Adaptive Appearance Rendering

by

Ruizhi Deng

B.Sc., University of Michigan - Ann Arbor, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

© Ruizhi Deng 2019
SIMON FRASER UNIVERSITY
Summer 2019

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.
# Approval

| Name:          | Ruizhi Deng                |
|---------------|---------------------------|
| Degree:       | Master of Science (Computing Science) |
| Title:        | Adaptive Appearance Rendering |
| Examining Committee: | Chair:  Mark Drew  
 Profesor |
|               | Greg Mori  
 Senior Supervisor  
 Professor |
|               | Yasutaka Furukawa  
 Supervisor  
 Associate Professor |
|               | Ping Tan  
 Internal Examiner  
 Associate Professor  
 School of Computing Science |
| Date Defended: | July 11, 2019 |
Abstract

We propose an approach to generate images of people given a desired appearance and pose. Disentangled representations of pose and appearance are necessary to handle the compound variability in the resulting generated images. Hence, we develop an approach based on intermediate representations of poses and appearance: our pose-guided appearance rendering network firstly encodes the targets’ poses using an encoder-decoder neural network. Then the targets’ appearances are encoded by learning adaptive appearance filters using a fully convolutional network. Finally, these filters are placed in the encoder-decoder neural networks to complete the rendering. We demonstrate that our model can generate images and videos that are superior to state-of-the-art methods, and can handle pose guided appearance rendering in both image and video generation.

Keywords: pose; appearance; disentanglement; generation
Acknowledgements

I would like to thank my advisor Dr. Greg Mori. He provided not only valuable suggestion for this thesis but advice on my research and career as a whole during my master program. I would also like to thank the other members of my defense committee, Dr. Mark Drew, Dr. Yasutaka Furukawa and Dr. Ping Tan for their assistance. Their help makes organizing the thesis defense event a smooth experience. I would like to give special thanks to my co-authors of the paper Adaptive Appearance Rendering published on BMVC 2018 upon which this thesis is based, Mengyao Zhai, Jiacheng Chen, Lei Chen and Zhiwei Deng. Finally, I want to thank all the other cooperators in my research projects during this master program.
Table of Contents

Approval ii
Abstract iii
Acknowledgements iv
Table of Contents v
List of Tables vi
List of Figures vii

1 Introduction 1

2 Related Works 4

3 Adaptive Rendering Network 7
  3.1 Network Structure ............................................. 8
  3.2 Loss Function ................................................... 9

4 Experiments 10
  4.1 Experiments on DeepFashion ................................... 10
  4.2 Experiments on Volleyball ..................................... 13

5 Limitations and Discussions 15

6 Conclusion 16

Bibliography 17

Appendix A FCN Structure 22
### List of Tables

| Table | Description                                      | Page |
|-------|--------------------------------------------------|------|
| Table 4.1 | Quantitative measures on DeepFashion.                | 12   |
| Table 4.2 | Quantitative measures on Volleyball dataset.            | 14   |
| Table A.1 | FCN used for DeepFashion dataset                     | 22   |
# List of Figures

| Figure  | Description                                                                 | Page |
|---------|-----------------------------------------------------------------------------|------|
| 1.1     | Introductory figure of adaptive appearance rendering.                       | 1    |
| 3.1     | Overview of the adaptive appearance rendering network.                      | 7    |
| 4.1     | Visualization of generation results of Ada-R using different loss terms on DeepFashion. | 12   |
| 4.2     | Visualization of generation results of different state-of-the-art approaches on DeepFashion. | 13   |
| 4.3     | Visualizations on Volleyball dataset.                                       | 14   |
| 4.4     | Visualizations on whole frame generation on Volleyball dataset.             | 14   |
Figure 1.1: We develop architectures for forecasting person images of various appearances. At the core of the method is adaptive rendering modules. A series of generation examples of various reference images rendered into new poses is shown above.

