Motion Transfer

Hsuan Ouyang, Jiaming Liu, Sheng Xu, Zichen Li

March 2, 2020

Abstract

The objective of this project is to research on the task of motion transfer and to present a combination of generative adversarial networks (GANs) in order to achieve better video-to-video translation result. We propose a motion transfer 3-G pipeline composed of three GANs: a full-body GANs, followed by two specialised generative models, face and deblur GANs to address the distortion issue around face and foot region of target subject and to reduce the blur in the background of the synthesised video.

1 Introduction

Motion transfer offers an efficient method for animation, movie and game making. Given a source video of a person, a motion transfer model could extract the poses of the person and synthesize another video where a target person, animation or game character could imitate the same poses. Unlike motion capture, motion transfer does not need special equipment or software. Motion transfer also enables common people to perform difficult dances or movements like professionals.

Recently, generative adversarial network (GAN) has been claimed to be the best approach to perform motion transfer tasks in many publications. The most common approach is generally a two-phase process: image-to-pose detection and then image synthesis based on detected poses. However, we found that in most papers, the generated video still has blur and distortion problems, especially in the face area and around the edges like fingers and feet. Another issue is that most of the models take a long time to train.

In this work, we aim to research the task of motion transfer and to present a model combination to achieve better synthesis results with relatively shorter training time, especially to address the blur issue and the distortion problem around face area in the synthesized video. To achieve the objectives, we extracted pose from target video we shot using the latest OpenPose [1], trained the full body GAN model to generate high resolution images of target subject from pose figures, and added face GAN and deblur GAN to optimise results.

2 Related Work

2.1 Face & Motion Transfer

One study implemented a Liquid Warping GAN with Liquid Warping Block (LWB) that propagates the source information in both image and feature spaces, and synthesizes an image with respect to the reference. The model generated background, generated target people, and synthesized the motion separately and then combined the three streams.[4]

One study achieved their goals with GAN model, and divided the work into three stages, which are pose detection, global pose normalization, and mapping from normalized pose stick figures to the target subject. During pose to video translation, they also integrated temporal smoothing and Face GAN to enhance video quality.[2]

Another paper synthesized faces from poses using conditional GAN based on the pix2pix method. They improved the original pix2pix framework by building the generator with two parts: local enhancer and global enhancer. They first trained the global enhancer on low-resolution images, then added a local enhancer and trained the two networks on high-resolution images.[6]

2.2 Pose Estimation

For pose estimation, one study used Part Affinity Fields(PAFs) to represent the locations and orientation of multi-person 2D poses.[1] Another study focused on hand poses detection, and implemented multiview bootstrapping training: removing outliers triangulated by multiview geometry from produced noisy labels and using left labels to iterate training.[5]

2.3 Motion Deblur

For motion deblurring, Kupyn et al. designed and trained a conditional GAN with a loss function combining perceptual loss (which was L-2 loss between CNN feature maps of sharp target images and generated deblurred images) and WGAN-GP loss, to sharp blurred images.[3]

3 Method

3.1 Baseline Model

Our baseline model architecture is from the ICCV 2019 paper 'Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis'. The model is a combination of three separate modules which are body mesh recovery, flow composition, and liquid warping GAN.

The first body mesh recovery module will take both the source and reference image as its inputs, and obtain the body kinematic pose features of each picture. This is achieved by firstly getting the 2048 dimensional embedding of

the image using a pretrained ResNet-50 and then parsing the embedding to a SMPL regression network to get the pose and shape parameters. Secondly, the obtained parameters would go through a flow composition module to calculate the correspondences between the parameters of source image and reference image. In this stage, we also separate the image into front image and background image. Lastly, we use a Liquid Warping Block which contains three parts. The first part is a generator which takes the background image and color information as inputs and reconstruct the background image by filling in colors to the unseen pixels. The second part is a CNN auto-encoder that extracts the high-level features of the foreground of the source image. The third part is another generator that combines the encoded source foreground image and target foreground image to synthesize the final pose image.[4]

After implementing and experimenting the baseline model, we found the model did not give us satisfactory result (as shown in the figures in Experiment section). Therefore, we propose a generative adversarial models combination to produce better image-to-image motion transfer: Full-body GAN, Face GAN, and Deblur GAN.

