Adaptive Appearance Rendering

Mengyao Zhai
mzhai@sfu.ca
Ruizhi Deng
ruizhid@sfu.ca
Jiacheng Chen
jca348@sfu.ca
Lei Chen
chenleic@sfu.ca
Zhiwei Deng
zhiweid@sfu.ca
Greg Mori
mori@cs.sfu.ca

School of Computing Science
Simon Fraser University
Burnaby, Canada

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.

1 Introduction

We may not be able to play soccer like Lionel Messi, but perhaps we can train deep networks to hallucinate imagery suggesting that we can. Consider the images in Fig. 1. In this paper we describe research toward synthesizing realistic images or sequences that forecast the appearance of people performing desired actions. The model can take a pose and a reference image of a person, and generate a novel view of the person in the given pose, while capturing fine appearance details.

Image or video generation has emerged as an important problem in many domains. Within robotics, work has explored predicting the consequences after interactions between an agent and its environment [5]. In natural language processing, approaches [18, 22] have been proposed to tackle tasks such as text to image or image to text synthesis. Generative models of video sequences [28, 29] and fashion images [17, 31] are a core part of visual understanding and have received renewed attention from the vision community.

© 2018. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

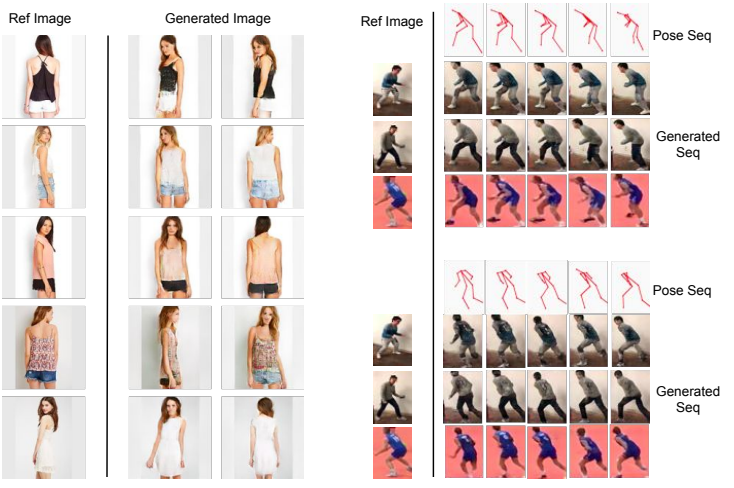

Figure 1: We develop architectures for forecasting person images of various appearances. At the core of the method is adaptive rendering modules. A series of generation examples of various reference images rendered into new poses is shown above.

In this paper, we focus on generating realistic images of a particular person striking a desired pose. 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, 4, 29]. We follow in this paradigm, using body poses to generate person images and videos. Simple networks [29] may generate blurry and distorted images. Stylistic methods [10] 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 [3], 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.

To sum up, we contribute a new state of the art 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.

2 Related Work

Image generation: Image-to-image translation has achieved great success since the emergence of GANs [7]. Recent work produces promising results using GAN-based models [10, 32]. Stylized images can be generated by using feed-forward networks [6] with the help of perceptual loss [11]. The recent work of [3] proposes a structure to disentangle style and content for style transfer. Styles are encoded using a stylebank (set of convolution filters).

Visual analogy making [23, 25] generates or searches for an new image analagous to an input one, based on other previously given example pairs.

Generation from intermediate representations: 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 [12, 16, 28, 29, 33] adopt intermediate representations. This type of approach 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 [29]. It's worth noting that Ma et al. propose a two-stage approach to solve a problem similar to our work [16]. 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 [19, 20, 21, 27]. These methods successfully synthesized low-resolution videos with relatively simple semantics, such as moving MNIST digits or human action videos with very regular, smooth motion. Subsequent work has 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. [15]).

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.

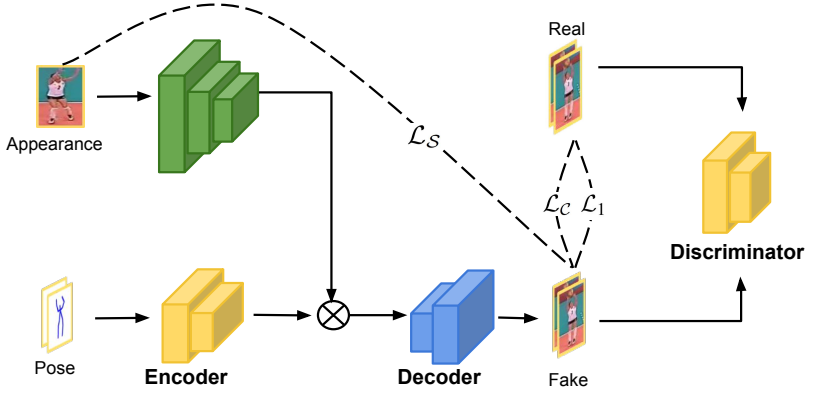

Figure 2: Overview of the adaptive appearance rendering network. Given input posemap image and appearance reference image, the appearance is encoded into convolutional filters, then these filters are placed in the image-to-image translation network to transfer the appearance to images of each person striking the desired pose.