The problems of image and video generation have regained popularity in the machine learning and computer vision communities since the emergence of modern deep neural networks. In the machine learning community, deep generative models that can learn a meaningful latent space and map it to photo-realistic images have been actively studied since the introduction of generative adversarial networks (GANs) [9] and variational autoencoders (VAEs) [19]. Controllable conditional generation that can produce output consistent with the input conditions still remains an open problem in many tasks. It usually requires the model to disentangle different sources of variations in the input space and map them to correct outputs. Meanwhile another stream of deep generative model works focus on improving the visual quality of generated results as many early works can only
generate low-resolution and low-quality pictures. Such works [3, 29, 22] usually have sophisticated models and engineering details and could consume large amount of computing resources. When these studies are combined with natural language processing, approaches have been proposed for text to image synthesis or image to text synthesis [24, 31]. Works [27, 7] have also explored predicting the future consequence conditioned on the interaction between agent and environment in the literature of robotics and reinforcement learning. In the computer vision realm, generative models of video sequences [37, 38] and photo-realistic images [42, 22, 23] are a core part of visual understanding of high-level semantics and temporal dynamics. They have received renewed attention from the vision community recently. Such models could also enable modern AI systems to assist human with many creative works, including fashion design [42] and artistic style transfer [8, 14].

Imagine we can train a deep networks to hallucinate imagery suggesting that we are able to play soccer like Lionel Messi even though we may not be able to do so. Consider the images in Fig. 1.1.

In this thesis, we describe research toward synthesizing realistic images or sequences that forecasts the appearance of people performing desired actions. More specifically, we focus on generating realistic images of a particular person striking a desired pose. The model can take a pose and a reference image of a person as input, and generate a novel view of the person in the given pose, while capturing fine appearance details. Human body pose is a natural intermediate representation for this generation, and hence utilized in many previous methods for synthesizing human motion and video [1, 6, 38]. We follow this paradigm, using body poses to generate person images and videos. Simple networks [38] may generate blurry and distorted images. Stylistic methods [13] have shown great success in generating realistic images, but lack control over the appearance of the generated images. Our task requires the model to be able to generate images of a person with a specific appearance. Inspired by [5], we propose a novel appearance rendering network which encodes appearance into convolutional filters. These filters are operationalized using a fully convolutional network, and utilized in an image-to-image translation structure that transfers the desired appearance to the generated image. Our model adopts a two-stream approach that explicitly disentangles two sources of variations in this generation task: human body structures are encoded into feature maps; appearance details are encoded into convolutional kernels separately and applied to the given pose through convolution operation. It avoids the complicated design of many conditional generative models on similar tasks but proves surprisingly effective.

To sum up, we contribute a simple yet effective generative model that performs adaptive appearance rendering to create accurate depictions of human figures in these human poses. We demonstrate our model on two applications: fashion image synthesis and multi-person video synthesis involving complex human activities. Comprehensive quantitative results are provided to facilitate the analysis of each component of our model, and qualitative visualizations are shown in complementary to quantitative results. The experiment results show that our proposed model poses a strong base model for the task of generating human images conditioned on pose and appearance.

Our work is one of the latest advancements in the study of controllable image generation. The simplicity and excellent performance of our model makes it a strong base model that can be com-
pared against and further developed on unleash more of its potentials. It could be an important step towards solving more complicated problems in computer vision, including video frame synthesis and developing intelligent systems to assist with image editing. Compared with previous approaches, our study also shows the importance and benefits of disentangling structure and appearance representations on certain image generation tasks.

This thesis is based on my research paper *Adaptive Appearance Rendering* [40]. The paper is result of collaboration between Mengyao Zhai, Jiacheng Chen, Lei Chen, Zhiwei Deng and me, and it is advised by Dr. Greg Mori. The model and approach were mainly proposed by Mengyao Zhai and I helped formulating the ideas of this paper by participating in discussions with Mengyao. I proposed the novel evaluation metric of perceptual pose score inspired by the latest advancement of human pose estimation and the perceptual score in evaluating generative models. Compared with other evaluation metrics, this metric is specifically designed for our tasks and better demonstrates the strengths of our model. I also helped with the implementations and experiments of different model design changes for performance improvements and part of the experiments on ablated models and baseline models. I sincerely thank all the co-authors of this paper.
Chapter 2

Related Works

**Image generation:** Image-to-image translation has achieved great success since the emergence of generative adversarial networks (GANs) [9]. GAN is a generative model consisting of two parts: a generator mapping a latent space to images and a discriminator discriminating real images from the generated ones. The training objective of GAN is to maximize the probability of correctly classifying generated images and real ones with respect to the generator and minimize it with respect to the discriminator. Recent works produce promising results using GAN-based models on various tasks. Pix2pix [13] trains a conditional generation network using reconstruction and adversarial loss with paired input-output images. The model is able to perform image transfer from one domain to another. CycleGAN [41] performs a similar task of image domain transfer. It reconstructs input images from the outputs and uses cycle-consistency loss for training with unpaired images.

