Applying Generative Models for Pose Transfer

Prathik Srinivasan

The University of Texas at Austin

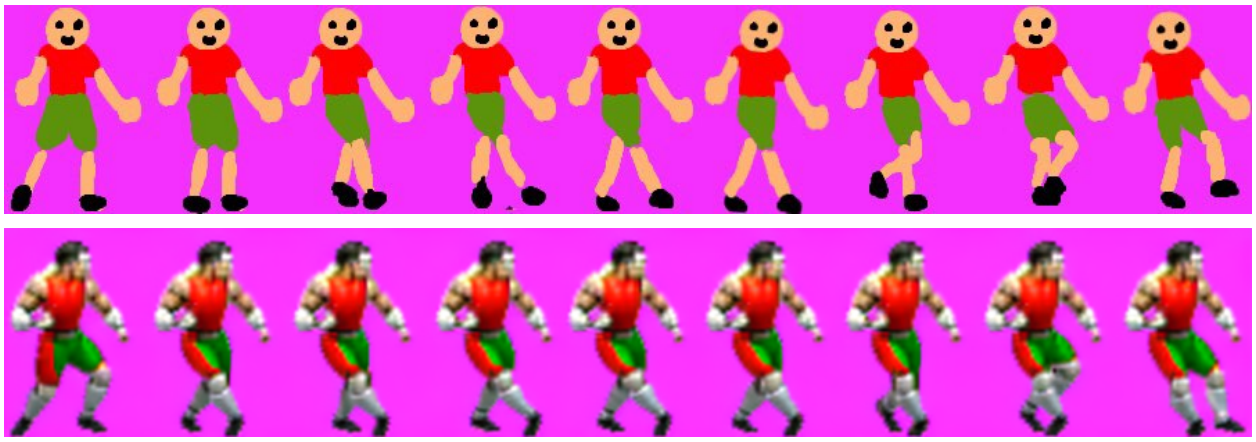

Figure 1. Source image (above); Diffusion model image-to-image transfer (below)

ABSTRACT

The primary goal of this project is to explore methods of generating target animations of human subjects using single images and target poses as inputs. This is extended to a generation of a higher-fidelity transfer as seen in figure 1, where a series of guiding poses are used as a basis for a higher-quality image generation. This paper explores two different methods of attempting the generation of novel poses based on a source image and target pose, and their effectiveness as applied to an animated sequence.

1. INTRODUCTION

In this paper, I explore various methods of achieving pose transfer, the process of using a source image and a target pose to generate a result that modifies the subject's pose to resemble the target while maintaining the original appearance [1]. The two methods explored are drastically different in implementation, but both work towards a general goal of using target poses to generate an animation from a static image of a person.

Pose transfer is a challenging task, as it requires a combination of two metrics to be considered successful - a resemblance in appearance to the source image, and a resemblance in pose structure to the target. Certain applications may also involve reconstruction of the image background, but this was not within the scope of this exploration. To achieve pose transfer between two subjects, the two methodologies explored in this paper were training a generative model and fine-tuning a pre-trained diffusion model.

2. RELATED WORK

2.1. Generative Adversarial Networks

Generative Adversarial Networks, or GANs [2], are networks that generally consist of two primary components, a generator and discriminator. In the training process, the generator attempts to mimic the images within the training dataset, while the discriminator attempts to distinguish real images from the dataset from 'fake' generated images. The training loss is dependent on the discriminator's

success in identification, which is then fed back into both systems in each iteration. Typically, CGANs or Conditional GANs are effective in generating images with specific constraints, such as specific poses or viewpoints. A variety of efforts have been made to achieve pose reconstruction using GANs, such as combining VAE and U-Net [3] to separate appearance from pose and reconstruct either separately. However, the most successful example was by *Zhu et al*. [4], who have created a unique system using *Pose Attention Transfer Blocks* (PATBs) to progressively adjust masks based around specific key points in the image.

2.2 Diffusion Models

Diffusion Models operate differently than GANs, despite having a common goal of acting as a generative image model. At the core of diffusion models is a denoising system, where the model learns how to best remove gaussian noise from an image based on a set of parameter keywords [5]. During the training process for a diffusion model, images annotated with keywords have gaussian noise added iteratively until the original image is completely unrecognizable, a process referred to as reverse diffusion. In the forward diffusion process, the network learns how best to denoise the image using the related keywords as parameters, attempting to achieve a result close to the original image. Once the process is complete, the model is capable of using the weights associated with keywords to apply a denoising process to random noise and generate an output image [6]. Stable Diffusion is one of the currently most popular pre-trained diffusion models, used in a variety of applications and as a basis for several products. It is trained on the LAION-5b dataset.

