Grenze International Journal of Engineering and Technology, Jan Issue

Human Imitation in Images and Videos using GANs

Dr. Keshetti Sreekala[1], Vishal Das[2] and Amulya V[3]
[1]Asst. Professor, Dept of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India
[2-3]Student, Dept of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India

***Abstract*—The paper aims to generate a pseudo-Image/Video from a given input Image and a reference video where the features of the human in the reference video are transferred to the human in the input Image. The source features, such as texture, style, color, and face identity, can be preserved using a Liquid Warping GAN. A three dimensional body mesh of the subject can be given to the GAN so that it propagates the source features in both image and feature spaces, and synthesizes an image with respect to the reference image or video. Full body motion imitation is a challenge in Artificial Intelligence and requires several modules in order to perform effectively and efficiently. To combine to different sources of appearance and motion into a single cohesive video joining them both is both interesting and useful for many applications. iPER dataset is used to train the GAN. The dataset is taken from Shanghai Vision and Intelligence Perception Lab. Within the dataset, there are different conditions of shape, height and gender dispersed among 30 human subjects. Each human subject wears separate clothes and performs a video with random actions. This technique can be employed for animations where an animator doesn't have to draw and render multiple sequences where an actor can just perform the sequence physically and this can be applied to multiple humans, thus reducing time and effort. This method can also be used to transfer the style of clothing which has several applications while shopping online.**

***Index Terms*— Computer Vision, Generative Adversarial Networks, Deep Learning, Artificial Intelligence.**

I. INTRODUCTION

Imitation refers to synthesizing a new image or video of a given person in a target pose or target motion. Imitating and impersonating humans in images and videos is a challenging problem for editors and software in computer vision, which has great application potentials in editing images, video generation, virtual try-on, etc. Generative Adversarial Networks (GANs) is the leading technology for achieving success in human imitation, pose and appearance transfer. Some methods employ neural networks to directly learn the mapping from the source image and pose to the target image. Most methods use keypoints as the pose representation, but with keypoints it is difficult to predict a sharp and reasonable image with sparse correspondences when the source pose and the target pose have large differences. Flow-based methods estimate an appearance flow to obtain correspondences, which is used to warp the source image to align with the target pose thus solving the above problem. The output image delivered by refining and enhancing the warped image or decoding the warped image feature is achieved by movement of the source image elements. Thus, the initial details of the source image are preserved by the generated image, but the invisible region is not recovered. Some techniques use introduce

Grenze ID: 01.GIJET.8.1.10

human parsing maps for human pose transfer in order to predict the invisible regions. Semantic correspondence is provided by the human parsing map to synthesize the final image and enables applications of person image editing. However, the category and texture of clothing cannot be disentangled by this method and it fails to preserve spatial context relationships. The following are some disadvantages present in existing systems:

- Does not retain features such as body shape etc;
- Unable to decouple the type, shape and style of clothing.

The proposed system is a novel two-stage generative model for Person Image Synthesis and Editing, which is able to generate realistic person images with desired poses, textures, or semantic layouts. To achieve human motion transfer, a human parsing map aligned with the target motion is synthesized in order to represent the shape of clothing by a parsing generator, and then generate the final video using a generator. The shape and style of clothing can be decoupled using a combined global and local per-region normalization to predict the type and style of clothing for invisible regions. To retain the spatial context relationship in the source image a spatial-aware normalization is used. The texture transfer and efficiency of the model helps human image and video editing an also for rendering animations.

II. Literature Review

Zhang et al. [1] proposed "PISE: Person Image Synthesis and Editing with Decoupled GAN" which uses the Decoupled GAN and Region encoding and normalization methodology. It is able to generate realistic person images withdesired poses, textures, or semantic layouts but unable to retainsource features of the person resulting in inaccurate results.

Wen Liu et al. [2] proposed "Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis" using the Liquid Warping GAN and Novel View Synthesis methodology. It tries to preserve the source information, such as texture, style, color, and face identity and can also be used for serial image motion transfer but requires high VRAM memory even for low resolution images.

Guha Balakrishnan et al. [3] proposed "Synthesizing Images of Humans in Unseen Poses" using Modular generative neural network and UNet architecture. It helps in Producing a depiction of a person in desired pose, retaining the appearance of both the person and background. In this technique the appearance transfer of the person limits only to clothing but notfor body shape or figure.