Outside the literature of GAN, Gatys et al. [8] is the first to propose neural network based image artistic style transfer by jointly optimizing the style loss and content loss with respect to model’s input. Johnson et al. [14] directly use a content reconstruction loss and a style reconstruction loss to supervise the training of a feed-forward network for style transfer. The content and style loss of both works are defined by the feature maps of a pretrained deep neural network. The recent work of Chen et al. [5] proposes a structure to disentangle style and content for style transfer. Styles are encoded using a stylebank (a fixed set of convolution filters). Visual analogy making [30, 34] generates or searches for a new image analogous to an input one, based on other previously given example pairs.

**Generation from intermediate representations:** Image generation directly in low-level pixel space is difficult and these types of approaches tend to generate blurry or distorted future frames. To tackle this problem, hierarchical models [38, 37, 42, 16, 22, 15] adopt intermediate representations.

Walker et al. [38] and Villegas et al. [37] both tackle the problem of future frame forecasting in human motion videos. They adopt similar two-stage pipelines by first predicting the future pose sequence and generating frames conditioned on the poses. The work of Walker et al. [38] uses a sequence VAE [19] to model the uncertainty about future motions and a conditional GAN to generate photo-realistic video at pixel level while the work of Villegas et al. [37] uses a deterministic RNN model to predict future poses and a visual-structure analogy model to turn poses into video
frames. Ma et al. [22] propose a two-stage generator to generate human images given desired poses and appearances. The first stage leverages a U-Net [33] shaped model to make a coarse prediction given the target pose map and condition appearance image. In the second stage, the model produces a difference map to refine the coarse prediction.

Zhu et al. [42], Karacan et al. [16], Johnson et al. [15] and Ma et al. [22] share a similar spirit of generating images conditioned on high-level semantic annotations. Zhu et al. [42] studied the problem of manipulating the appearance of human’s wearing using natural language. It uses predicted semantic segmentation map of human body as the intermediate representation. Karacan et al. [16] propose the attribute and layout conditioned GAN model which takes scene layout represented by semantic segmentation mask and transient attributes as conditional inputs to generate diverse images. Johnson et al. [15] take scene graph as a high-level semantic input to their model and predict bounding boxes of objects in scene layout to generate photo-realistic images.

This type of approaches can alleviate image blur, however the quality of generation largely depends on the the image generation network. Simple generation networks can still produce blurry images as shown in [38]. It is worth noting that the approach of Ma et al. [22] solves a problem similar to our work. However, the main contribution of this work is to propose a simple yet well-performing base model to the problem of generating images conditioned on pose and appearance.

Video generation: Data-driven video generation has seen a renaissance in recent years. One major branch of methods uses RNN-based models such as encoder-decoder LSTMs for direct pixel-level video prediction [28, 36, 27, 25]. Ranzato et al. [28] borrow idea from language modeling literature and treat next frame prediction as a classification problem. Their approach discretizes videos frames by applying K-means on non-overlapping patches and uses language model based on recurrent neural networks to predict the next frame. The model forms a strong baseline of filling and prediction tasks on video data. Srivastava et al. [36] explored using different LSTM models [10] for unsupervised representation learning on video data. The work uses auto-regressive approach to predict long sequences into the future as a regression problem. Mathieu et al. [25] specifically target the problem of blurry prediction resulting from using mean squared error (MSE) in training. It proposes the use of multi-scale prediction, adversarial training and image gradient difference loss to improve the quality of predicted future frames. Oh et al. [27] studied the problem of video prediction in vision-based reinforcement learning settings. Their model predicts the next frame of video game given the past history of observation and the current action of controller. It can make long-term action-conditional prediction without diverging from the game. It also shows usefulness in improving the training of deep-Q-network [26] via informed exploration. These methods successfully synthesize low-resolution videos with relatively simple semantics, such as moving MNIST digits or human action videos with very regular, smooth motion.

Subsequent works have attempted to expand the quality of predicted video in terms of resolution and diversity in human activity. Earlier efforts were focused on optical flow-timescale prediction; further work pushed past into more complex motions (e.g. [21]). Liu et al. [21] propose a model that predicts 3D voxel flows from input video frames and uses trilateral interpolation to fill in missing
frames or extrapolation to predict future frames. Experiment results show that the model is able to synthesize frames with better quality than direct prediction or prediction based on traditional optical flow.

