Pose-Guided High-Resolution Appearance Transfer via Progressive Training

Ji Liu *
Carnegie Mellon University
jiliu@andrew.cmu.edu

Heshan Liu *
Carnegie Mellon University
heshanl@andrew.cmu.edu

Mang-Tik Chiu
UIUC
mtchiu2@illinois.edu

Yu-Wing Tai
Tencent
yuwingtai@tencent.com

Chi-Keung Tang
HKUST
cktang@cs.ust.hk

Abstract

We propose a novel pose-guided appearance transfer network for transferring a given reference appearance to a target pose in unprecedented image resolution (1024^2), given respectively an image of the reference and target person. No 3D model is used. Instead, our network utilizes dense local descriptors including local perceptual loss and local discriminators to refine details, which is trained progressively in a coarse-to-fine manner to produce the high-resolution output to faithfully preserve complex appearance of garment textures and geometry, while hallucinating seamlessly the transferred appearances including those with dis-occlusion. Our progressive encoder-decoder architecture can learn the reference appearance inherent in the input image at multiple scales. Extensive experimental results on the Human3.6M dataset, the DeepFashion dataset, and our dataset collected from YouTube show that our model produces high-quality images, which can be further utilized in useful applications such as garment transfer between people and pose-guided human video generation.

1. Introduction

Learning 3D appearance from 2D images is challenging because images are 2D projections of the corresponding 3D world where objects can undergo complex deformation and occlusion. Existing relevant work in computer vision and machine learning either uses some forms of 3D model to produce quality results, or produces relatively low-resolution output image if only 2D images are allowed.

This paper focuses on images of human, whose different poses introduce complex non-rigid deformation and self-occlusion. Specifically, given a reference image of a person, our method seamlessly transfers the reference appearance to the person at the target pose while preserving high-quality garment texture of the reference person, and at the same time hallucinating realistically their complex appearance under the target pose, see Figures 1 and 2. Note that the network should not only move the corresponding body parts to match the target pose, but also realistically inpaint or hallucinate exposed body/garment parts unseen in the input due to occlusion. This is particularly challenging for human images due to the non-rigid nature of 3D human

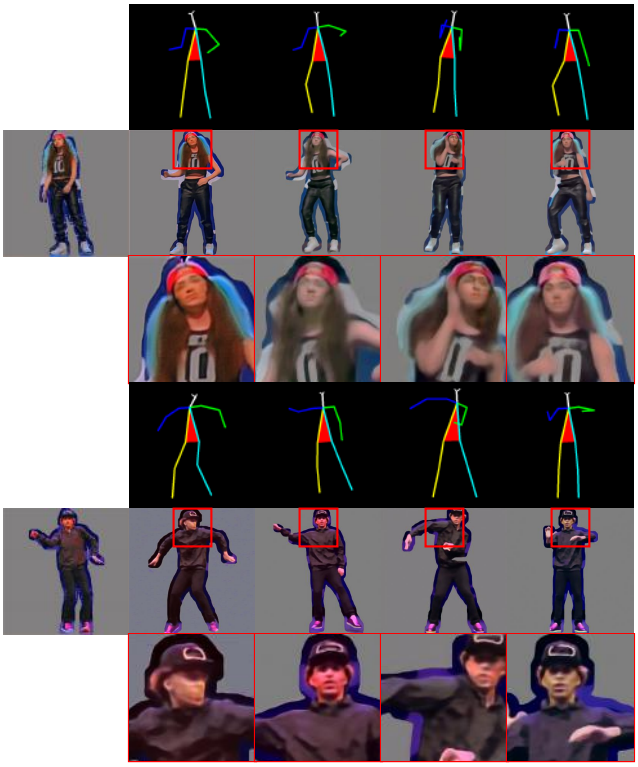

Figure 1: Pose transfer on YouTube dataset. Test results on our self-collected high-resolution (1024^2) dataset. Given a reference image (leftmost column) and target poses as input which contains self-occlusion with complex appearance in texture and geometry, our method transfers the reference appearance to target pose in high resolution while faithfully preserving complex appearance and facial features under large pose variations.