3 Adaptive Rendering Network

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 [10], 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. 2 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. 2 is a sketch $\rightarrow$ image model.

We use the same input size and encoder-decoder structure as in [10]: both generator and discriminator use modules of the form convolution-BatchNorm-Relu [9], 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 $\boldsymbol{K}_{ada\text{-}app}$ encoding appearance information of a person is derived from an input appearance reference image $\boldsymbol{I}_{app}$ using a fully-convolutional network

$$\boldsymbol{K}_{ada\text{-}app} = FCN(\boldsymbol{I}_{app}) \tag{1}$$

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

$$\boldsymbol{F} = \mathcal{E}(\boldsymbol{I}_{pose}) \tag{2}$$

$$\bar{\boldsymbol{F}} = \boldsymbol{F} * \boldsymbol{K}_{ada\text{-}app} \tag{3}$$

where $\mathcal{E}$ is the encoder network, $\boldsymbol{I}_{pose}$ is the posemap image and $*$ is a regular 2D convolution operator. $\boldsymbol{F}$ is the feature map of size $w \times h \times C_{in}$ generated by the encoder network. We set

the size of $\boldsymbol{K}_{ada\text{-}app}$ to $k \times k \times C_{in} \times C_{out}$ to be compatible with $\boldsymbol{F}$, where k, C_{in} and C_{out} denote the kernel size, input channel number and output channel number of this adaptive convolution operation. $\bar{\boldsymbol{F}}$ is the feature map after applying the adaptive appearance filter to the feature map $\boldsymbol{F}$. The person $\boldsymbol{I}_{gen}$ with desired appearance is finally produced with

$$\boldsymbol{I}_{gen} = \mathcal{D}(\bar{\boldsymbol{F}}) \tag{4}$$

where $\mathcal{D}$ 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 $\boldsymbol{I}_{goal}$ be the target image that we try to produce, and $\boldsymbol{I}_{gen}$ be the image that the Ada-R network generated. The loss of our Ada-R network is defined as

$$\mathcal{L}_{\mathcal{CGAN}}(G,D) + \mathcal{L}_{\mathcal{T}} \tag{5}$$

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

$$\mathcal{L}_{\mathcal{T}} = \alpha\mathcal{L}_1(\boldsymbol{I}_{gen},\boldsymbol{I}_{goal}) + \beta\mathcal{L}_{\mathcal{C}}(\boldsymbol{I}_{gen},\boldsymbol{I}_{goal}) + \gamma\mathcal{L}_{\mathcal{S}}(\boldsymbol{I}_{gen},\boldsymbol{I}_{app}). \tag{6}$$

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

$$\mathcal{L}_1(\boldsymbol{I}_{gen},\boldsymbol{I}_{goal}) = ||\boldsymbol{I}_{gen} - \boldsymbol{I}_{goal}||. \tag{7}$$

$\mathcal{L}_{\mathcal{C}}$ and $\mathcal{L}_{\mathcal{S}}$ are the content and style loss, defined the same as Gatys et al. [6]. They aim to preserve image structure and colour distributions respectively:

$$\mathcal{L}_{\mathcal{C}}(\boldsymbol{I}_{gen},\boldsymbol{I}_{goal}) = \sum_{l\in l_c} ||F_l(\boldsymbol{I}_{gen}) - F_l(\boldsymbol{I}_{goal})||^2 \tag{8}$$

$$\mathcal{L}_{\mathcal{S}}(\boldsymbol{I}_{gen},\boldsymbol{I}_{app}) = \sum_{l\in l_s} ||G_l(\boldsymbol{I}_{gen}) - G_l(\boldsymbol{I}_{app})||^2 \tag{9}$$

where F_l is the feature map from layer l of a pretrained VGG-19 network [26]. 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

$$G^\star = \arg\min_G \max_D \mathcal{L}_{\mathcal{CGAN}}(G,D) + \mathcal{L}_{\mathcal{T}} \tag{10}$$

[1] https://github.com/wisdomdeng/AdaptiveRendering

4 Experiments

We demonstrate our model on the DeepFashion dataset [14] and Volleyball dataset [8]. 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 1^{st} frame as appearance reference image.