In summary, our approach builds on the substantial body of related work in pose analysis and style/analogy-based image generation. We contribute a novel, simple and effective method for adaptive appearance rendering model for image/video generation from human poses.
Chapter 3

Adaptive Rendering Network

Figure 3.1: Overview of the adaptive appearance rendering network. Given input posemap image and appearance reference image, the appearance is encoded into convolutional filters, then these filters are placed in the image-to-image translation network to transfer the appearance to images of each person striking the desired pose.

Given pose estimations as intermediate representations, the goal of our model is to synthesize realistic image of person with desired appearance and pose, where the pose of every person is a posemap image in which “white” body joint points are drawn on a black background canvas. Great success has been shown in generating realistic images given a sketch as inputs [13], which is similar to our task since the posemap image can be treated as a sketch. However, for our task we cannot generate random appearance and we need control over appearance of generated images to make sure that the generated person is wearing the desired clothing.

To accomplish this goal, we propose an adaptive rendering structure where the appearance filters are adaptively computed from an input reference image using a fully convolutional neural network (FCN) and supervised using style loss to enforce similar color distributions as the input reference image. By incorporating this FCN into an image-to-image translation network a realistic image of a target consistent with the desired action and appearance can be generated.
3.1 Network Structure

Fig. 3.1 shows our adaptive rendering network (Ada-R Network) architecture, which consists of two branches: an encoder-decoder branch, and an adaptive rendering branch. The network requires two input images: a posemap image, and a reference image which provides the appearance of the same person in a different pose. The goal of the network is to generate a realistic image of a person consistent with the posemap and having appearance consistent with the reference image.

- **Encoder-Decoder**: Instead of training an encoder-decoder network which can reconstruct input images, our encoder-decoder branch shown in Fig. 3.1 is a sketch → image model. We use the same input size and encoder-decoder structure as in [13]: both generator and discriminator use modules of the form convolution-BatchNorm-Relu [12], the encoder consists of convolutional layers with stride 2 and symmetrically the decoder consists of convolutional layers with fractional stride $\frac{1}{2}$.

- **Adaptive Rendering**: The encoder-decoder network takes binary posemap images as inputs which do not contain any information about the appearance of the person. Hence, we propose to use another network to learn appearance information. By combining these two networks together we are able to generate realistic images of a person wearing the desired clothing. Here we introduce our Ada-R network.

To transfer the desired appearance to the encoder-decoder branch, we replace the last convolutional filter in the encoder-decoder branch with our adaptive appearance transfer filter. The adaptive appearance filter $K_{\text{ada-app}}$ encoding appearance information of a person is derived from an input appearance reference image $I_{\text{app}}$ using a fully-convolutional network

$$K_{\text{ada-app}} = FCN(I_{\text{app}}) \quad (3.1)$$

For image generation, the inputs are appearance-posemap image pairs. For video generation, the rendering of one person’s posemap sequence only requires one reference image, and the frames could be obtained by performing adaptive appearance rendering frame by frame. The filter is applied to the rendering procedure by

$$F = \mathcal{E}(I_{\text{pose}}) \quad (3.2)$$

$$\bar{F} = F \ast K_{\text{ada-app}} \quad (3.3)$$

where $\mathcal{E}$ is the encoder network, $I_{\text{pose}}$ is the posemap image and $\ast$ is a regular 2D convolution operator. $F$ is the feature map of size $w \times h \times C_{\text{in}}$ generated by the encoder network. We set the size of $K_{\text{ada-app}}$ to $k \times k \times C_{\text{in}} \times C_{\text{out}}$ to be compatible with $F$, where $k$, $C_{\text{in}}$, and $C_{\text{out}}$ denote the kernel size, input channel number and output channel number of this adaptive convolution operation. $\bar{F}$ is the feature map after applying the adaptive appearance filter to the feature map $F$. The person
$I_{gen}$ with desired appearance is finally produced with

$$I_{gen} = D(\hat{F})$$

(3.4)

where $D$ is the decoder network.

We release a reference implementation$^1$ to provide the full details on the model architecture and parameter settings.