Haoye Dong et al. [4] proposed "Soft-gated Warping-GAN for pose-guided person image synthesis" using Warping-GAN and Target part segmentation map. It uses spatial layouts for guiding image synthesis with higher-level structure constraintsbut applying this technique on videos is not feasible.

Aliaksandr Siarohin et al. [5] proposed "Deformable GANs for Pose-based Human Image Generation" for synthesizing a novel image of a person in a target pose, given an image of a person and a target pose. However this technique works efficiently only when a significant number of 2D keypoints are provided.

Liqian Ma, et al. [6] proposed "Pose Guided Person Image Generation" which utilizes an U-Net like network to generate imitating images. However pose transfer using this techniques can be very expensive computationally as it has to generate several pixels for filling appearance details, so using this method for video generation is not feasible.

III. Methodology

A. iPER Dataset Analysis

We are using the iPER dataset for training and testing on theGAN architecture. In the dataset there are 30 human subjects of varied conditions of shape, height and gender. Each human subject wears separate types of clothes and performs an A-pose and a video with random actions.

iPER contains 206 video sequences with 241,564frames. The distribution of videos of each category is shown in the figure 1 and figure 2.

The figure 1 represents the number of videos with respect to thetype of action performed by the person in each particular video. The X-axis represents the number of actions (each video contains a single action) and Y-axis represents the type of actionperformed. Some humans might wear multiple types of clothing, and there are a total of 103 clothes.

The figure 2 represents the number of clothing pieceswith respect to the style of clothes worn by the person in each particular video. The X axis represents the number of clothing pieces (each video may contain multiple pieces of clothing) andY axis represents the type of cloth worn.

The dataset is split into training: testing set at the ratioof 8:2 according to the different clothes.

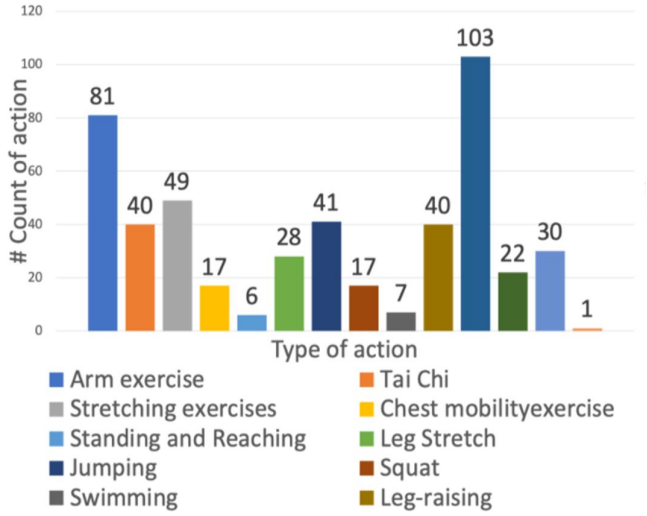

Fig..1: Dataset bar graph representing number of videos as per actions

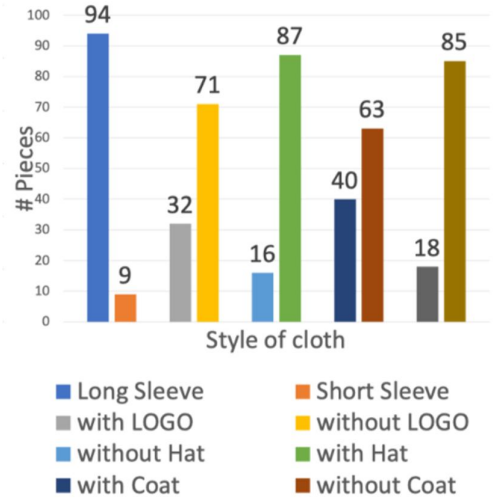

Fig..2: Dataset bar graph representing number of clothing pieces as per type of cloth

Dataset Format:

There are 206 videos in total, all of those have name format,as "xxx_yyy_zzz.mp4", where:

- xxx: actor name
- yyy: appearance id, an id specifies a particular cloth
- zzz: action type.

B. Generative Adversarial Network (GAN)

A generative adversarial network (GAN) has two parts:

- The generator learns to generate plausible data. The instances which are generated are used by the discriminator as negative training examples.
- Overtime the discriminator trains to distinguish the fake data from real data as produced by the generator and punishes the generator for producing implausible results.