* Equal contribution. Authorship order was determined by rolling dice.

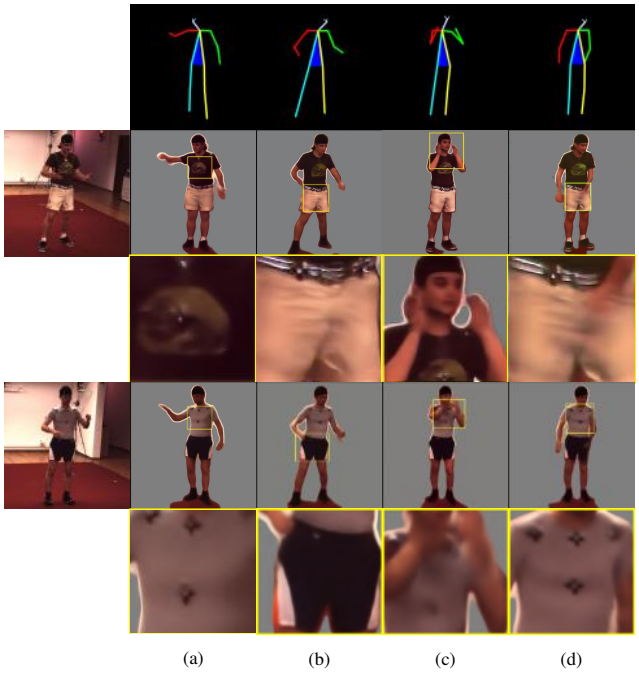

Figure 2: Pose transfer on Human3.6M dataset. Test results on human3.6M dataset. (a), (b) demonstrate the effect of dis-occlusion; (c), (d) demonstrate the effect of transferring to other self-occluding poses with zoom-in views.

body and complex texture and geometry distortion on 3D garment worn by humans.

To address these challenges, we propose to learn appearance information inherent in a given reference image and transfer the original appearance of the person according to the target pose representation, by injecting the target pose representation into the bottleneck of the encoder-decoder architecture. More importantly, to transfer detailed appearance, we enforce both global and local loss to encourage the network to learn both global coherency and local details. In addition, progressive growth is employed on both encoder and decoder to increase output resolution. To validate our approach, we conduct extensive experiments on the Human3.6M [13] dataset, the DeepFashion [25] dataset and our dataset collected from YouTube. We apply our method to other applications such as high-quality garment transfer and pose-guided human video generation, demonstrating its huge potential in many challenging tasks.

Our contribution consists of a new encoder-decoder architecture that successfully enables appearance transfer to a target pose. To enable high-resolution appearance transfer, 1) we propose novel local descriptors (progressive local perceptual loss + local discriminators at the highest resolution (1024^2) to enhance local details and generation quality; 2) we apply progressive training to our autoencoder architecture to achieve outputs at unprecedented high resolution (1024^2). To our knowledge, this is the first progressive, deep encoder-decoder transfer network that can realistically hallucinate in such high resolution at the target pose the complex appearance of the worn garment, including the portion that was previously occluded in the reference image.

2. Related Work

We address the problem of high-resolution pose-guided appearance transfer, which is also a problem of conditional image generation. Therefore, in this section, works related to conditional image generation and pose transfer will first be discussed, followed by recent approaches that can produce high-resolution images.

Conditional image generation Generative models including Variational Autoencoders [19] (VAEs) and Generative Adversarial Networks [8] (GANs) have demonstrated success in image generation. Although VAEs can generate target images complying a given reference image, they may not faithfully preserve high-quality details because a lower bound is optimized.

Conditional GANs [14] have been exploited to solve many challenging tasks. Zhao *et al.* [43] integrated GANs and other inference models to generate images of people in various clothing styles from multiple views. Reed *et al.* [31] proposed a conditional generative model that used pose and text as conditions to generate images. Lassner *et al.* [20] also presented a generative model that could generate realistic images conditioning on clothing segmentation.