### 3.2 Loss Function

Our network is trained in an adversarial setting, where the Ada-R network is the generator $G$, and a discriminator $D$ is introduced to discriminate between real and generated images. The overall architecture is a conditional generative adversarial network conditioned on the reference image. Let $I_{goal}$ be the target image that we try to produce, and $I_{gen}$ be the image that the Ada-R network generated. The loss of our Ada-R network is defined as

$$\mathcal{L}_{CGAN}(G,D) + \mathcal{L}_T$$

(3.5)

where $\mathcal{L}_{CGAN}(G,D)$ is the standard adversarial loss. $\mathcal{L}_T$ is our appearance transfer loss, defined below.

$$\mathcal{L}_T = \alpha \mathcal{L}_1(I_{gen}, I_{goal}) + \beta \mathcal{L}_C(I_{gen}, I_{goal}) + \gamma \mathcal{L}_S(I_{gen}, I_{app}).$$

(3.6)

$\mathcal{L}_1$ is an $L_1$ loss that encourages pixel-level agreement between the generated image and the target image, defined as:

$$\mathcal{L}_1(I_{gen}, I_{goal}) = ||I_{gen} - I_{goal}||.$$  

(3.7)

$\mathcal{L}_C$ and $\mathcal{L}_S$ are the content and style loss, defined the same as Gatys et al. [8]. They aim to preserve image structure and colour distributions respectively:

$$\mathcal{L}_C(I_{gen}, I_{goal}) = \sum_{l \in l_c} ||F_l(I_{gen}) - F_l(I_{goal})||^2$$

(3.8)

$$\mathcal{L}_S(I_{gen}, I_{app}) = \sum_{l \in l_s} ||G_l(I_{gen}) - G_l(I_{app})||^2$$

(3.9)

where $F_l$ is the feature map from layer $l$ of a pretrained VGG-19 network [35]. $l_c$ are layers of VGG-19 used to compute the content loss. $G_l(\cdot)$ is the Gram matrix which learns the correlations of color distribution given two input images. $l_s$ are layers of VGG-19 used to compute the style loss.

The final objective is defined as

$$G^* = \arg \min_G \max_D \mathcal{L}_{CGAN}(G,D) + \mathcal{L}_T$$

(3.10)

$^1$https://github.com/wisdomdeng/AdaptiveRendering
Chapter 4

Experiments

We demonstrate our model on the DeepFashion dataset [20] and Volleyball dataset [11]. For the DeepFashion dataset, the goal is to render a given appearance to different poses of same person. Due to the completeness of poses, we perform comprehensive quantitative evaluations and an ablation study on the DeepFashion dataset. We also demonstrate a novel application of our model on the Volleyball dataset where the goal is to synthesize short videos (5 frames) containing groups of people given the people in the 1st frame as appearance reference image.

To train our Ada-R network, we compute content loss at layer $\text{relu4-2}$ and style loss at layer $\text{relu1-2}$, $\text{relu2-2}$, $\text{relu3-2}$, $\text{relu4-2}$ and $\text{relu5-2}$ of the pre-trained VGG-19 network. We set the learning rate to $1e^{-3}$, loss weights are set to bring the $L_1$, content and style losses to a similar scale. For fashion dataset $\alpha = 100$, $\beta = 1e^{-4}$, $\sigma = 1e^{-14}$. For Volleyball dataset, $\alpha = 100$, $\beta = 0.1$, $\sigma = 1e^{-12}$. To make the training stable, for DeepFashion dataset, in each iteration the generator is updated three times and the discriminator is updated one time; and for Volleyball dataset, in each iteration the generator is updated twice and the discriminator is updated one time.

We adopt Mean Square Error (MSE), Peak Signal-to-noise Ratio (PSNR), and Structural Similarity (SSIM) [39] as evaluation metrics. MSE and PSNR are pixel-wise measurements for quality of reconstructed or generated images. SSIM measures the quality of a generated image by considering the image’s structural information instead of only pixel-wise errors. For both datasets, the OpenPose detector [4] is used to obtain corresponding poses for each target in a given image.

4.1 Experiments on DeepFashion

DeepFashion contains fashion images of different person IDs, and most IDs contain 4 views: front, side, back and full body. We filtered out IDs without upper body view or with less than 3 views, resulting in 3418 person IDs. To train our model, we use 2395 person IDs, resulting in 14370 pose-appearance pairs. To test our model, we use 1023 person IDs, resulting in 6138 pose-appearance pairs. Original images are resized to $128 \times 128$ for both training and testing.

To analyze the strength of our model as well as the importance of each component in both the network structure and loss, we compare our Ada-R approach with the following approaches:
• Ada-R w/o style loss: In this baseline, we remove the style loss from the loss function.

• Ada-R w/o content loss: In this baseline, we remove the content loss from the loss function.

• Ada-R w/o $L_1$ loss: In this baseline, we remove the $L_1$ loss from the loss function.

