view fitted 3D models. A novel approach for rendering video-realistic interactive character animation from a database of 4D actor performance captured in a multiple camera studio is proposed in [2]. The approach successfully reconstructs captured dynamic shape appearance of human-like character motion. A MoCoGAN proposed framework is depicted in which generates a sequence of video frames consisting of content and motion parts [3]. Hence, allowing them to generate videos with either different motion of different content. In an RNN architecture with forwarding kinematics, the layer is proposed, which translates input motion into target characters with varying figures of the skeleton [4]. A video-based cloning technique using deep generative networks is proposed in [5]. Another similar approach is presented in which also focuses on spatiotemporal constraints for effective retargeting, which uses a novel motion descriptor based on optical flow for retargeting[6][7].

Detecting and generating poses is a crucial step in the process of motion translation. A Deep neural network-based approach for high precision pose estimation [8]. Approaches in [9, 10] employ adversarial networks for human pose estimation and generating 3D pose annotations. An open-source full-body estimation for 2D pose estimated in presented in [11][13]. An end to end, trainable GAN is used in which focuses on fitting 3D poses to 2D images[12].

In recent works, multiple generative networks such as human synthesis net and fusion net are used for enhancing realism into the video to video translation[14]. Other works include feeding medium quality-controlled 3D models for synthesizing human character images[15]. With the recent development in the field of multiple different types of GANs, polished unsupervised image generation is possible. CycleGAN introduces a cycle consistency loss by coupling inverse mapping between target and source domains, cycle consistency maintained by mapping F and G functions, $F(G(x)) \approx x$ and $G(F(y)) \approx y$ in X and Y domains, cycle consistency loss along with adversarial loss is used[16].

High-resolution condition GAN is used in which uses a multi-scale generator and discriminator and incorporates object instance segmentation, the technique presented in the paper allows users to interact

Accepted 20th August 2020

with object appearance[17]. Another approach uses a DiscoGAN to learn cross-domain relations while preserving identity and orientations. The DiscoGAN present in the paper doesn't use extra annotated supervision[18]. CoGAN or Coupled GAN uses marginal samples to generate joint distribution[22]. GAC-GAN is used for appearance-controlled movement translation. GAC-GAN uses ACGAN loss and shadow extraction modules and is used for generating new appearances from video records. Two types of GANs are used in making GAC-GAN, layout GAN and appearance GAN. Layout GAN helps in describing motion in videos at pixel level [23]. SimGAN or simulated and unsupervised learning GAN doesn't use random noisy vectors. Still, it uses a synthetic image, which is later refined while preserving the annotation information; this type of GAN also uses adversarial loss while training the refiner network[24]. Pix2Pix is another type of conditional GAN used for the image to image translations, this type of GAN uses UNet architecture which contains two both encoder, decoder network, the discriminator used here is PatchGAN architecture also known as markovian discriminator[25].

Xu et al. use multi-see catches of an objective subject performing straightforward movements to make a database of pictures and movement translation through a fitted 3D skeleton and relating surface work for the objective[21]. Work by Casas et al. utilize 4D Video Textures to minimalistic ally store a layered surface portrayal of an examined target individual and utilize their transiently intelligible work and information portrayal to render video of the objective subject performing novel movements[20]. Interestingly, our methodology investigates movement transfer between 2D video subjects. What's more, stay away from information adjustment and lifting into 3D space. A UNet-like architecture is used in Pose Guided Person Generation Network for synthesizing arbitrary but blurry results, which are later refined for generating high-quality images. The technique proposed in the paper contains two major steps for synthesizing such images, pose integration stage consists of generator G1 and pose mask loss, in the second stage image refinement, a generator G2 is used which is completely tested on Deep fashion dataset[19].

III. METHODOLOGY

The methodology for the process is shown below:

Step 1: Preprocessing and cleaning of the input Video

Step 2: Pose Detection and Estimation from all frames of the video

Step 3: Training DCGAN to generate various poses for the input image

Step 4: Combining images generated from DCGAN to form the augmented vide**o**.