To train our Ada-R network, we compute content loss at layer *relu4-2* and style loss at layer *relu1-2*, *relu2-2*, *relu3-2*, *relu4-2* and *relu5-2* of the pre-trained VGG-19 network. We set the learning rate to *1e-3*, loss weights are set to bring the $L1$, 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) [30] 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 [2] 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 128×128 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. [29], 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 [23, 28]. 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 [2] 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. 1 and visualizations of all approaches are shown in Fig. 3 and Fig. 4.

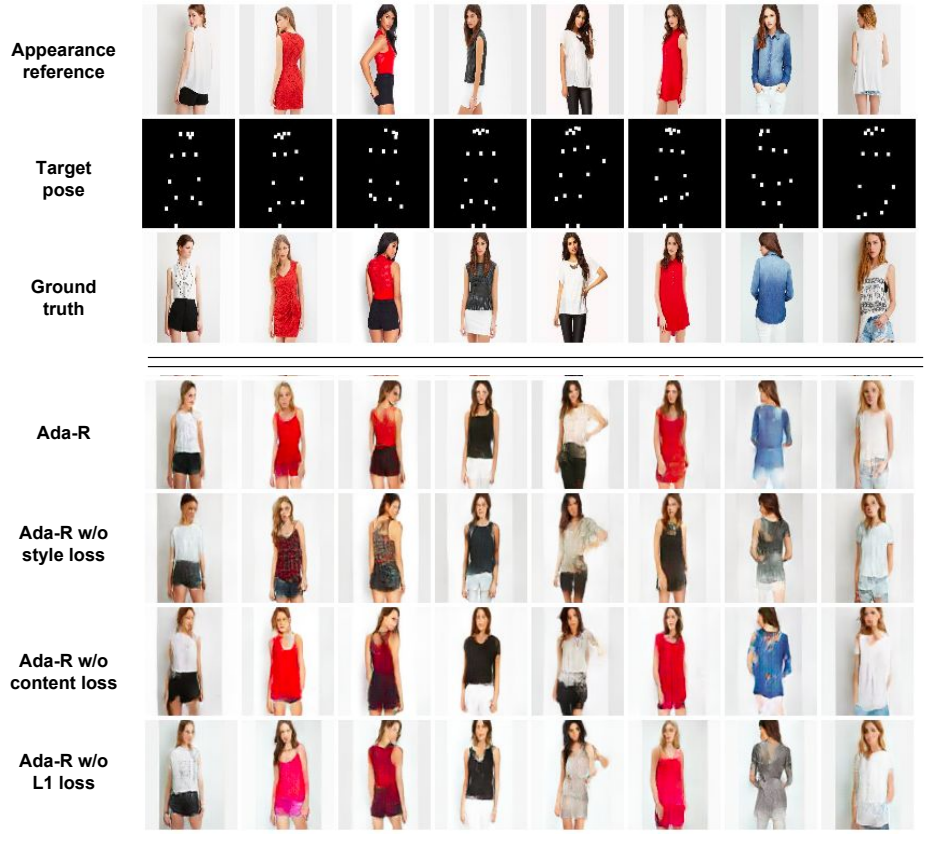

Figure 3: Visualization of generation results of Ada-R using different loss terms on DeepFashion.

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. 3. As we can see from Fig. 3, 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. 1, removing the style loss from the training objective causes relatively smaller impacts on the model's performance and the model can