3. METHODS

3.1 CGAN Model

To create the generative model, I followed a similar process to that of *Zhu et al*. [4], using PyTorch to create the CGAN architecture shown in figure 2. The condition image P_c and the condition pose heatmap S_c are encoded with 2 downsampling convolutional layers. The PATN or *pose-attentional transfer network* consists of T PATBs (T = 9 in my case) that guide the pose transfer by instructing where to sample attention mask areas from based on the condition and target pose. The output is then decoded through 2 deconvolutional layers.

The model training process involves two discriminators to determine the loss produced by the generator. The appearance Discriminator D^A and shape discriminator D^S feed the output

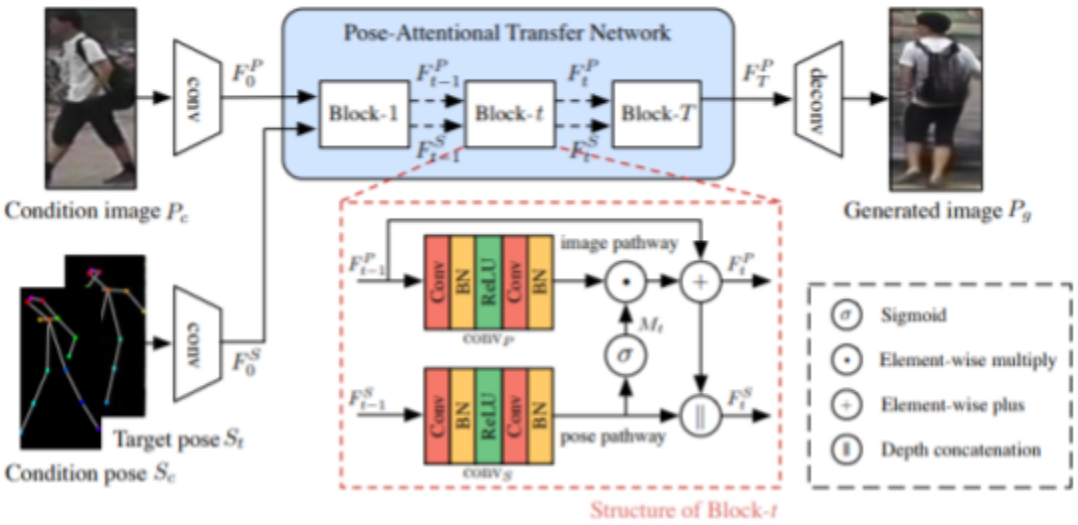

Figure 2. Architecture of GAN generative model; via *Zhu et al.* [4]

image P_g into a CNN to determine the likelihood of P_g containing the same person as input P_c, and the likelihood of P_g containing the same pose present in S_c. These appearance consistency and shape consistency scores are represented as R^A and R^S respectively, with the final score R used for the loss function of the generative model computed as the product of these two scores. The model was trained for 90k iterations on each dataset, creating two networks applicable to the Market-1501 and DeepFashion datasets [8, 9].

3.2 Fine-tuned Diffusion Model

To create the fine-tuned diffusion model, I followed the DreamBooth training process, due to its quickness and versatility, able to introduce a concept with only a handful of training set images. The DreamBooth method typically involves 3-5 images with a unique classifier (one as of yet unrepresented within the model itself) as well as a classifier that represents the category of the subject or concept [7]. This process can be used to introduce both specific subjects (e.g. a specific person or pet) to a model as well as a specific style (e.g. vincent van gogh's painting style) such as in figure 3, where a model fine tuned with 4 Vincent Van Gogh paintings was prompted with *Ice Cream Sundae drawn in the style of <Van-Gogh>*. DreamBooth is able to achieve more diverse instances of the subject with a higher level of fidelity due to its paired process, where the unique identifier and class-specific identifier are trained in parallel.

Figure 3. Fine-tuned diffusion model output

To implement pose transfer using DreamBooth, I began with the v1.5 Stable Diffusion pretrained model. Rather than create an identical condition as was present in the GAN system, I decided to attempt a specific type of pose transfer that was more relevant to my overall goal of applying pose transfer to create an animated sequence. As my training data for fine tuning, I used a series of walk cycle sprite sheets from the original *Mortal Kombat* games, as they use a Full-Motion Video sprite style that I wanted to emulate in my results. These images were labeled with the unique label '*walkcycle*' and fed into the fine tuning process, trained for 3000 iterations.