• Pose-GAN* (PG*): We adopt the generation structure used by Walker et al. [38], namely the posemap image and appearance image are concatenated as input to the image-to-image translation network and the FCN is removed. The same loss is used as our Ada-R approach.

• Pose-GAN* (PG*) w/o style loss: We adopt the generation structure of PG and remove the style loss from the loss function.

• Visual Analogy Making* (VAM*): We adopt the analogy based generation structure used in [37, 30]. Same loss is used as our Ada-R approach.

For evaluation, posemaps and images are normalized to $[-1, 1]$. To measure the quality of the poses of the generated images, we propose a new perceptual pose score which uses a state-of-the-art pose estimator [4] to extract pose from generated images and compares each pose joint with the corresponding ground truth pose joint using Euclidean distance. Quantitative results showing comparisons among our approach with all baselines are shown in Tab. 4.1 and visualizations of all approaches are shown in Fig. 4.1 and Fig. 4.2.

The quantitative evaluation results show that our Ada-R models, except for the ablated one without $L_1$ loss, uniformly outperform Pose-GAN and Visual Analogy Making models by significant margins across all measures. The ablation study results for Ada-R models on different components of the loss function also show that each component contributes positively to the model’s performance. Since the target image directly supervises the model training through the content loss and $L_1$ loss, removing one of them, especially the $L_1$ loss, causes the largest degradation of model performance. This degradation in performance is also reflected in visualization results presented in Fig. 4.1. As we can see from Fig. 4.1, removing $L_1$ or content loss results in: (1) the pose of the target in the generated images being inferior (missing arms or hands); (2) images generated being more blurry; (3) the appearance of the generated images showing more errors in details (color transfer to arms, wrong colors, uneven colors in clothes). As seen from Tab. 4.1, removing the style loss from the training objective causes relatively smaller impacts on the model’s performance and the model can generate person images striking the correct pose. However it is observed that many examples generated by Ada-R w/o style loss do not reflect the colors of the style reference image correctly and the colors tend to be flat and unvaried compared with our full model.

When compared with PG* models that learn the representation for the pose and the style together, Ada-R which adopts a two-stream approach and encodes the pose and style separately also shows clear advantages. As shown in Fig. 4.2, PG* tends to generate images that show both the pose of the input posemap and the pose in the style reference image so as to match the color distribution of the style image, resulting in two overlapping targets in the generated images. While PG*
Figure 4.1: Visualization of generation results of Ada-R using different loss terms on DeepFashion.

w/o style loss tends to generate blurry images with incorrect poses or appearance as shown by the pose score and examples in Fig. 4.2. The comparison demonstrates the advantage of disentangling the representations of pose and style and targeting each representation with a specific loss. Visual Analogy Making* (VAM*) [30], also uses disentangled representations for pose and style and we use the same loss function as our model for training. However, we find that it cannot generate images with correct poses most of the time. Failure examples of Visual Analogy Making* are presented in Fig. 4.2.

|                | Ours | Ours w/o $L_s$ | Ours w/o $L_C$ | Ours w/o $L_1$ | PG w/o $L_s$ | PG w/o $L_C$ | VAM* |
|----------------|------|----------------|----------------|----------------|--------------|--------------|------|
| MSE            | 0.0849 | 0.0884 | 0.1070 | 0.1136 | 0.1276 | 0.1236 | 0.1652 |
| PSNR           | 17.2338 | 17.0346 | 16.1756 | 15.8898 | 15.5597 | 15.6150 | 14.3356 |
| SSIM           | 0.6508 | 0.6484 | 0.6153 | 0.5810 | 0.5836 | 0.5708 | 0.5562 |
| Perceptual     | 0.0310 | 0.0371 | 0.0671 | 0.0538 | 0.1046 | 0.0859 | 0.0910 |

Table 4.1: Quantitative measures on DeepFashion.
4.2 Experiments on Volleyball

The Volleyball dataset contains sequences of volleyball games. For this dataset, our model is trained to observe players in the 1st input frames and predict their future appearances from frame 6 to frame 10. To get training sequences of each player, we run person detection [32] and tracking [18] to get tracklets of each player in each clip. We follow the data split of original dataset and preprocessing is conducted to filter out instances with less than 10 joints and clips containing less than 10 targets. We get 1262 clips for training and 790 clips for testing. Images of players are cropped and then resized to $256 \times 256$ pixels for both training and testing. For evaluations, we compare our model with state-of-the-art approaches including Pose-GAN and VAM as described in Chapter 4.1. Comparisons among our approach and two state-of-the-art approaches are shown in Tab. 4.2 and visualizations of all approaches are shown in Fig. 4.3.
Table 4.2: Quantitative measures on Volleyball dataset.