Numerous researchers [36, 29, 39, 41, 46] introduced their respective methods to enable more control on the appearance of the generated images in generative processes by providing different intermediate information such as labels and texts. Models such as Conditional GANs [14] and CycleGAN [45] also demonstrated their efficacy in image-to-image translation.

However, it is difficult for the above methods to simultaneously encode different factors such as pose and appearance. To transfer the pose-invariant human appearance, disentangling pose and appearance from the reference image becomes an essential step. Many previous studies [4, 5] attempted to use GANs [8] and autoencoders [1] to disentangle such factors, including writing styles from character identities. Recently, Tran *et al.* [37] proposed DRGAN, which can disentangle pose from identity by learning the representation of human face followed by synthesizing the face with preserved identity at the target pose.

Appearance transfer Recent work produced high-quality transfer results by employing 3D human models and information (in the data synthesis stage) [22] or by estimating 3D human model as an intermediate step [40]. With no 3D information used, approaches for pose transfer [6, 27] used encoder-decoders to attempt disentangling the pose and appearance of the input image to perform pose transfer. Esser *et al.* [7] explored a variational U-Net [32] on transferring the pose of a reference image invariant with its appearance.

The PG^2 [26] was a more related work that aims at generating images of a subject in various poses based on an im-

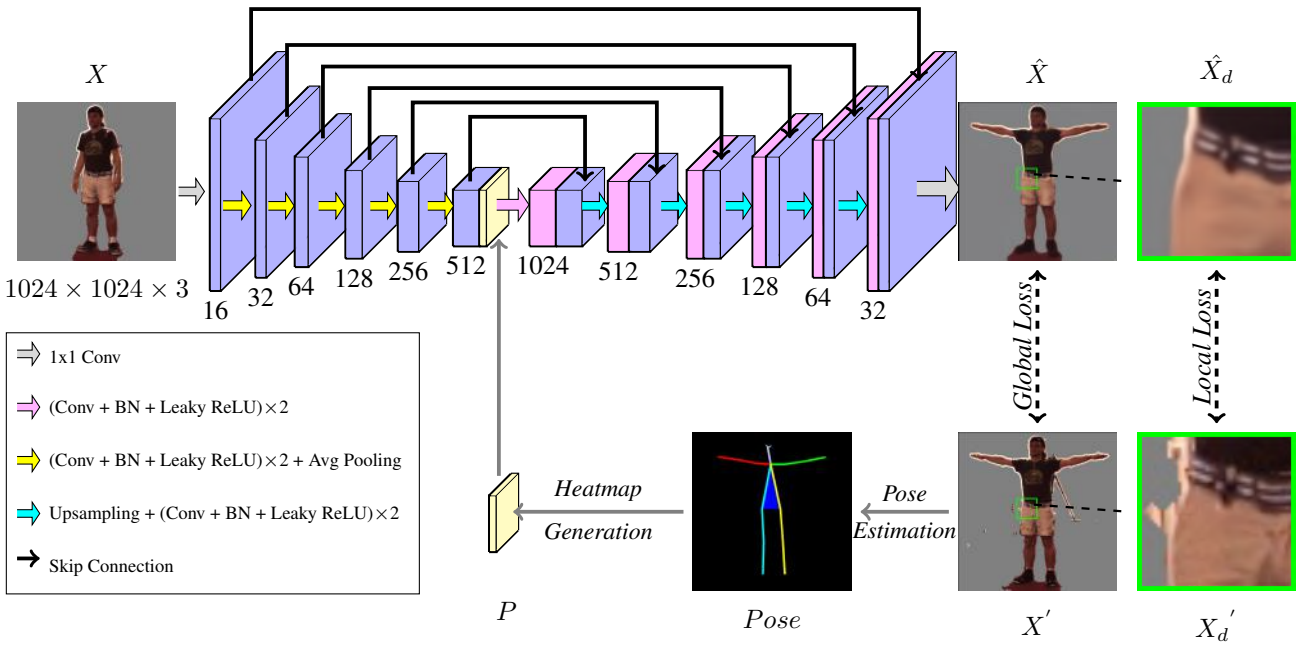

Figure 3: Overall Network Architecture. The reference image X is first passed through an encoder to generate a latent representation. In the lower branch, 18 keypoints are estimated from the ground truth image $X^{'}$ to produce an explicit pose representation P. P is then concatenated with the latent representation, which is further decoded into the output $\hat{X}$. Global perceptual loss is enforced between $X^{'}$ and $\hat{X}$. To enable high-resolution appearance transfer, two types of local loss are also enforced on the corresponding local regions ($X_d{}^{'}$, $\hat{X}_d$), indicated by the bounding boxes. Local details are progressively refined. See section 3.2 and section 3.3 for technical contributions.

