Person Video Synthesis using Denoising Diffusion Model

Adeel Yousaf
adeel.yousaf@knights.ucf.edu

Prudvi Kamtam
prudvikamtam@knights.ucf.edu

Mukund Dhar
mukund.dhar@knights.ucf.edu

Manu S Pillai
manu.pillai@knights.ucf.edu

Abstract

Person image synthesis is the task of generating realistic images of humans in various poses using pose guidance. There are several works involving generative adversarial networks, which can struggle to maintain realistic textures and handle complex deformations and occlusions. In this work, we extend the work titled "Person Image Synthesis via Denoising Diffusion Model" (PIDM) for person video synthesis. Specifically, in this work, we try to synthesise fashion videos using a pretrained person image diffusion model guided by a sequence of poses and a style image as in PIDM. We propose a temporal texture diffusion block that is an extension to the texture diffusion block proposed in PIDM. We add a self attention block on the space and time dimension separately that allows the UNet model to aggregate information from neighbouring frames to generate frame sequences that are temporally consistent and robust to artifact effects.

1. Introduction

The goal of the task of Pose-guided person image synthesis is to create an image of a person that has a specific pose and appearance. The appearance of the person is based on a source image, while the pose is defined by a set of keypoints. The problem of person synthesis is typically approached in the literature using Generative Adversarial Networks (GANs), which attempt to produce an image of a person in a specific pose using a single forward pass. However, it is a difficult task to maintain a consistent structure, appearance, and overall body composition when generating the image in one step. As a result, the generated images often suffer from distorted textures and unrealistic body shapes, particularly when dealing with occluded body parts. The work by Bhunia et. al mitigates the issues with GAN based method and proposes a probabilistic diffusion based method for Person Image synthesis.

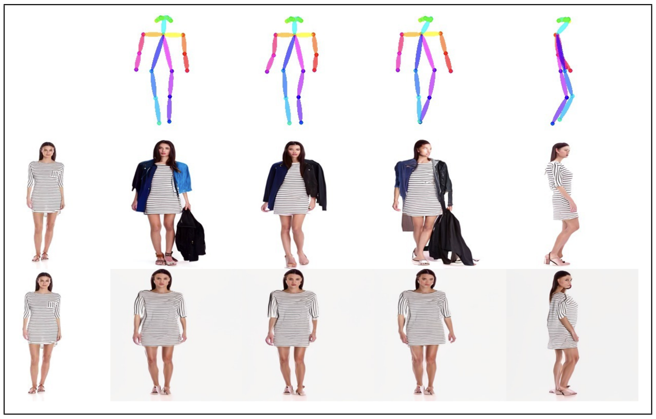


Figure 1. Sequence of frames generated using PIDM (second row) given the poses (first row) and style (first column). Here we can see the artifact effects in play, the generated frames have hallucinated objects that are not present in the style guidance and are not consistent across frames. Frames synthesized with the proposed method (second row). Its evident that the proposed method removes the artifacts and is able to generate consistent sequences.

In this work, our objective is to generate fashion videos by employing a pre-trained person image diffusion model that is directed by a sequence of poses and a style image, similar to the approach used in PIDM. To achieve this, we introduce a new component called the temporal texture diffusion block, which builds upon the texture diffusion block utilized in PIDM. We augment this by incorporating a self-attention block separately on both the space and time dimensions. This modification enables the UNet model to collect information from adjacent frames, resulting in a video sequence that is both temporally coherent and resistant to distortion/artifact effects.

2. Related Work

2.1. Diffusion Models

Diffusion models are a class of generative models that have gained significant attention in recent years due to their ability to generate high-quality samples and perform image synthesis tasks. These models operate by iteratively refining a noise source through a series of stochastic diffusion steps, where the noise is transformed into the final image. The diffusion process can be reversed to generate samples by iteratively refining the image until it becomes noise.

One of the advantages of diffusion models is their ability to model long-range dependencies in images, which can be challenging for other generative models. Additionally, diffusion models can generate high-resolution images with a high degree of fidelity, making them suitable for synthesis tasks such as image super-resolution and inpainting.