|       | Ours | PG w/o $\mathcal{L}_s$ | VAM |
|-------|------|-------------------------|-----|
| MSE   | 0.1670 | 0.1854 | 0.2091 |
| PSNR  | 14.0191 | 13.5037 | 13.0174 |
| SSIM  | 0.2825 | 0.2333 | 0.2178 |

As shown in both quantitative and qualitative results, our model performs very well on this application. Our model can generate clean and sharp person image sequences with correct appearance generated and consistent poses as in the posemap input, despite the complex scenario in terms of delicate pose changes, motion blur, and appearance variation.

We could also generate frames in original resolution $1280 \times 720$ by replacing players in a frame by generated ones. Examples of the frame synthesis results are shown in Fig. 4.4, where the generated players are highlighted using blue bounding boxes.

Figure 4.3: Visualizations on Volleyball dataset.

Figure 4.4: Visualizations on whole frame generation on Volleyball dataset. The generated players are highlighted using blue bounding boxes in each frame.
Chapter 5

Limitations and Discussions

The input and output images of our model have relatively low resolutions and quality compared with some recent works on image generation [2, 17]. They demonstrate the possibility of generating images with much larger size and realistic fine detail using generative adversarial networks. Their model design and engineering work could be incorporated into our model for high-resolution image generation on more complicated datasets.

Even though we demonstrate the potential of generating short video clips using our model, the video data are still processed frame by frame. Therefore we do not explicitly enforce consistency between frames in training. This could cause inconsistency in colors and texture between consecutive frames on more complex video data. Each individual frame may look similar to the other by itself. However, such inconsistency can be much more obvious in video playback. To fully extend our model to the domain of video data, we need to take such consistency between frames into the consideration of training objective.

Finally, our model uses ground truth images in the training. The target image may not be available in some datasets. A training objective that does not requires direct supervision from the target image can make the model applicable to more datasets. Such training objective is worthy of studying in future works. One possible method to achieve such unsupervised learning is to replace the $L_1$ and content loss terms with the pose perceptual score estimated using a fully differentiable pose estimator. The adversarial loss will now encourage the model to generate images with similar distributions to the training data which should contain humans. The pose perceptual score can force the model to generate images with the same poses in the input sketches. The style loss term makes sure that the output and the appearance image look visually similar.
Chapter 6

Conclusion

We proposed a novel approach for synthesizing realistic images or sequences that generate the appearance of people taking the desired poses. Our approach encodes the appearance into convolutional filters. These filters are learned using a fully convolutional network, and utilized in an image-to-image translation structure that transfers the desired appearance to the desired pose. Both quantitative and qualitative results show that our model is superior to state-of-the-art approaches and can generate better images involving complex appearance and better videos involving complex human activities. The success of our model demonstrates that the combination of appearance filters and style loss can render the desired target appearance while adapting to the desired pose.
Bibliography

[1] Matthew Brand and Aaron Hertzmann. Style machines. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000, pages 183–192, 2000. doi: 10.1145/344779.344865. URL https://doi.org/10.1145/344779.344865.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. CoRR, abs/1809.11096, 2018. URL http://arxiv.org/abs/1809.11096.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. CoRR, abs/1809.11096, 2018. URL http://arxiv.org/abs/1809.11096.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 1302–1310, 2017. doi: 10.1109/CVPR.2017.143. URL https://doi.org/10.1109/CVPR.2017.143.

[5] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 2770–2779, 2017. doi: 10.1109/CVPR.2017.296. URL https://doi.org/10.1109/CVPR.2017.296.

[6] Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In 9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France, pages 726–733, 2003. doi: 10.1109/ICCV.2003.1238420. URL https://doi.org/10.1109/ICCV.2003.1238420.

[7] Chelsea Finn, Ian J. Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 64–72, 2016. URL https://arxiv.org/abs/1605.07157.

[8] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 2414–2423, 2016. doi: 10.1109/CVPR.2016.265. URL https://doi.org/10.1109/CVPR.2016.265.

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets.
[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[11] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1971–1980, 2016. doi: 10.1109/CVPR.2016.217. URL https://doi.org/10.1109/CVPR.2016.217.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 448–456, 2015. URL http://proceedings.mlr.press/v37/ioffe15.html.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976, 2017. doi: 10.1109/CVPR.2017.632. URL https://doi.org/10.1109/CVPR.2017.632.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, pages 694–711, 2016. doi: 10.1007/978-3-319-46475-6_43. URL https://doi.org/10.1007/978-3-319-46475-6_43.