age of that person and one novel pose. Combining GANs and autoencoders, PG^2 was trained through an encoder-decoder network followed by a refinement network given the pose and person image as input. Siarohin *et al.* [34] proposed a generative model similar to PG^2, where a discriminator was included at the end of the autoencoder to help generate realistic images. Instead of using a discriminator, the pose transfer network presented by Natalia *et al.* [28] attempted to produce the seamless result by blending the synthesized image and warped image through end-to-end training. Though not aiming at transferring human pose, the landmark learning network recently proposed by Jakab *et al.* [15] demonstrated acceptable results on pose transferring, which was achieved by using a simple encoder-decoder network with the learning landmarks concatenated in an intermediate representation.

Although [26, 28, 15] performed well on changing pose at low-resolution (128^2) reference images while keeping their rough identity, they could not preserve but significantly blur complex textures after pose transfer. In contrast, we are dealing with a more challenging task compared to their work since we want to preserve as many details as possible at the target pose presented in the high-resolution reference image.

Progressive training In the generative model, producing high-resolution and high-quality results is difficult since the training process becomes unstable and hard to converge as the output dimension increases.

Recently, Tero *et al.* [17] proposed a progressive training methodology for GANs to generate high-quality results. They started training from low resolution and added layers to the model progressively to obtain satisfactory high-resolution results. Tero's work focused on GANs and cannot be directly applied to autoencoders while our goal is conditional image generation using autoencoders for even higher resolution output (1024^2).

3. Method

Figure 3 details the autoencoder architecture for pose-guided appearance transfer with technical contributions to transfer appearance details at 1024^2 to be detailed.

Specifically, given a reference image X of a person and another image $X^{'}$ of the same person which is in the target pose, we first extract the explicit pose representation P from $X^{'}$ using a state-of-the-art pose estimator (section 3.1). We then inject P into the autoencoder's bottleneck by concatenating it with the deepest feature map generated by the encoder. Finally, the concatenated feature block is passed through a decoder to generate an image with the person in the target pose, denoted as $\hat{X}$. Reconstruction loss is enforced globally between $X^{'}$ and $\hat{X}$ (section 3.1). To enable high-resolution appearance transfer, we employ novel local descriptors to refine output details. Local descriptors are applied under the guidance of keypoint locations from the pose estimator (section 3.2). To generate images in a