2.2. Person Image Synthesis via Denoising Diffusion Model

In recent years, there has been significant progress in the field of person image synthesis, especially with the use of GAN-based models. One common problem in this field is human pose transfer, which has been extensively studied. Previous approaches have used various methods to address the issue of feature misalignment caused by concatenating the source image, source pose, and target pose as inputs. These methods include VAE-based design [5], UNet-based skip connections, deformable skip connections, flow-based deformation, geometric models, and progressive transformation.

In recent works, a diffusion-based framework named PIDM [3] for pose-guided person image synthesis has been introduced. Diffusion models have shown success in unconditional generation and have been extended to work in conditional generation settings, demonstrating competitive or even better performance than GANs. The pose transformation process is broken into several conditional denoising diffusion steps, each of which is relatively simple to model. This diffusion-based approach offers a different way of modeling the complex transformation of pose. By breaking the process into simpler steps, it achieves high-quality image synthesis results while maintaining computational efficiency. The use of texture diffusion block and disentangled classifier-free guidance also offers tighter alignment between different aspects of the input and output images, which can lead to more accurate and realistic results.

2.3. Pose Guided Human Video Generation

Pose-guided video generation is a challenging task that aims to generate realistic videos of a person performing various actions based on a single input image and a pose condition. While there have been several works in conditional video generation, the most well-studied task is future frame prediction, where the goal is to generate future frames of a video given previous frames as input. However, recently, researchers have focused on more complex tasks such as video-to-video translation, where the goal is to translate one video into another based on a given condition.

One approach to pose guided video generation is to use intermediate representations, such as learned key points, to guide the generation process. Two recent works, [4] and [7], have suggested pose-guided video generation, where a separate network is trained for each person. However, this approach is not scalable and may require extensive training data to cover various poses and actions.

Another recent work by Siarohin et al. [6] tries to learn a representation of a subject in an unsupervised manner for video generation. However, the results obtained are suboptimal due to the lack of supervision in the learning process. Another work on video generation is the Dense warp-based network [8], a GAN-based architecture, that generates videos iteratively leveraging the dense intermediate pose-guided representation to warp the required appearance from a source image into a desired pose.

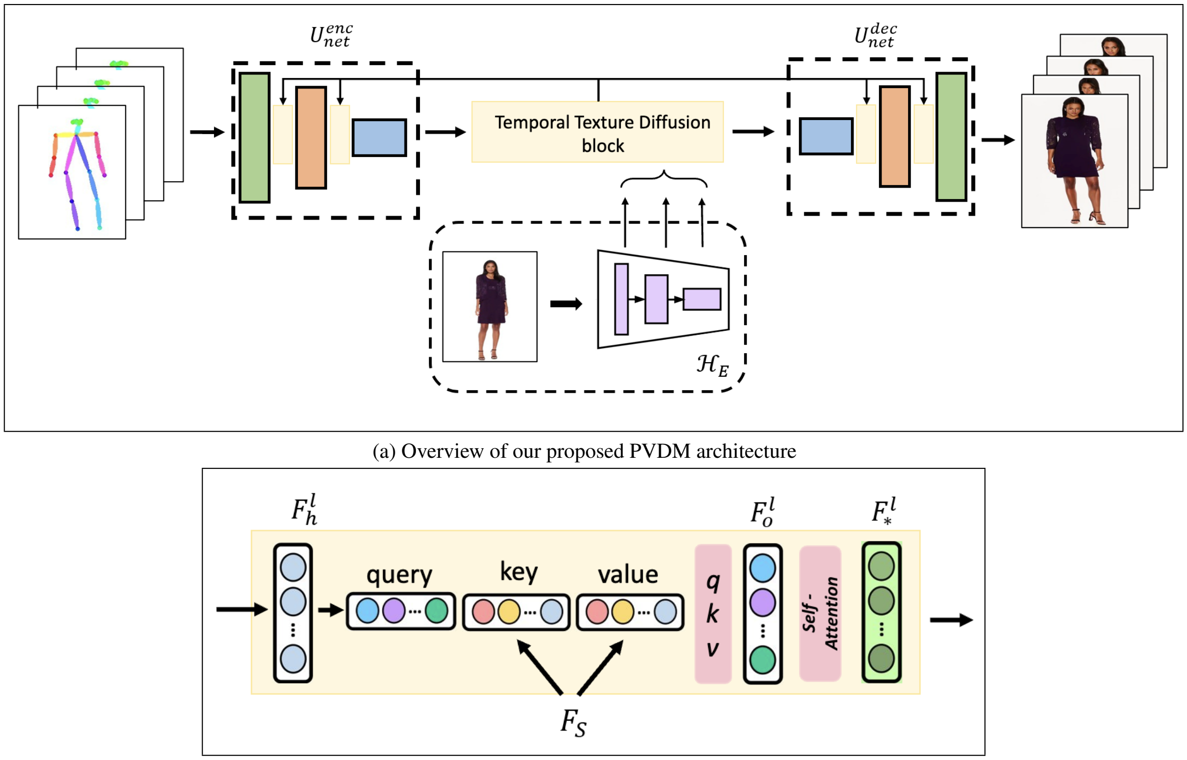
We aim to generate videos given an input-style image conditioned on the set of poses achieved using the diffusion process similar to the PIDM model for image synthesis. The difference is the newly introduced temporal texture diffusion block that takes care of the temporal inconsistencies. We show that with the attention provided within this block, the model achieves outstanding video generation results.