3.2 Proposed 3-G Pipeline

Instead of further complicating the model architecture, which are already sophisticated enough, our experiments focused on a different direction. The key idea is to simplify the model architecture used during each stage but add more consecutive stages to the pipeline. Each of the GANs in our proposed model combination has different focuses and operates on the result of the previous stage. Such a longer, but much more compact pipeline improved the synthesized image quality with less amount of training time required. Also, different from our baseline model, we continued with the classic approach, using pose extracted from target subject as the intermediate representation of body structure. For pose detection, we found that OpenPose [1] can produce the best 2D full-body estimation. Especially, the latest version of OpenPose significantly enhanced the hand and foot detection, which could help reduce the distortion problem around hand and foot area, typically found in related works.

Our proposed network works as shown in Figure 1. We first feed the detected pose after normalization and the real images of our target subject to train the Full-body GAN. This stage focuses on global pose imitation. Until the generator can synthesize the target subject image based on pose estimation, we start to train the Face GAN to optimize the synthesis effects around face area. Then, we use the generated images, which typically have blur in the background, as input to train the Deblur GAN, which outputs final synthesized image. Last, the transferring process is shown in Figure 2, we use the trained network to transfer the dance movements in the source video to our target subject.

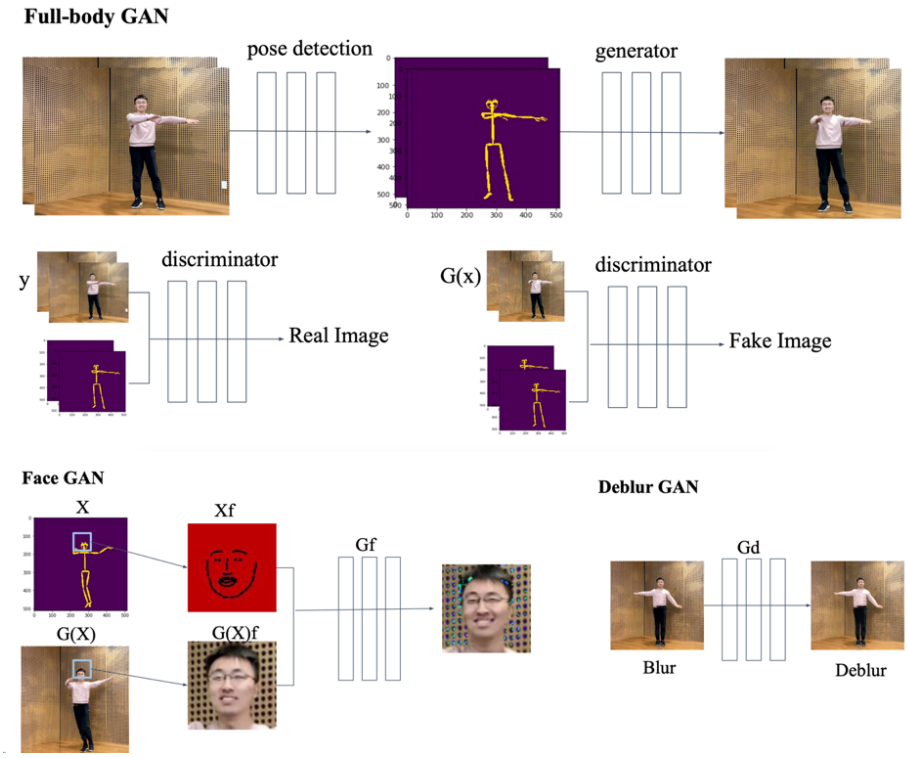

Figure 1: Training 3-G pipeline (full-body GAN, face GAN, deblur GAN)

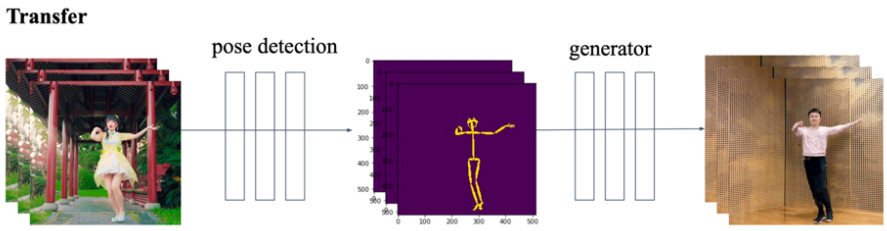

Figure 2: Motion transfer from source video to target subject)

3.3 Full-body GAN

The implementation of full-body GAN is based on the pix2pixHD model architecture [6]. The generator network is composed of a global generator and a local enhancer as shown in Figure 4. The global generator operates on higher resolution images whereas the local enhancer operates on lower resolution images. We plugged this network into our pipeline in two different ways to compare the

results: (1) only the global network; (2) global and local enhancer network. For the second way, we follow the same training technique as proposed in the original paper. First, we train the global generator and then train the local enhancer with the global generator frozen, then jointly fine-tune together before entering the next stage of the pipeline. The generator is trained against multi-scale discriminators, which have the same architectures but work on different scales of generated images. With the multi-scale discriminator setting, the generator is best trained for our task, which requires both global pose imitation and human body characteristics reproduction in finer details.