When the GAN training starts, the generator produces randomdata which appears fake from the original data, and the discriminator classifies the data as fake. Overtime with the progression of the training, the generator producing output true to ground truth that can fool the discriminator. The training ends when the generator

training proceeds well enough, thediscriminator fails at telling the difference between the real image and the fake image. Therefore it starts to classifythe fake image as real which means the generator training is successful.
Since a GAN has to train two separate networks, the trainingalgorithm has to address two complications:

- GANs must multitask with two different kinds of training i.e. generator and discriminator training.
- GAN convergence is hard to identify.

The generator output is a direct input to the discriminator. The classification performed by the discriminator provides a signal to the generator which the generator uses to update its weights using backpropagation. The design flow of a Generative Adversarial Network (GAN) Architecture is shown in the figure 3.

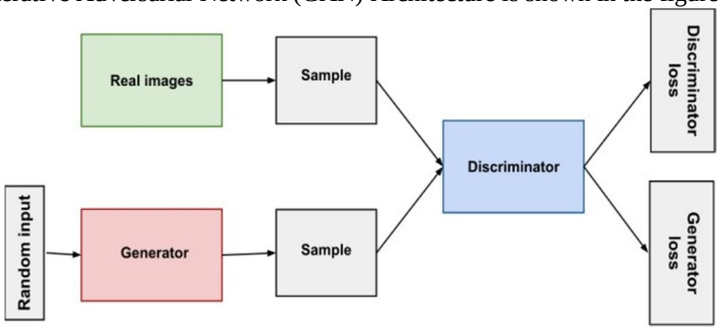

Fig..3: Generative Adversarial Network Architecture

The discriminator performance gets worse with improvement in generator training because the discriminator can't identify thedifference between real and generated image. The minimum accuracy the discriminator can have after successful training is 50% as the generated image is almost same as the real image.

C. Architecture of Human Imitation

To train the model for the appearance transfer task, we use motion imitation denoting the source image as Is and the reference image I*r*. The first body mesh recovery module will estimate the 3D mesh of Is and Ir, and render theircorrespondence maps, Cs and Ct. Based on two correspondencemaps, Cs and Ct and their projected mesh in image space, the flow composition module will calculate transformation flow T as shown in figure 4.

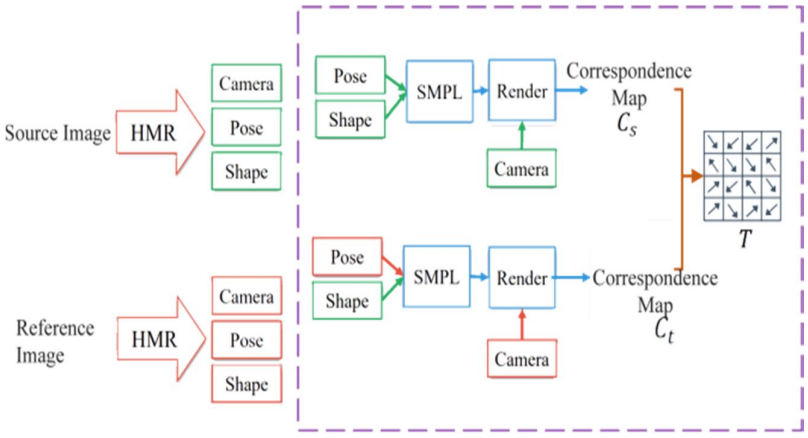

Fig..4: Calculating transformation flow

As shown in the above figure 4 the correspondence maps aregenerated from source and reference image using features suchas the camera perspective, pose and shape of the human. The generated correspondence maps are used to calculate the Transformation flow (T). The source image I_s is thereby decomposed as front image I_{ft} and masked background I_{bg}, andwarped to I_{syn} based on transformation flow T.
The Liquid Warping GAN module has a generator with threedifferent streams for reconstruction and synthesis. I_t separatelygenerates background image by G_{BG}, reconstructs the source image I'_s by G_{SID} and synthesizes the image I'_t under reference condition by G_{TSF}. To preserve the details of source image, a Liquid Warping

Block (LWB) is used. LWB propagates the source features of GSID into GTSF at several layers and preserve the source information, in terms of texture, style and color.

D. Modules

The model trains and learns in three phases, each phase consisting of deconstruction and reconstruction of the source and reference images. The design and flow of the model from one phase to another is represented in the figure 5.