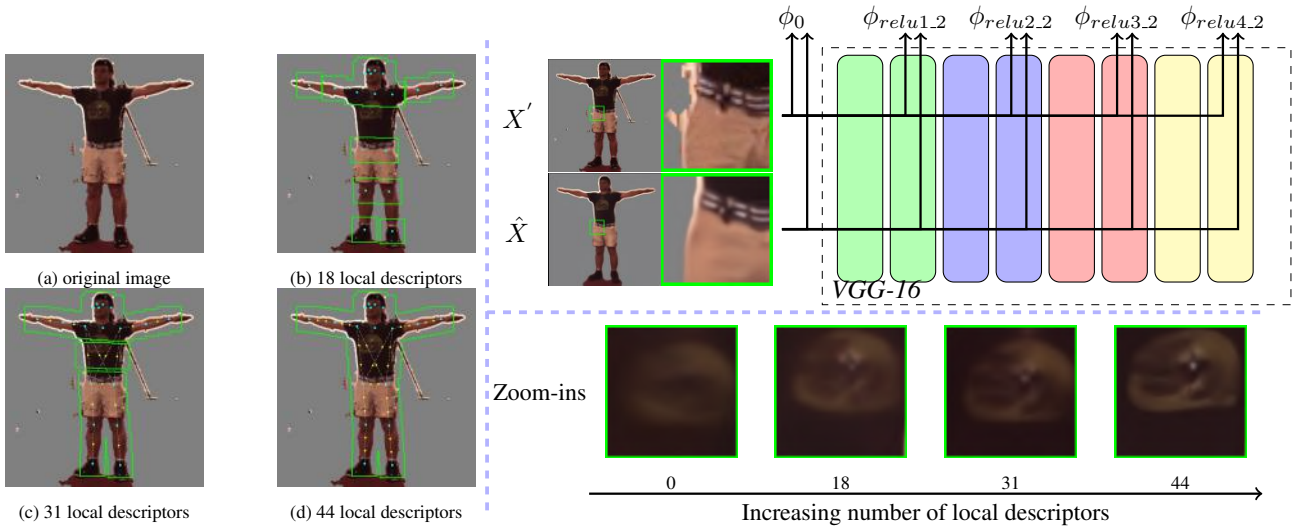

Figure 4: Left: the distribution and coverage of different numbers of local descriptors. Local descriptors are centered at the dots and their coverage is indicated by green bounding boxes. In particular, blue dots denote the 18 keypoints generated by a pose estimator and yellow dots denote the interpolated keypoints. Denser local descriptors introduce higher coverage of human body. **Right-top:** the mechanism of local perceptual loss back-propagation. Two corresponding local regions $\hat{X}_d$ and X'_d are respectively cropped from generated image $\hat{X}$ and ground truth image X'. $\hat{X}_d$ and X'_d are then separately passed through a pre-trained VGG-16 to generate activations ϕ at different layers l. A customized criterion $C(\phi, \phi')$ measures the distances between corresponding activations ϕ. Local descriptors intensify local loss back-propagation and thus enhance local details: see the sharper wrinkles and belt depicted in X'_d. **Right-bottom:** the detail enhancement introduced by increasingly denser overlapping local descriptors. Specifically, the four zoom-in views of logo are respectively cropped from outputs of four models trained using 0, 18, 31, 44 local descriptors (from left to right). Subtle but evident improvement can be identified in the process of increasing the number of overlapping local descriptors.

high resolution (1024^2), the encoder and decoder are grown progressively as training proceeds (section 3.3) and a super resolution (such as SRGAN [21], SinGAN [33]) can be additionally applied to further sharpen simple garment textures.

3.1. Pose-guided high-resolution appearance transfer

Pose representation To represent human pose information in an explicit manner, we employ a state-of-the-art pose estimator [3], which gives the locations of 18 keypoints of a person in 2D coordinates. To let the network leverage the keypoint information effectively, these 18 keypoints are separately represented by a gaussian distribution map with a fixed standard deviation. Specifically, we denote each keypoint as $k = 1, \cdots, 18$ and their respective 2D coordinates as $u(k)$. Then the pose representation P, which is the concatenation of 18 gaussian distribution maps, is encoded as:

$$P(\mathbf{x}; k) = \exp(-\frac{1}{2\sigma^2}||\mathbf{x} - u(k)||^2) \tag{1}$$

The result is an explicit pose representation $P \in \mathbb{R}^{H \times W \times 18}$ whose 18 maxima represent the locations of the 18 keypoints. P is then concatenated into the bottleneck of autoencoder.