3. Methodology

Fig.3a shows an overview of the proposed generative model. Our objective is to train a conditional diffusion model $p_{(y|x_s,x_p)}$ where the source frame x_s and a sequence of target pose x_p are given. The aim is to generate a final output video sequence y that not only possesses the same style as the source frame x_s but also matches the target pose sequence requirement i.e. follows the same motion as given in the pose sequences. We start by summarizing PIDM [2], which is the main motivation and baseline for our work.

3.1. Overview of Person Image Diffusion Model (PIDM)

The denoising network θ used in PIDM [2] is based on a UNet architecture and comprises two main components: a noise prediction module H_N and a texture encoder H_E. The texture encoder H_E is responsible for encoding the texture patterns found in the source image x_s. In order to obtain multi-scale features, authors extract output from various layers of H_E, resulting in stacked feature representations $Fs = [f1, f2, ..., fm]$. To effectively transfer rich multi-scale texture patterns from the source image distribution to the noise prediction module H_N, authors utilize



(a) Overview of our proposed PVDM architecture

(b) Proposed temporal attention block

Figure 2. Proposed PDVM framework

Texture Diffusion Blocks (TDB). This approach allows the network to take full advantage of the similarities between the source and target appearances, resulting in images that are free from distortions.

3.2. Temporal Consistency

One way to address the temporal aspect is to use PIDM [2] frame-by-frame. However, this approach has a major drawback as it ignores the fact that the frames are interconnected and results in output that is temporally inconsistent which leads to artifacts as discussed in the experiment section. To address this issue, we propose a simple extension of PIDM [2] that incorporates an attention block within Texture Diffusion Block (TDB) as shown in Fig.3b.

First, the sequence of poses is passed through the UNET encoder to get the pose features for each frame. Then, the reference frame is passed through the texture encoder H_E for encoding the texture patterns. Then, the target pose features of each frame are cross-attended to the reference texture features using the standard TDB [2]. We introduced a temporal attention block within the Texture Diffusion Block (TDB). It takes cross-attended features from the standard TDB and applies self-attention between features from different time indices i.e. from different poses. The attention block makes sure that model has an explicit understanding that the input frames are connected with each other temporally. Our experimental results indicate that this attention mechanism significantly enhances the model's ability to maintain temporal consistency.