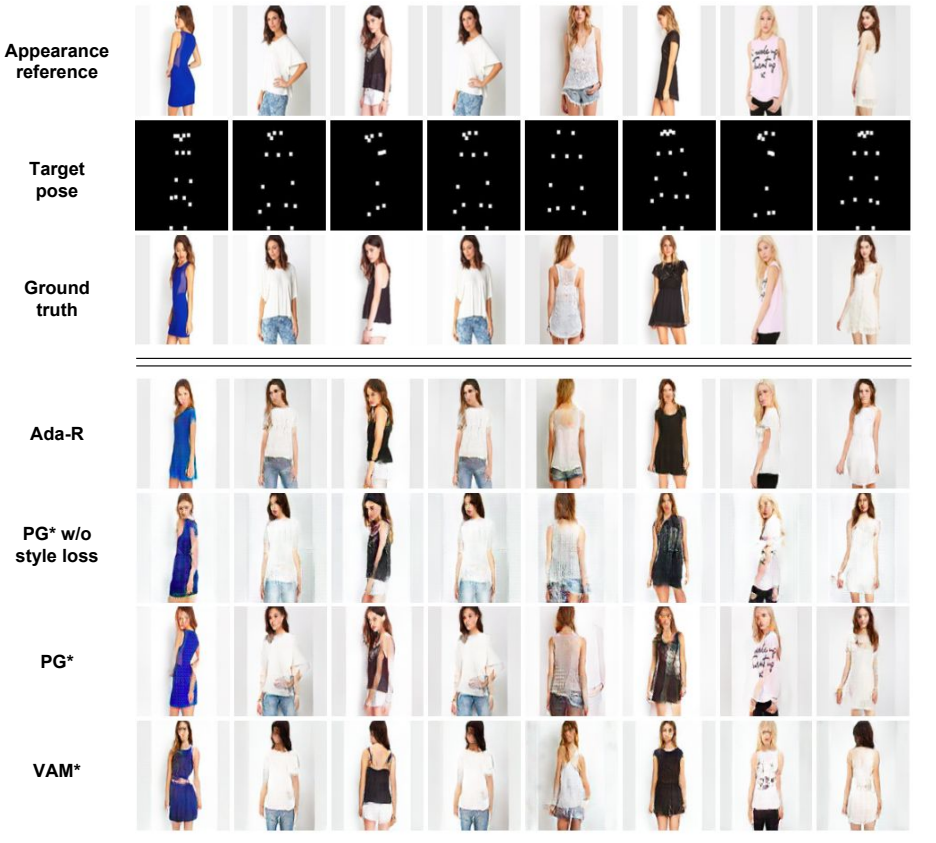

Figure 4: Visualization of generation results of different state-of-the-art approaches on DeepFashion.

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, *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** *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. 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***)* [23], 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.

	Ours	Ours w/o $\mathcal{L}_s$	Ours w/o $\mathcal{L}_C$	Ours w/o L_1	PG	PG w/o $\mathcal{L}_s$	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 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 1^{st} input frames and predict their future appearances from frame 6 to frame 10. To get training sequences of each player, we run person detection [24] and tracking [13] 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 256×256 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 Sec. 4.1. Comparisons among our approach and two state-of-the-art approaches are shown in Tab. 2 and visualizations of all approaches are shown in Fig. 5.

	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

Table 2: Quantitative measures on Volleyball dataset.

Figure 5: Visualizations on Volleyball dataset.

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 1280×720 by replacing players in a frame by generated ones. Examples of the frame synthesis results are shown in Fig. 6, where the generated players are highlighted using blue bounding boxes.

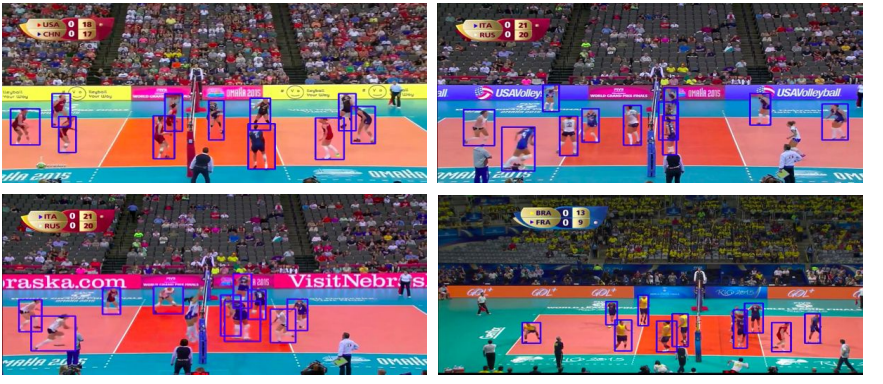

Figure 6: Visualizations on whole frame generation on Volleyball dataset. The generated players are highlighted using blue bounding boxes in each frame.

5 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 predicted pose.

References

[1] M. Brand and A. Hertzmann. Style machines. *Conference on Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 2000.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Dongdong Chen, Lu Yuan, Jing Liao, Nenghai Yu, and Gang Hua. Stylebank: An explicit representation for neural image style transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] A.A. Efros, A.C. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *International Conference on Computer Vision (ICCV)*, 2003.

[5] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

[6] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[8] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[12] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *CoRR*, 2016.

[13] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research(JMLR)*, 10:1755–1758, 2009.

[14] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

[16] 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*, pages 406–416, 2017.

[17] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CoRR*, 2017.

[18] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *International Conference on Learning Representations (ICLR)*, 2016.

[19] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2016.

[20] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[21] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint arXiv:1412.6604*, 2014.

[22] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, 2016.

[23] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[25] Fereshteh Sadeghi, C Lawrence Zitnick, and Ali Farhadi. Visalogy: Answering visual analogy questions. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[27] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning (ICML)*, 2015.

[28] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *International Conference on Machine Learning (ICML)*, 2017.

[29] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *International Conference on Computer Vision (ICCV)*, 2017.

[30] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[31] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, and Jiashi Feng. Multi-view image generation from a single-view. *CoRR*, abs/1704.04886, 2017.

[32] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017.

[33] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. *CoRR*, 2017.