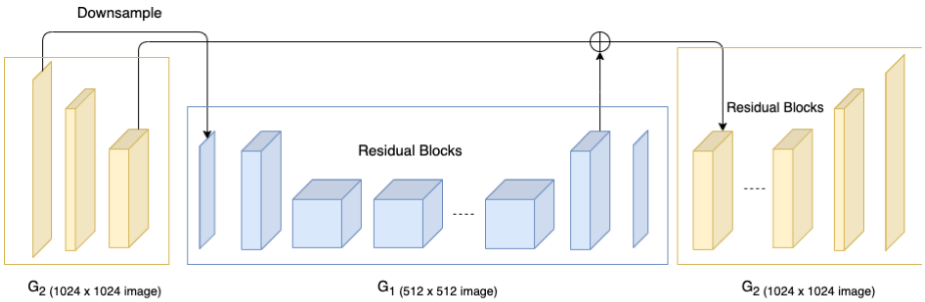

Figure 3: Full-body GAN generator architecture (pix2pixHD [6])

Since our objective is using image-to-image translation to ultimately realize video-to-video motion transfer, we need to make sure the smooth transition from frame to frame. Through our experiments, we found that train the generator to produce one frame at a time does not produce desired synthesized video terms of smoothness. Therefore, we tried using two and three consecutive frames extracted from our target subject basic movement video as input to train the generator. It turns out that using two images at a time produces the best visual effects.

The training objective of full-body GAN has three components: the basic GAN loss, feature matching loss [6] of all discriminators, and perceptual VGG loss. The full objective function is given as following.

$$\min_{\mathbf{G}}((\max_{\mathbf{D1,D2,D3}} \sum_{k=1,2,3} \mathcal{L}_{GAN}(G, D_k))$$

$$+\lambda \sum_{k=1,2,3} \mathcal{L}_{FM}(G, Dk) + \lambda_p(\mathcal{L}_p(G(x_{t-1}), y_{t-1}) + \mathcal{L}_p(G(x_t), y_t)))$$

3.4 Face GAN

The next stage of our proposed pipeline is to optimize the facial characteristics of the synthesized image generated in the previous stage. Using the same model structures as full-body GAN, we crop a smaller section of the pose figures and

generated images centered around face area, which are fed into the generator as input to predict the face region of the real target image (as shown in Figure 1). The final output is the addition of face residual predicted by Face GAN and the cropped face region predicted by the full-body GAN. The training objective for Face GAN is the basic GAN loss and the perceptual VGG loss as shown in the following function.

$$\min_{\mathbf{G}_f}((\max_{\mathbf{D}_f} \mathcal{L}_{face}(G_f, D_f)) + \lambda_p \mathcal{L}_p(r + G(x)_F, y_F))$$

3.5 Deblur GAN

To further enhance the resolution and fluency of the motion transfer result, we implemented Deblur GAN proposed by Kupyn et al. We feed the output of the face GAN model, which are slightly blurred images, to the Deblur GAN model, before converting them to final video result.

The Deblur GAN generator consists of two convolution blocks, nine residual blocks, and two transposed convolution blocks. The loss function consists of two parts, which are adversarial loss and content loss. Instead of using vanilla GAN objective, the author used WGAN-GP as the adversarial loss in the training process, which was more robust. To avoid the problem of blurry artifacts, the author chose perceptual loss as the content loss, which performed much better than MAE and MSE. The main purpose of the adversarial loss is to restore textual details, while the content loss is to restore general information. [3]

Table 1: Data set summary

Name	Quantity/Size	Source
iPER	1.2 GB	(Liu, Piao, Min, Luo, Ma, & Gao, 2019)
Source short dance videos	4	Source short dance videos
Target basic movement video	1	Shot by team
Target subject images (512 * 512)	60GB	Extracted from target video
Target subject images (1024 * 1024)	220GB	Extracted from target video

4 Data sets

For the baseline model, we used iPER data set provided by Shanghai Tech University, our own pictures, and dancing videos on Bilibili as training and testing data. We will only select random pose videos in the iPER data set. There are 103 short videos of random poses performed by 30 actors in total. [4]

The data set we used in the final model includes two parts: target person images for training and source dance videos for motion transfer. Target subject images are from the target basic movement video performed by one of our team members. Source dance videos are from Bilibili and Youtube. Each video are performed by one dancer with fixed background.