Once a fine-tuned model integrating the concept and style of my original dataset was complete, I experimented with further fine-tuning to see if I could use additional iterations to add subjects and perform true pose transfer between a subject and a series of poses. To achieve this, I used a series of images of myself, with my aforementioned fine-tuned model as a starting point. I repeated the training process to introduce myself as a subject, training for an additional 2000 iterations.

4. RESULTS

While it is possible to set quantitative metrics for evaluating the results of both the GAN and diffusion model, due to the specific nature of my overall goal with this study, I decided to embrace a more qualitative approach to evaluation. The reasoning for this was twofold - first, the quantitative validity and effectiveness of the GAN model and DreamBooth technique have already been evaluated by [4] and [7] respectively. Second, as my initial goal was uniquely generative in nature to begin with, I lack the objective target results to serve as a comparative basis for a quantitative metric.

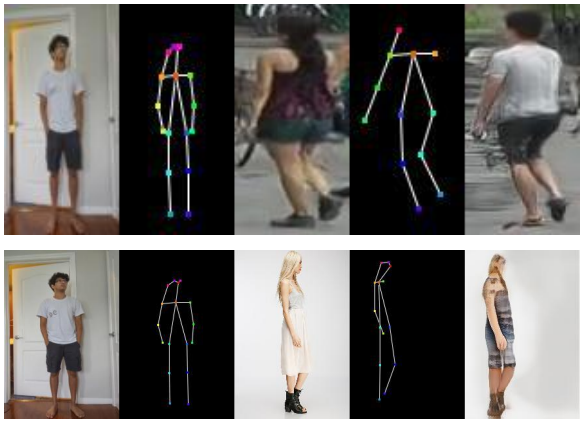

Figure 4. GAN pose transfer results from data-extraneous condition image

4.1. CGAN Results

Similar to the results found by *Zhu et al.* [4], the GAN model was able to accurately reshape poses when operated on condition images within the dataset itself. However, a phenomenon observed within both the Market-1501 and DeepFashion versions of the model was their inflexibility with new data. When introduced to an extraneous condition image, the models generally failed to deliver a reasonable amount of appearance consistency, despite exhibiting reasonable amounts of shape consistency.

Figure 5. Fine-tuned Diffusion model results, prompted with text prompt “Superman walkcycle”

Figure 6. Original image (above); Result of image-to-image transfer with text prompt “Batman walkcycle” (below)

4.2. Diffusion Results

The diffusion model was naturally much more diverse in condition images, owing to its massively larger dataset, allowing for a much wider range of applicable data to be reshaped with the custom model. When prompted with only text inputs, the fine-tuned model was able to generate images that contained many elements reminiscent of the sprite sheets used in the training data, applied to the desired subject. Most notably, the repeated similar poses as well as the presence of specific pose features such as one leg extending in front of the other for a forwards step, as can be seen in figure 5. With image-to-image transfer, a process that uses an existing image as a source for randomized noise, the model was able to generate results that were extremely similar in structure to that of the training data. Results often were accurate to the point of consisting of the exact correct number of sprites present in the original image, with similar spacing between poses, as can be seen in figure 6.

Figure 7. Results of text prompt “Prathik walkcycle (above); Results of image-to-image transfer with text prompt “Prathik walkcycle” (below)

Once trained to introduce a specific subject to the model, the results of fine-tuned stable diffusion were significantly more accurate in pose transfer for a specific subject in comparison to the GAN model. As can be seen in figure 7, In both text-based generation and image-to-image transfer, many key appearance consistencies (e.g. purple clothing, dark skin) are present within the generations, while the unique

structure of the walk cycle concept is maintained.

5. REFERENCES

[1] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In Proc. NIPS, pages 405–415, 2017.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proc. NIPS, pages 2672–2680, 2014.

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

[4] Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., & Bai, X. (2019). Progressive pose attention transfer for person image generation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[5] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18208–18218, 2022.

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B.. (2021). High-Resolution Image Synthesis with Latent Diffusion Models.

[7] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K.. (2022). DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation.

[8] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable Person Re-identification: A Benchmark. In Computer Vision, IEEE International Conference on.

[9] Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10] TheLastBen, “fast-stable-diffusion Colabs” [Online], 2022. https://github.com/TheLastBen/fast-stable-diffusion