3.1 Preprocessing

In the first step of our approach, the source video is split into multiple frames first. We follow this by removing unwanted noise from our split images using bilateral filters. It is highly effective in noise removal while keeping edges sharp. But the operation is slower compared to other filters. The Gaussian filter takes the neighborhood around the pixel and finds its Gaussian weighted average. This gaussian filter is a function of space alone; that is, nearby pixels are considered while filtering. It doesn't consider whether pixels have almost the same intensity. It doesn't consider whether a pixel is an edge pixel or not. So it blurs the edges also, which we don't want to do since sharp edges around the image, helps in detecting and estimating poses accurately. Some starting and ending frames of the video are removed from the dataset because it contains static noise which contains excessive noise and less movement.

3.2 Pose Detection and Estimation

In the second step of our approach, we need to generate the poses of the detected people from the multiple preprocessed frames of the video. The procedure of pose detection is divided into four major steps, at first, a DCGAN model is created, this model is designed to predict the heatmaps of the areas detected/estimated poses with input as the video frame having source person. Using the heatmap generated from the above model, body parts are extracted, peak values from the heatmap are used to link the body parts. For pose estimation, a DCGAN is used to generate the pose of the source person from the image. Two major stages of pose generation and estimation is posing integration and image refinement. A U-Net like generator and discriminator is used in the stages mentioned above and the source image and target image are inputted to generate the estimated pose [19]. The architecture used is mentioned in figure 3.

3.3 Pose to Video Translation

This is performed using a conditional GAN whose input is the generated pose from the previous step along with the random Noise. In this step, the generator tries to fool the discriminator. Discriminator tries to figure out whether the image is fake or real, And during this fight between the generator and the discriminator accuracy of

Accepted 20th August 2020

the model's increases which in results into two independent models and these can be used to either generate the target image or to check whether an image is real or fake.

The loss function used is a combination of Adversarial and Content loss, which is a modification of the Perceptual Loss function. The Adversarial loss gives a weight of 10-3 to the discriminator loss, but in our implementation, the weight of this loss is increased to 10^{-2}. The generator network architecture is shown in figure 4.