Table 1 shows the summary of the data set that we use in our models.

Figure 4: Source images (Japanese Style Dance)

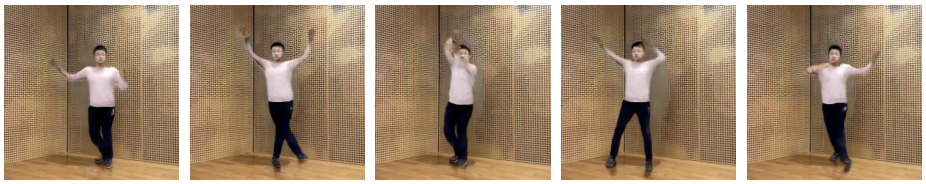

Figure 5: Baseline results (Japanese Style Dance)

Figure 6: Our results (Japanese Style Dance)

5 Experiments

Setup To train our model, and compare the result with the baseline model, we firstly prepared and processed our data to satisfy the requirement of the

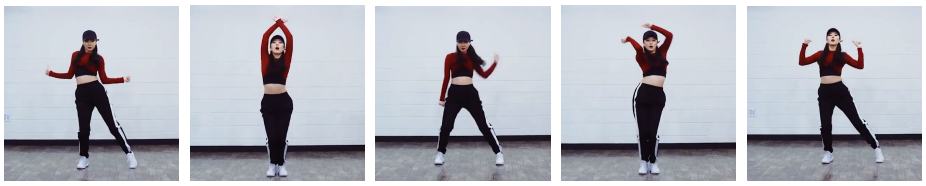

Figure 7: Source images (Korean Style Dance)

Figure 8: Baseline results (Korean Style Dance)

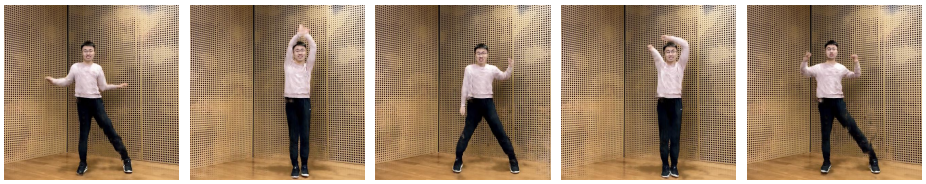

Figure 9: Our results (Korean Style Dance)

two models. We collected source videos from YouTube and Bilibili and then converted them into two sets of frames. The first one contains frames of *512*x*512* resolution, and the second contains *1024*x*1024* frames. We use the target video shot by ourselves, which consists of different poses of one of our team members.

Baseline method - Liquid Warping GAN This method takes a single target image, and frames of source pose to produce corresponding imitated images. The final video is generated by merely concatenating each frame into motions.

Our method - OpenPose+Simplified Pix2PixHD+Temporal Smoothing+Face GAN+Deblur GAN In this experimental model, it firstly synthesizes the smoothed images using the skeletons generated by the latest OpenPose and optimizes the images to keep the consistency between the current frame and the last frame. Secondly, the Face GAN model optimizes the face to contain more details of the facial expression. Lastly, the Deblur GAN model tries to enhance the quality of the images by reducing jitter and blur.

Evaluation Metrics We use both qualitative and quantitative methods to evaluate our models. The major quantitative evaluation metrics that we use

are **SSIM** Structural Similarity and **MSE** Mean Squared Error.