Various attention architectures are proposed recently in the literature [1]. In this paper, we propose two such variants a) spatio-temporal attention and, b) factorized attention. In spatio-temporal attention, the tokens from different spatial and temporal indices interact with each other jointly. However, in factorized attention, spatial and temporal attention is divided into two sequential steps. The first, is spatial attention, where tokens extracted from the same temporal index interact with each other. Then, in temporal attention tokens from different temporal indices interact with each other. Recently, literature studies [1] have shown that factorized attention is able to get better results with less number of FLOPS. And, our experiments also validate this finding.

4. Experiments

4.1. Dataset

For this paper, we utilized the Fashion Video Dataset to fine-tune the PIDM model into one that generates videos. This dataset contains 500 training and 100 test videos, each consisting of approximately 350 frames. The videos feature a single human subject, with a static camera providing a consistent visual context. One of the most notable aspects of this dataset is the diverse range of clothing and textures present, covering a large space of possible appearances. The high resolution of the videos ensures that fine-grained details of the clothing are captured and can be used to train a model effectively. The Fashion Video Dataset is publicly available and is an excellent choice for our purposes due to its similarity with the original dataset used in the diffusion-based image model. The dataset was obtained from the University of British Columbia's computer vision lab and can be accessed through their website at https://vision.cs.ubc.ca/datasets/fashion/.

4.2. Implementation Details

In accordance with the original paper, our PVDM model was trained with T = 1000 noising steps and a linear noise schedule. During training, we adopted an exponential moving average (EMA) of the denoising network weights with a decay of 0.9999. In all experiments, we used the Adam optimizer with a learning rate of $\alpha = 2 \times 10^{-5}$. For disentangled guidance, we used $\eta = 10$, and set the values of w_p and w_s to 2.0 for sampling.

When training our model, we followed the approach of using the style image from the first half of the clip, while the series of pose frames were taken from the second half of the clip. The source image and the source skeleton were extracted from the first frame of the video, while the target skeletons and target images were taken from the second half of the video. Due to computational constraints, we selected a clip length of 4 with a sample rate of 2. Additionally, we froze the entire model except for the temporal attention blocks that we introduced. As a result, our model had approximately 186 million parameters, with approximately 5.7 million being trainable parameters and 180 million being non-trainable parameters.

4.3. Qualitative Comparision

To ensure a fair comparison between the original model and our adapted model, we set all the sampling parameters to be the same for both models. We sampled the video as a batch, where the inputs to the model were the source image (the first frame of the video) and a batch of target skeletons from the second half of the video, applying the model frame by frame. For all qualitative comparisons, we used a batch size of 4 to simplify visualizations.

Our qualitative analysis focused on two versions of outputs: short-term (consecutive frames) and long-term (frames with a step size of 20). Upon analysis, we observed that the original model generated images that were well-aligned with the pose and style. However, we noticed inconsistent artifacts in the generated frames, indicating little temporal consistency. In contrast, our adapted model not only generated artifact-free frames but also produced highly consistent frames in both the short-term and long-term.

5. Conclusion

The goal of person image synthesis is to create realistic human images in different poses using pose guidance. Current approaches rely on generative adversarial networks, which struggle with complex deformations and occlusions, and maintaining realistic textures. This study builds on previous work titled ”Person Image Synthesis via Denoising Diffusion Model” (PIDM) by applying it to fashion video synthesis. Specifically, we aim to generate fashion videos by utilizing a pretrained person image diffusion model that is guided by a sequence of poses and a style image, similar to PIDM. To improve the model's performance, we introduce a temporal texture diffusion block that extends the original texture diffusion block from PIDM includes a self-attention block on the space and time dimension, which allows the UNet model in the backward process to aggregate information from nearby frames to produce frame sequences that are both temporally consistent and robust to artifact effects.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[2] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. *arXiv preprint arXiv:2211.12500*, 2022. 2, 3