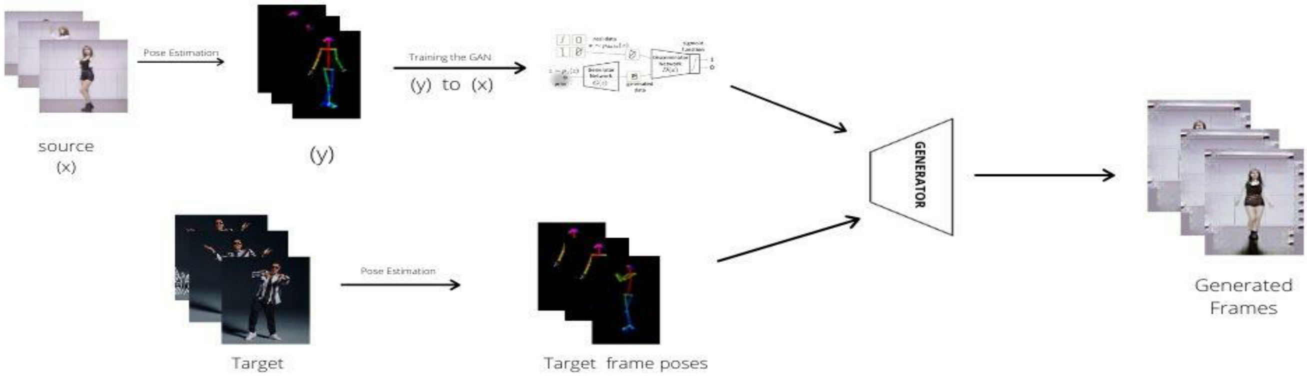

Fig. 3: Architecture

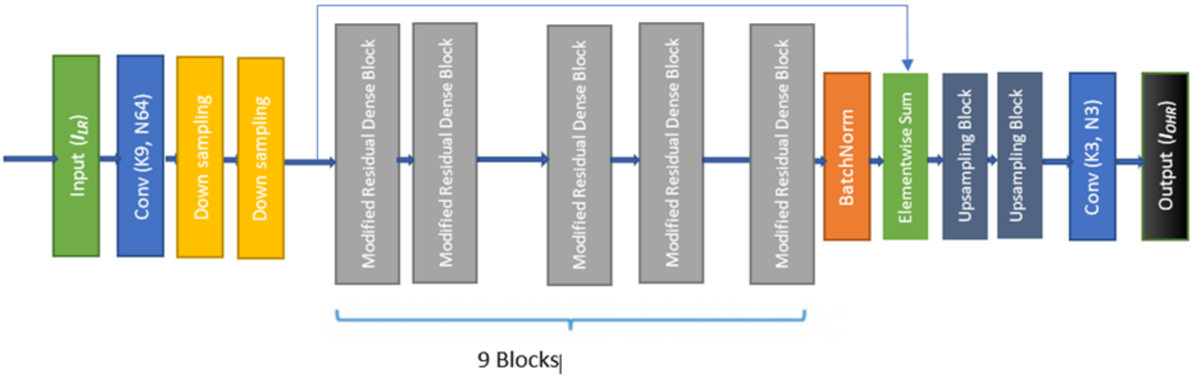

Fig. 4: Generator Network Architecture

IV. RESULTS AND ANALYSIS

The DCGAN modal proposed has been trained on 19675 images of a single dancer YouTube video, and a total of 19000 iterations were performed. The final discriminator loss is 0.0003, and the generator loss is 5.8206. The graph of discriminator loss and generator loss is plotted for every iteration, as shown in figures 5 and 6. As shown in the figure, the decrease in losses indicates that the model is making better and optimized results. In every iteration, we use these losses to train the model better and further improve accuracy and decrease losses.

Accepted 20th August 2020

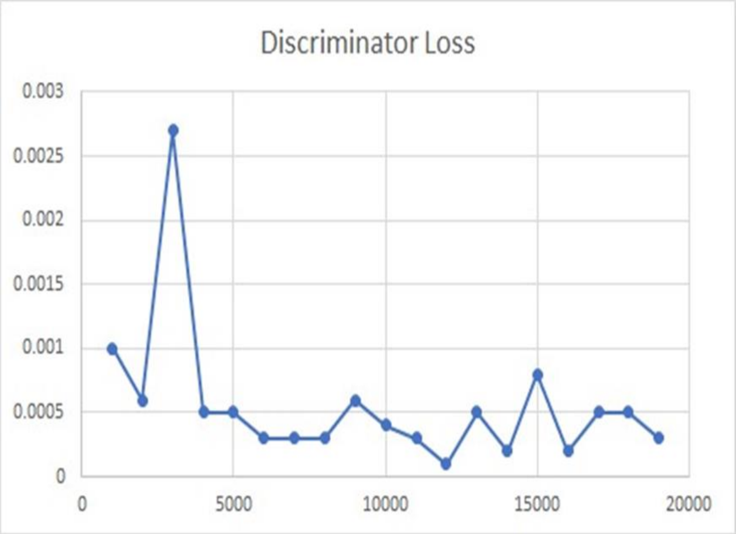

Fig. 5: Discriminator loss

Discriminator Loss is determined using the adversarial Loss function that is described by equation (i) and equation (ii) :

$$Loss_{adv}(G, D_y, X) = \frac{1}{m}\sum_{i=1}^{m} \quad (1 - D_y(G(x_i)))2 \qquad \text{(i)}$$

$$Loss_{adv}\ (F, D_x, Y) = \frac{1}{m}\sum_{i=1}^{m} \quad (1 - D_x\big(F(y_i)\big)2 \qquad \text{(ii)}$$

Let's consider the generator loss, which is the generator's main motive is to maximize this functionality. In other words, It tries to make discriminators wrong every time and try to increase the probability of getting discriminators wrong.

D(x) is the Discriminator’s output for a real instance.
G(z) is the generator's output.
D(G(z)) is the Discriminator output for a fake instance.