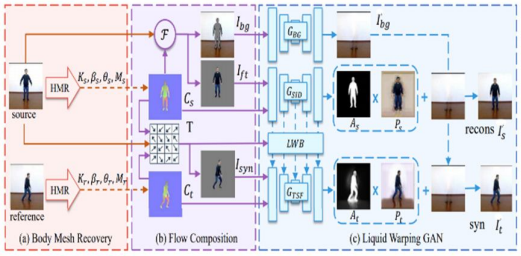

Fig. 5: Architecture of human imitation in images and videos using GANs

The source image is denoted as I_S and the reference image isdenoted as I_r. A body mesh recovery module will estimate the 3D mesh ofeach image, and render their correspondence map, C_S and C_t. The flow composition module will first calculate the transformation flow (T) based on two correspondence maps and their projected vertices in image space. In the last GAN module, the generator performs the image reconstruction and generates new images by imitating the motion

Body Mesh Recovery Module:

Given source image I_S and reference image I_r, the role of thisstage is to predict the kinematic pose (rotation of limbs) and shape parameters, as well as 3D mesh of each image. A pre- trained model of Human Mesh Recovery (HMR) is used for 3Dpose and shape estimation due to its good trade-off between accuracy and efficiency. In HMR, an image is firstly encoded into a feature by a ResNet-50 and then followed by an iterative 3D regression network that predicts the pose (θ) and shape (β) of SMPL, as well as the weak-perspective camera (K). Skinned Multi-Person Linear model (SMPL) is a 3D body model that can be defined as a differentiable function M (θ, β) and it parameterizes a triangulated mesh with pose parameters θ and β. Here, shape parameters β are coefficients of a low- dimensional shape space learned from thousands of registered scans and the pose parameters θ are the joint rotations that articulate the bones via forward kinematics. With such process, we will obtain the body reconstruction parameters of source image, $\{K_S, \theta_S, \beta_S, M_S\}$ and those of reference image, $\{K_r, \theta_r, \beta_r, M_r\}$, respectively.

Flow Composition Module:

In this stage a correspondence map of source mesh Ms and that of reference mesh Mr under the camera view of Ks is rendered. The source and reference correspondence maps are denoted as Cs and Ct, respectively. In this project, we use a fully differentiable renderer, Neural Mesh Renderer (NMR). We thereby project vertices of source Vs into 2D image space by weak perspective camera. Then the barycentric coordinates of each mesh face is calculated to obtain. Next, we calculate the transformation flow T by matching thecorrespondences between source correspondence map with its mesh face coordinates fs and reference correspondence map. Consequently, a front image I_{ft} and a masked background image I_{bg} are derived from masking the source image I_S based on C_S. Finally, we warp the source image I_S by thetransformation flow T, and obtain the warped image I_{syn}.

Liquid Warping GAN:

The Liquid Warping GAN stage synthesizes high-fidelity human image under the desired condition. More specifically, itperforms the following along three streams:

1. Synthesizes background from the source image.
2. The color of invisible parts is predicted based on thecolor of visible parts.
3. Using the SMPL reconstruction it generatesmissing and new pixels.

The generator network works in a three-stream manner and the discriminator network follows the architecture of Pix2Pix which is based on conditional GANs. The three streams of the generator network are as follows:

- Background stream, namely G_{BG}

- Source identity stream, namely G_{SID}
- Transfer stream, namely G_{TSF}

Background Generator Stream, G_{BG} , works on the concatenation of the masked background image I_{bg} and the mask obtained by the binarization of C_s in color channel (4 channels in total) to generate the realistic background image I'_{bg}.

Source Identity Generator Stream, G_{SID}, is a de-noising convolutional auto-encoder which aims to guide the encoder toextract the features that are capable to preserve the source information. Together with the I'_{bg}, it takes the masked source foreground image I_{ft} and the correspondence map C_s as inputs, it reconstructs source front image I'_s.

Transfer Generator Stream G_{TSF} stream synthesizes the final result, which receives the warped foreground by bilinear sampler and the correspondence map C_t as inputs. To preservethe source information, such as texture, style and color, LiquidWarping Block (LWB) that links the source with target streamsis used. It blends the source features from G_{SID} and fuses them into transfer stream G_{TSF}.