[15] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1219–1228, 2018. doi: 10.1109/CVPR.2018.00133. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Johnson_Image_Generation_From_CVPR_2018_paper.html.

[16] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *CoRR*, abs/1612.00215, 2016. URL http://arxiv.org/abs/1612.00215.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

[18] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. URL https://dl.acm.org/citation.cfm?id=1755843.

[19] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.
[20] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pages 1096–1104, 2016. doi: 10.1109/CVPR.2016.124. URL https://doi.org/10.1109/CVPR.2016.124.

[21] Ziwei Liu, Raymond A. Yeh, Xiaou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In International Conference on Computer Vision (ICCV), 2017.

[22] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 405–415, 2017. URL http://papers.nips.cc/paper/6644-pose-guided-person-image-generation.

[23] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 99–108, 2018. doi: 10.1109/CVPR.2018.00018. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Ma_Disentangled_Person_Image_CVPR_2018_paper.html.

[24] Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. URL http://arxiv.org/abs/1511.02793.

[25] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016. URL http://arxiv.org/abs/1511.05440.

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. CoRR, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

[27] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L. Lewis, and Satinder P. Singh. Action-conditional video prediction using deep networks in atari games. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 2863–2871, 2015. URL https://arxiv.org/pdf/1507.08750.pdf.

[28] Marc’Aurelio Ranzato, Arthur Szlam, Joan Bruna, Michaël Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. CoRR, abs/1412.6604, 2014. URL http://arxiv.org/abs/1412.6604.

[29] Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. CoRR, abs/1906.00446, 2019. URL http://arxiv.org/abs/1906.00446.
[30] Scott E. Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 1252–1260, 2015. URL http://papers.nips.cc/paper/5845-deep-visual-analogy-making.

[31] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, pages 1060–1069, 2016. URL http://proceedings.mlr.press/v48/reed16.html.

[32] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. volume 39, pages 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031. URL https://doi.org/10.1109/TPAMI.2016.2577031.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III, pages 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

[34] Fereshteh Sadeghi, C. Lawrence Zitnick, and Ali Farhadi. Visalogy: Answering visual analogy questions. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 1882–1890, 2015. URL http://papers.nips.cc/paper/5777-visalogy-answering-visual-analogy-questions.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL http://arxiv.org/abs/1409.1556.

[36] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pages 843–852, 2015. URL http://proceedings.mlr.press/v37/srivastava15.html.

[37] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, pages 3560–3569, 2017. URL http://proceedings.mlr.press/v70/villegas17a.html.

[38] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, pages 3352–3361, 2017. doi: 10.1109/ICCV.2017.361. URL https://doi.org/10.1109/ICCV.2017.361.
[39] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Processing*, 13 (4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

[40] Mengyao Zhai, Ruizhi Deng, Jiacheng Chen, Lei Chen, Zhiwei Deng, and Greg Mori. Adaptive appearance rendering. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 247, 2018. URL http://bmvc2018.org/contents/papers/0837.pdf.

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251, 2017. doi: 10.1109/ICCV.2017.244. URL https://doi.org/10.1109/ICCV.2017.244.

[42] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1689–1697, 2017. doi: 10.1109/ICCV.2017.186. URL https://doi.org/10.1109/ICCV.2017.186.
Appendix A

FCN Structure

Table A.1: FCN used for DeepFashion dataset

| FCN for compute the adaptive convolutional filter $K_{ada-app}$ | Input: reference image |
|---|---|
| Conv1, filter=11×11×96, stride=4, padding="same" | Output size 32×32×96 |
| MaxPooling, ksize =3×3×1, stride=2, padding="valid" | Output size 15×15×96 |
| Conv2, filter=5×5×256, stride=1, padding="same" | Output size 15×15×256 |
| MaxPooling, ksize =3×3×1, stride=2, padding="valid" | Output size 7×7×256 |
| Conv3, filter 3×3×384, stride=3, padding="same" | Output size 3×3×384 |
| Conv4, filter 3×3×96, stride=1, padding="same" | Output size 3×3×96 |
| Conv5, filter 3×3×64+64, stride=1, padding="same" | namely, 3×3×4096 |
| Reshape, Output Adaptive Convolutional Filter | Output size 3×3×64×64 |