The output of Discriminator D does not have to be between 1 and 0.

Discriminator Loss tells us about how well he can recognize the real and fake images, and as the loss decreases, our generator accuracy of generating the image increases. But since this is an iterative process, discriminator also has been improving their results as a generator, so the result gets benefited from this, and our images will become more realistic as the iterations/epochs increases .

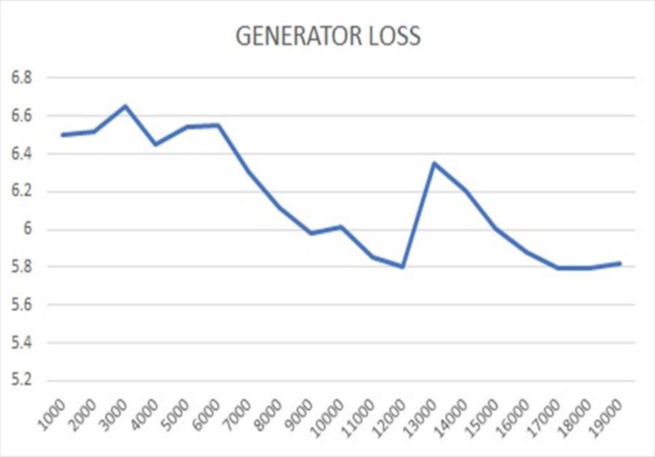

Fig. 6: Generator loss

The final results for the generated frames and alongside the original video is shown in figure 6.
The Loss is consisting of two losses that are adversarial loss and the cyclic loss,

The Cyclic Loss is as follows, as shown in equation (iii). So here in figure 6, We can see generator loss is continuously decreasing. Still, the rate is less, which means our generator is struggling to confuse our discriminator because the generator is also learning at the same time.

The combination loss is shown in equation (iv). Here the value of λ is taken as 100. to somewhat equalize the impact of generator and discriminator both. As shown in figure 7, the frame in the first column is the motion we are willing to transfer to our target image and column in the second frame is the estimated pose, and the third and last column is the motion transfer from the source object to the target image. As we can see, we can transform the images to particular pose well, but still, there's an issue with the image quality (figure 8).

$$Loss_{cyc}(G, F, X, Y) = \frac{1}{m} \sum_{i=1}^{m} \left[F\left(G(x_i)\right) - x_i\right] + \left[G\left(F(y_i)\right) - y\right] \qquad \text{(iii)}$$