Liquid Warping Block (LWB) can address multiple sources, such as in human motion transfer, preserving the face of source one, and wearing some clothing items from the source two, while wearing some other clothing items from the source three. The different parts of features are aggregated into the transfer generator stream, G_{TSF}, independently using its own transformation flow.

E. Loss Functions for Human Imitation

In body recovery module, we follow the network architecture and loss functions of Human Mesh Recovery. For body mesh recovery a pre-trained model of Human Mesh Recovery is used. In the Liquid Warping GAN training phase, a randomlysampled pair of images from each video is used and one of themis set as source, and another as reference I_r. For this project, the model is trained for motion imitation and appearance transfer at the same time . The model does not need to train separately to perform these tasks. It is able to perform these tasks with only being trained once. The whole loss function contains the following terms:

- Perceptual loss
- Face identity loss
- Adversarial loss.

Perceptual Loss:

It regularizes the reconstructed source image I'_s and generated target image I'_t to be closer to the ground truth I_s and I_r in VGG-19 subspace. VGG-19 is a pre-trained 19 layer deep convolutional neural network. It is used to minimize the feature difference between the source image and the synthesized image and the motion or pose difference between the generated targetimage and the reference image.

Equation:

$L_p = \|f(I'_s) - f(I_s)\| + \|f(I'_t) - f(I_r)\|$

Here, f is a pre-trained VGG-19 network.

Face Identity Loss:

It regularizes the face from the synthesized targetimage I'_t to be similar to that from the image of ground truth, which pushes the generator to preserve the face identity. It is used to minimize the distance between the pixels of face in the generated target image and the reference image which was taken as an input.

Equation:

$L_f = \|g(I'_t) - g(I_r)\|$

Here, g is a pre-trained SphereFaceNet.

Adversarial Loss:

It pushes the distribution of synthesized images to the distribution of real images given as input. We use LSGAN−110 (Least Squares Generative Adversarial Network) loss in a way like PatchGAN for the generated target image I' . PatchGAN isa discriminator for GANs which penalizes structures in local image patches and classifies images with respect to each patch. The discriminator D is used to regularize I'_t to be more realistic-looking. A conditioned discriminator is used, and it takes generated images and the correspondence map C_t, whichhas 6 channels, as inputs.

Equation:

$L_{adv} = \sum D(I'_t, C_t)^2$

IV. Testing and Results

A. Results for Human Imitation

Input Image:

Fig. 6: Input image with random pose

An Image of a person with random pose is taken from the iPER dataset. The below figure 6 represents an image of a person in a random pose. The image has a resolution of 256 x 256 pixels. This is the source image which will be transposed onto the reference video in order to create the desired output.

Reference Video:

Fig. 7: Reference video frame

A video with an animated figure performing a jump is takenas the reference video. The below figure 7 represents a single frame of an animated persontaken from a video within the iPER dataset. It has a resolution of 256 x 256

pixels. This is the reference video onto which the source image is transposed upon in order to create the desired output. The pose in the frame represents the target pose.

Output Video:

Fig. 8: Output video frame

The output is the video of the person in the source image performing the actions performed by the animated figure in thereference video. The above figure 8 represents a frame of the output video inwhich the person in the source image imitates the motionperformed by that of the animated figure as shown in the reference video. The output is a video with a resolution of 256 x 256 pixels.

B. Benchmarks

TABLE I. EVALUATION AND BENCHMARKS FOR HUMAN IMITATION

NAME	**SSIM**	**LPIPS**
Pose guided person image generation (PG2)	0.854	0.135
Synthesizing images of humans in unseen poses(SHUP)	0.832	0.099
Deformable GANs for pose-based human image generation (DSC)	0.829	0.129
The Proposed Work	**0.844**	**0.093**

PG2:

Pose Guided Person Generation Network (PG2) is a network that allows the synthesis of images of people in random poses, based on an image of that person and an unseen pose. PG2 consists of the following stages: pose integration and image refinement.

The condition image and the target pose are provided as input to a network to generate an initial image of the person with the target pose in the pose integration stage. The initial and blurry image is then refined by training another generator network in an adversarial way.

It focuses on the tasks of image generation with a given reference image and generates an intended pose. It utilizes mask loss function which encourages the model to focus on the human body appearance transfer instead of the image background. For pose transfer, it works in two stages:

- First stage concentrates on the entire structure of the human body.
- Second stage concentrates on filling in appearance details left out in the blurry first image.