[3] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model, 2023. 2

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now, 2019. 2

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation, 2018. 2

[6] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer, 2019. 2

[7] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2

[8] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation, 2019. 2

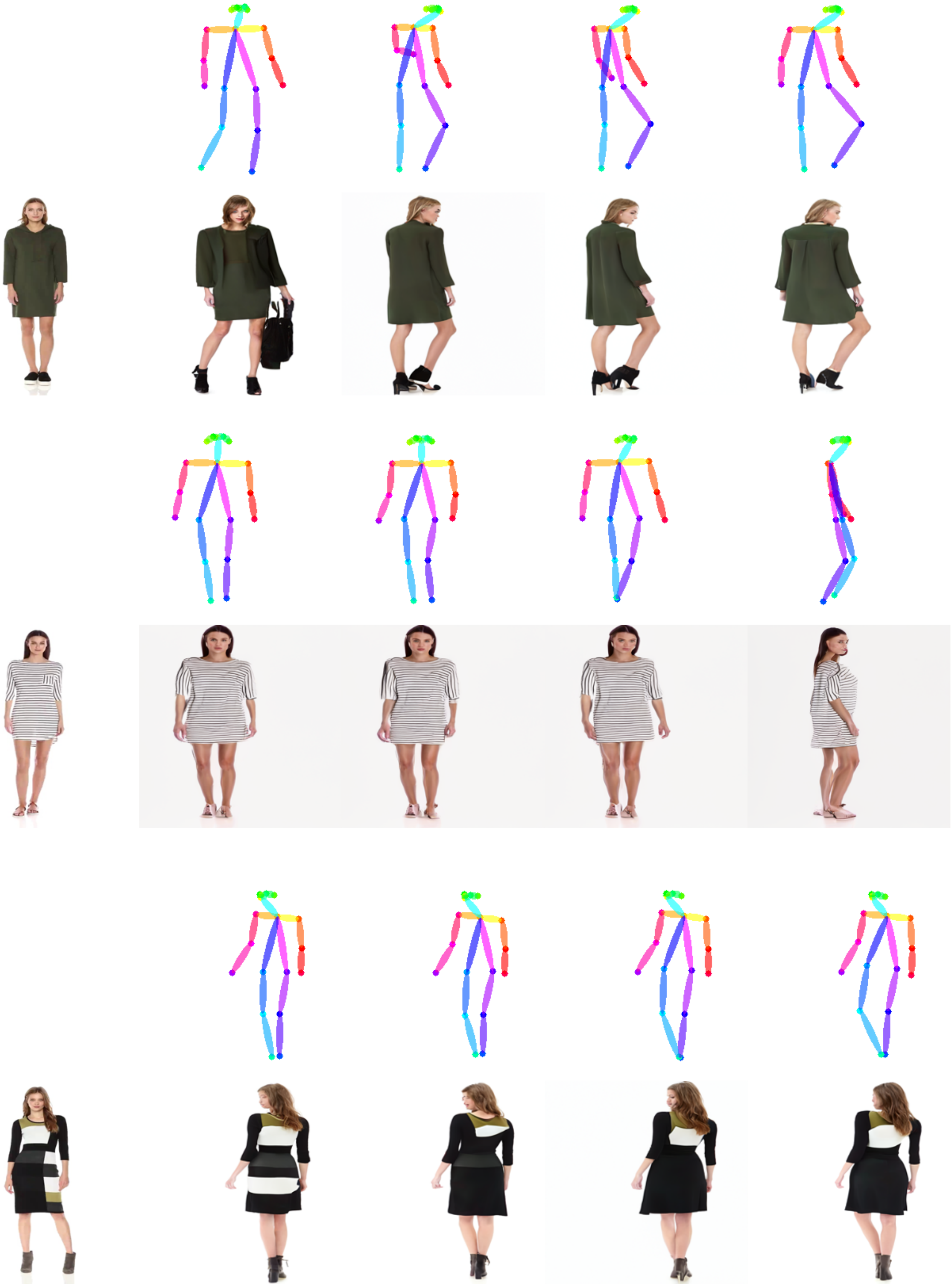


Figure 3. Qualitative results of our approach

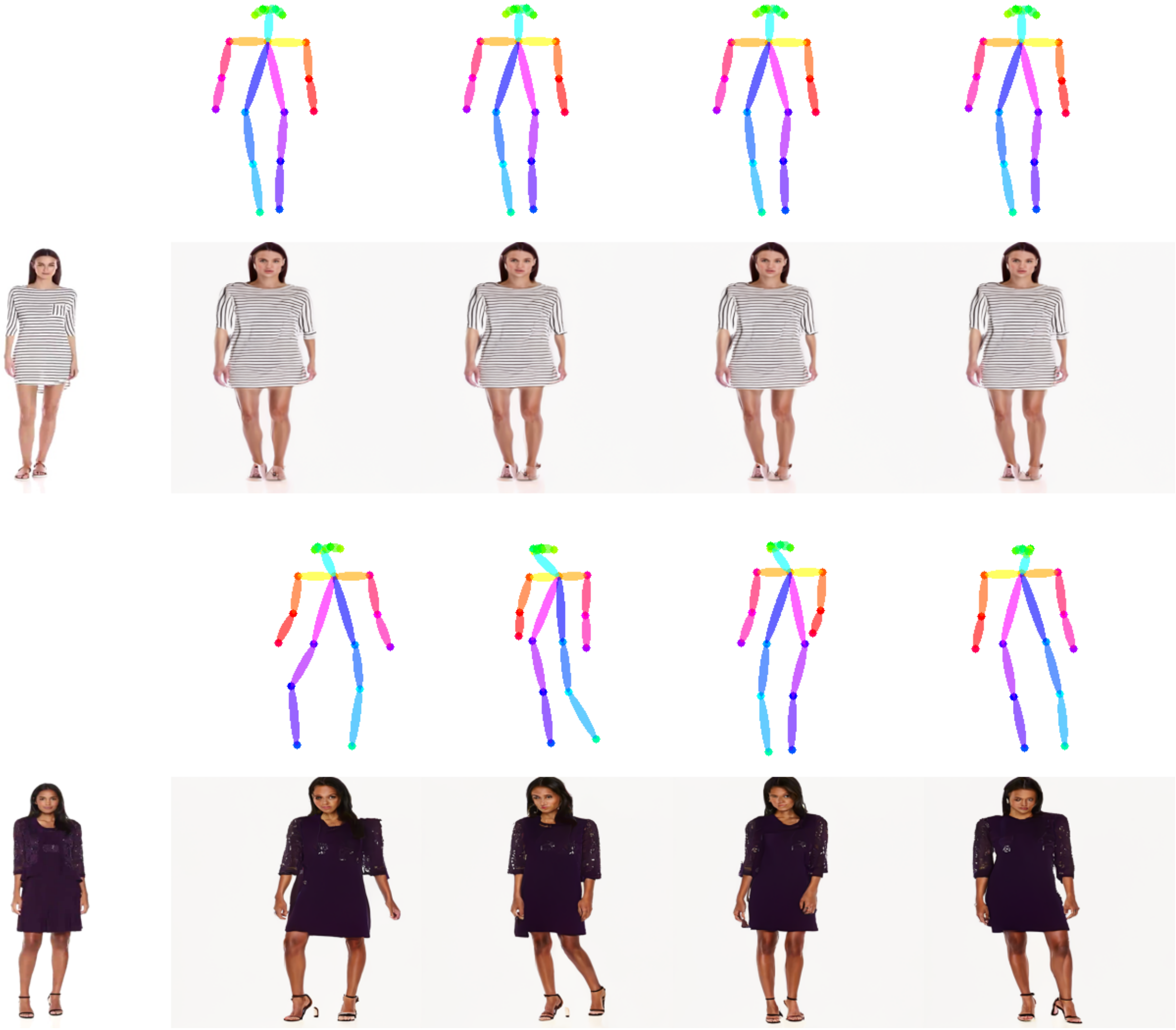


Figure 4. More qualitative results of our approach