The combination loss will be, as

$$Loss\_full = Loss\_adv + \lambda Loss\_cyc \qquad \text{(iv)}$$

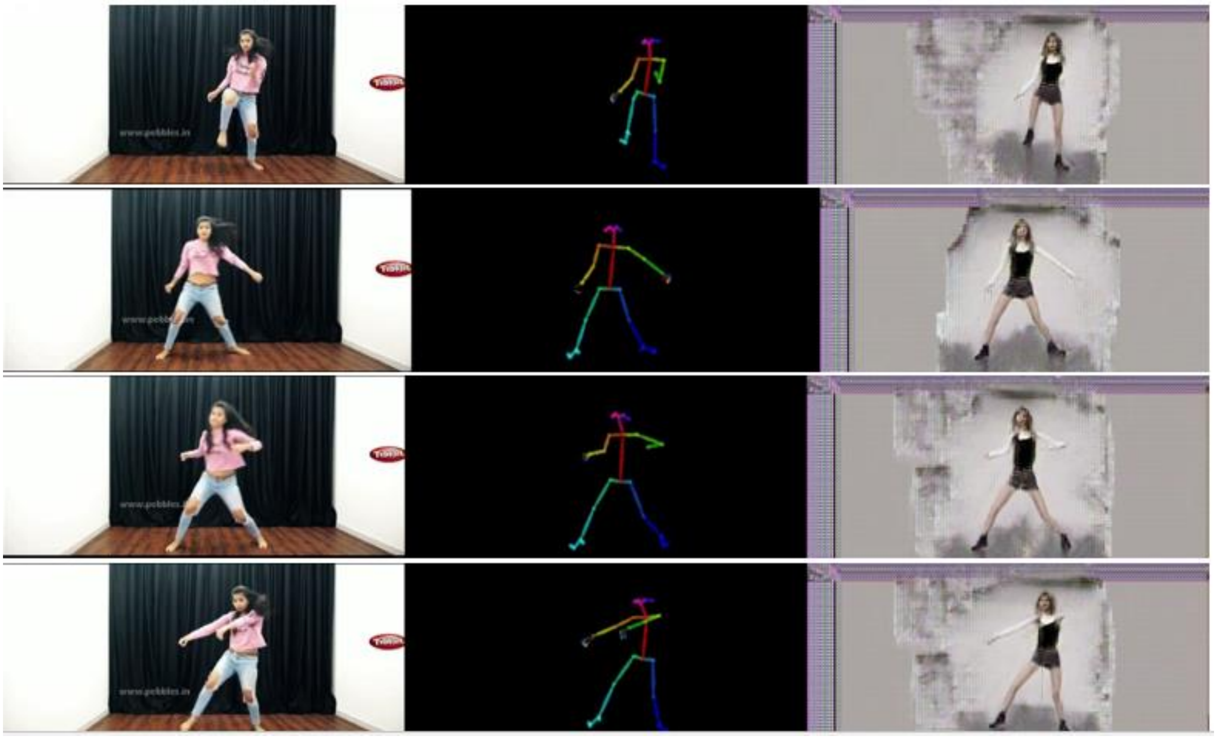

Figure 7: Generated results

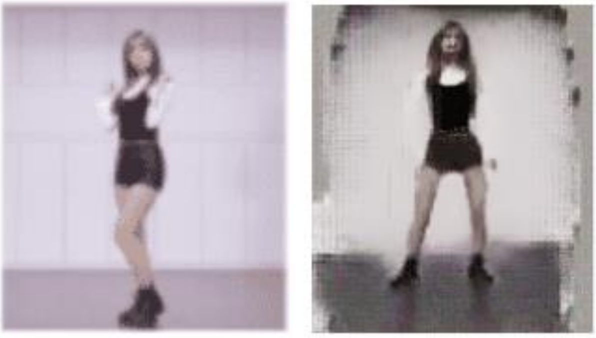

Fig. 8: Original Image and the Generated image

V. CONCLUSION AND FUTURE SCOPE

Our proposed approach was able to produce high-quality results of the generated images from the source video with the discriminator loss of 0.0003 and generator loss 5.8206. One can also use a densePose network to generate the pose of the source image and then use the generated poses to train the generator; this method is expected to give better results as the pose generated by densePose conver's whole body as compared to stick image generated by openPose. One of the use cases of this framework is to create a synchronized motion dance video of

Accepted 20th August 2020

multiple different subjects. Having trained models for multiple subjects, we can use the source video to generate the motion of all target persons. The second use case of this system can be to use discriminator to check whether the image is “fake” or “real.” One can also use a separate face GAN to generate better facial expressions, which would result in a more realistic video of the target person.

Trained discriminator can also be used to detect whether an image is real or fake as it was trained to determine if the pose is generated accurately. Realistic, this can be very useful in forensic science for detecting image morphing or modifications done on an image. Background augmentation using BAGAN can be used for effectively adding a background to the proposed model, which lacks in the background and spits out the gray background, as shown in figure 6. The advantage of using background GAN would help in augmenting various backgrounds for synchronized dancing videos.

VI. REFERENCES

[1]. G.K.M Cheung, S. Baker, J. Hodgins, and T. Kanade (2004). “*Markerless human motion transfer*.” In 3D Data Processing, Visualization, and Transmission, 2004. 3DPVT, 2004; Proceedings. 2nd International Symposium .pp. 373–378. IEEE, 2004.

[2]. D. Casas, M. Volino, J. Collomosse (2014). 4D *Video Textures for Interactive Character Appearance*. Computer Graphics Forum (Proceedings of EUROGRAPHICS), Vol. 33, Issue 2, pp. 371–380.