SHUP:

Synthesizing Humans in Unseen Poses (SHUP) address the computational problem of human pose synthesis in new poses. It generates a picture of a human in a desired pose, while retaining the source features of the human

and background in the input image. SHUP utilizes a modular generative neural network that generates novel poses which are trained on pairs of images and actions taken from various action videos.
The generative neural network works in layers. It divides the input image into separate layers such as the body layer, background layer etc. and moves the layers to novel locations and enhances their appearances, in a new background. SHUP uses a supervised learning model on pairs of images and their respective actions. SHUP takes a source image and a 2D pose as input, and a 2D target pose, and generates an output image.
SHUP uses segmentation on the source image and generates a background layer and multiple foreground layers corresponding to different body parts which allows fluidity of moving different parts of the body. The parts are then moved and adjusted to generate a new foreground, while the background is separately edited and filled with the appropriate pixels. The foreground and background are then combined to produce an output image.

DSC:
Deformable GANs for pose-based human image generation (DSC) address the problem of generating novel images of humans in a given target pose. An image of a human in a new pose can be generated by using source human and a target pose as input.
DSC employs deformable skip connections in the generator of a Generative Adversarial Network to adjust the pixel misalignments in the output image. This method used in the field of deformable object generation, given that the object is articulated and keypoints can be extracted from the image.
It addresses the problem of image generation where perspective variation or a deformable motion, changes the foreground object i.e. the human.
DSC aims to generate a human images conditioned on two variables:

- The appearance of a specific human in a given image
- The pose of the same human in another image.

This approach can be used with other deformable items such as human faces or animals with intelligible poses.

SSIM:
It stands for Structural Similarity Index Metric. A higher SSIM value indicates that the output image/video is closer to the source image in structural similarity i.e. a higher value is better.

LPIPS:
It stands for Learned Perceptual Image Patch Similarity. A lower LPIPS value indicates the similarity between the pixel patches of the output image/video and the source image i.e. a lower value is better.
There is no perfect measurement index that can be used to compare the results to that of a human observer but SSIM and LPIPS are viable among the known methods. On an average this model provides a better value than the existing systems judging by both Structural Similarity Index Metric and Learned Perceptual Image Patch Similarity values.

C. Explanation of perceptual loss, Face Identity loss and Adversarial loss

Perceptual loss:
Perceptual loss is employed to reconstruct the source image and provide source similarity to the image. Without utilizing this loss function the GAN does not converge and the resultant output video is distorted as several output frames consist of distorted images.

Face Identity loss:
It is used for regularizing the generator network to preserve the facial details of the subject in the source image. This loss is introduced within the loss functions to avoid warped faces. It converges when the facial similarity of the source image and the output image are same. So without this loss function the face of the subject in the above image will appear pixelated and blurry in the output image.

Adversarial loss:
It is the GAN loss function. It is a form of texture loss. It is used to for image patch similarity. The discriminator network classifies the images based on image patch similarity, so without this loss the source image details would not be preserved and the GAN generator network cannot converge. The output in this case would be pixelated as the generator cannot be trained properly without this loss function.

V. Conclusion

With the used GAN architecture along with Liquid Warping Block and loss functions implemented in PyTorch, thestability and performance of the model is increased and the GAN generates satisfactory results while being resource efficient.

While generator and discriminator are trained together, we only need the trained generator network for the inference of new outputs after training, i.e., the whole discriminator network canbe discarded making it a light weight and feasible approach forImitation.

Human Imitation is an intriguing concept in the field of computer vision and deep learning and can be used to solved various problems in image and video editing and we can conclude that the Human Imitation in Images and Videos usingGANs is a feasible approach to this problem.

References

[1] Jinsong Zhang, Kun Li, Yu-Kun Lai, Jingyu Yang. PISE: Person Image Synthesis and Editing with Decoupled GAN. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), March 2021.

[2] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, Shenghua Gao. Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), October 2019.

[3] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frdo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In The IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June 2018.

[4] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018.

[5] A. Siarohin, Enver Sangineto, Stephane Lathuiliere and Nicu Sebe, Deformable GANs for Pose-based Human Image Generation, CSCV, April 2018.

[6] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, Luc Van Gool, Pose Guided Person Image Generation, CSCV, Janauary 2018.