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad [7]$$

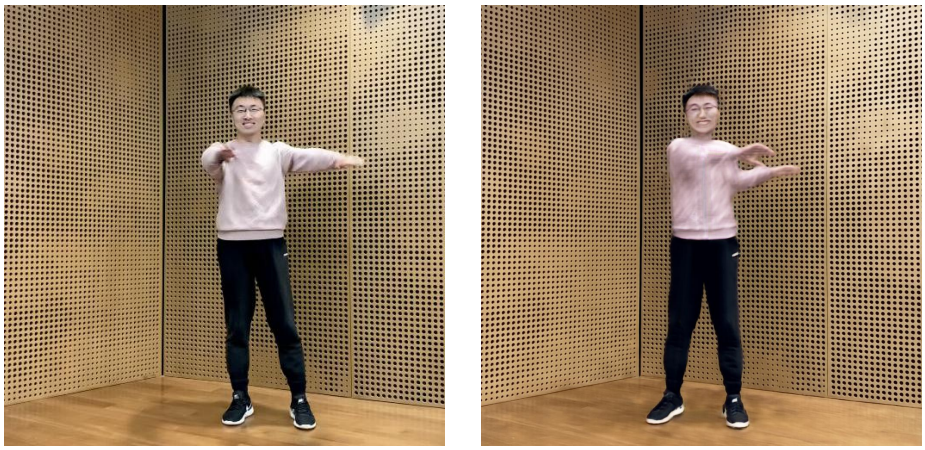

Figure 10: Global Generator vs Local Enhancer

We trained the baseline model for 20 epochs using *512*x*512* target frames as it is the maximum resolution that the model supports. For our improved model, we first generate the corresponding poses using the latest OpenPose. Then we trained the global Pix2PixHD network for 25 epochs using *512*x*512* frames. We also experimented with training the local enhancer network for 22 epochs using *1024*x*1024* frames after 6 epochs of the global network to refine the generator. However, the quality of the synthesized image was worse than only training the global network, as shown in Figure 10. So we decided only to use the global network as our generator. Lastly, we trained the Face GAN network for 8 epochs. For the Deblur GAN, we used the pretrained model weights that are provided by the original author.

As our baseline can only generate output resolution of *256*x*256*, and SSIM need the same resolution for both images. We manually rescale our synthesized image from *512*x*512* to *256*x*256*. We compare the improved model to the baseline model based on the same target video. The result is showed in Figure 11. Our improved model got not only a higher SSIM but also a lower MSE, which means that our result has a better imitation result on poses and maintains more information on details.

Moreover, by qualitatively judging the generated video on the same source frames, our improved model got higher resolution and smoother frames when switching motions compared to the baseline model. Also, the facial expression is more explicit and maintains the facial information of the source dancer in most of the frames.

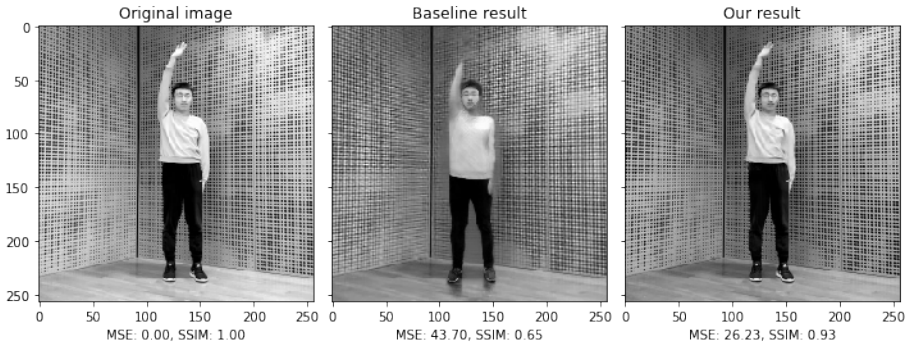

Figure 11: SSIM and MSE Comparison

6 Conclusion

In this work, we proposed a 3-G pipeline for the task of motion transfer, including full-body GANs, face GANs, and deblur GANs. Our proposed solution can achieve a surprisingly satisfying video-to-video motion transfer result with a simplified model structure, especially considering the limited computational resource and training time we have.

7 References

[1] Cao, Z., Simon, T., Wei, S., & Sheikh, Y. (2018). Realtime multi-person 2D pose estimation using part affinity fields. 2018-, 1302–1310.

[2] Chan, C., Ginosar, S., Zhou, T. & Efros, A., (2019) Everybody dance now. ArXiv, abs/1808.07371v2

[3] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin J. Matas, (2018) DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks.ArXiv, abs/1711.07064.

[4] Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., & Gao, S. (2019). Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis. ArXiv, abs/1909.12224.

[5] Simon, T., Joo, H., Matthews, I., Sheikh, Y. (2017). Hand keypoint detection in single images using multiview bootstrapping. 2017-, 4645–4653.

[6] Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B. (2018). High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. ArXiv, abs/1711.1158

[7] Wang, Z., Bovik, A., Sheikh H., Simoncelli, E. (2004) Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004.