[3]. S. Tulyakov, M. Liu, X. Yang, and J. Kautz (2018). *MoCoGAN: Decomposing motion and content for video generation*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4]. R. Villegas, J. Yang, D Ceylan, and H. Lee(2018). “*Neural kinematic networks for unsupervised motion retargeting*.” The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[5]. K. Aberman, M. Shi, J. Liao, D. Liscbinski, and B. Chen (2019). “*Deep video-based performance cloning. In Computer Graphics*” Forum, Vol. 38, pp. 219–233. Wiley Online Library.

[6]. A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh(2018*). Recycle-gan: Unsupervised video retargeting*. In ECCV.

[7]. A. Efros, A. Berg, G. Mori, and J. Malik (2003). *Recognizing action at a distance*. In IEEE International Conference on Computer Vision, pp. 726–733, Nice, France.

[8]. A. Toshev, and C. Szegedy (2014) . Deeppose: Human pose estimation via deep neural networks. In CVPR.

[9]. C. Chou, J. Chien and H. Chen(2018). "Self-Adversarial Training for Human Pose Estimation," *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Honolulu, HI, USA, pp. 17-30.

[10]. W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li and X. Wang(2018) "3D Human Pose Estimation in the Wild by Adversarial Learning," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, pp. 5255-5264.

[11]. Hidalgo, Gines, Yaser Sheikh, Kris M. Kitani, Aayush Bansal, Ramon Sanabria, Donglai Xiang, Xiu Li, and Haroon Idrees(2019) “*OpenPose: Whole-Body Pose Estimation*.”.

[12]. Chen, Xu, Jie Song, and Otmar Hilliges (2019) “Unpaired Pose Guided Human Image Generation.” *ArXiv* abs/1901.02284 (2019).

[13]. Cao, Zhe, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh(2018). “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields.” *IEEE transactions on pattern analysis and machine intelligence*.

[14]. Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. Berg (2019). “*Dance generation: Motion transfer for internet videos*.” arXiv preprint arXiv:1904.00129 .

[15]. L. Liu, W. Xu, M. Zollhofer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt (2019) “*Neural rendering and reenactment of human actor videos*.” ACM Trans. Graph., Vol. 38, Issue 5,pp. 139:1–139:14 .

[16]. Jun-Yan Zhu, Taesung Par*, Phillip Isola, and Alexei A. Efros(2017) "*Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*" in IEEE International Conference on Computer Vision (ICCV), .

[17]. Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. "*High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs*," in CVPR, 2018.

[18]. Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. (2017). *Learning to discover cross-domain relations with generative adversarial networks*. In Proceedings of the 34th International Conference on Machine Learning – Vol. 70 (ICML'17), Doina Precup and Yee Whye Teh (Eds.), Vol. 70. JMLR.org ,pp. 1857-1865.

[19]. Ma, Liqian & Jia, Xu & Sun, Qianru & Schiele, Bernt & Tuytelaars, Tinne & Van Gool, Luc. (2017). *Pose Guided Person Image Generation*.

[20]. Feng Xu, Liu, Carsten Stoll, Gaurav Bhara, James Tompkin, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Theobalt (2011). *Video-based characters: creating new human performances from a multi-view video database*. In ACM Transactions with Graphics (TOG), volume 30, page 32. ACM.

Accepted 20th August 2020

[21]. Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. *4D Video Textures for Interactive Character Appearanc*e. Computer Graphics Forum

[22]. Y. Liu, and O. Tuzel (2016). *Coupled generative adversarial networks*. In NIPS .

[23]. D. Wei, X. Xu, H. Shen, and K. Huang. (2020).GAC-GAN*: A General Method for Appearance-Controllable Human Video Motion Transfer*.

[24]. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang. (2017) *Learning from Simulated and Unsupervised Images through Adversarial Training*.

[25]. P. Isola Jun-Yan Zhu T. Zhou and A. Efros(2018). *Image-to-Image Translation with Conditional Adversarial Networks*.

Accepted 20th August 2020