Autoencoder The goal of the autoencoder is to reconstruct $\hat{X}$ in the target pose based on the appearance of the person in the reference image X and the pose representation P extracted from the same person in the ground truth image X', as shown in Figure 3. Since P contains no appearance information, the network is forced to utilize the appearance information in X. Furthermore, we add skip connections similar to those in a U-Net [32] to enable smoother gradient flow along the autoencoder. Then we use reconstruction loss between output $\hat{X}$ and ground truth image X' to encourage the network to generate appropriate appearance which matches the pose of the person in X'.

Perceptual loss The design of reconstruction error is critical for good performance. Since it is hard for the network to learn a pixel-to-pixel mapping only from X due to the inherent pose and appearance variation, we encourage the network to also learn high-level semantic meanings during training, which is pivotal for decoupling pose and appearance. Inspired by recent excellent practices [16], we adopt perceptual loss as the reconstruction loss between X' and $\hat{X}$. Apart from comparing only the raw pixel values, perceptual loss involves passing the output and the ground truth images individually through a pre-trained deep network and comparing the activations extracted from multiple layers inside the network. This process enables the network to better learn the decoupling of appearance and pose and alleviates overfitting. Specifically, we define perceptual loss as:

$$L(X', \hat{X}) = \sum_l C(\phi_l(X'), \phi_l(\hat{X})) \tag{2}$$

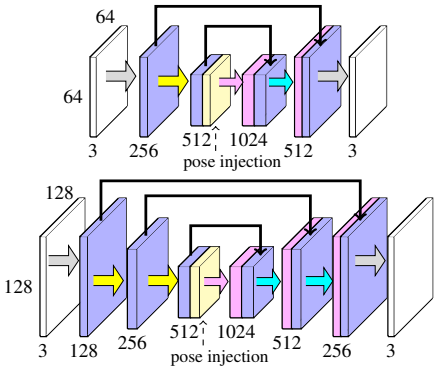

Figure 5: Progressive training. The bottleneck size of the autoencoder is 32×32. We start from a low spacial resolution of 64×64 pixels and incrementally add layers to encoder and decoder as training proceeds until we reach the ultimate resolution of 1024×1024. All existing layers remain trainable throughout the process. Here we illustrate a snapshot when the network increases its resolution from 64×64 to 128×128. During this transition, a new convolution block [(Conv + BN + Leaky ReLU) $\times$ 2] with corresponding up-sampling or down-sampling layer is introduced to encoder and decoder respectively. Also 1×1 convolution layer used to project RGB channels to/from feature space is replaced by a new one that fits the network.

where $\phi(x)$ is a pre-trained network, such as VGG-16 [35], and ϕ_l denotes the activation of the l^{th} layer of $\phi(x)$. Different from common practices which use L_2 loss as the criterion to evaluate $\hat{X}$, we customize the criterion $C(\phi, \phi^{'})$ to accelerate network convergence. Since L_2 loss has an optimal solution while L_1 loss enforces sharper output but is less stable, we designate $C(\phi, \phi^{'})$ as L_2 loss in the first half of the training process within each resolution level and L_1 loss in the second half. This practice enables stable convergence as well as high generation quality.

3.2. Local descriptors

Note that the adoption of global perceptual loss does not enforce sufficient preservation of local details. It is observed that sharp garment textures cannot be preserved well under the restriction of global perceptual loss only, as will be shown in the ablation study in Figure 7. To address this limitation, we introduce novel local descriptors which enable the generation of high-quality images. Local descriptors describe a set of regions telling the network where to focus and concentrate loss back-propagation. The locations of local descriptors are guided by the pose keypoints produced by the pose estimator. To ensure appropriate detail refinement and alleviate overfitting, the size of local regions is designed to be one-eighth of the input image resolution.

Figure 4 shows the distribution and coverage of local descriptors. Since higher resolution generally requires more local details, we increase the number of local descriptors adopted by interpolating between existing keypoints as input image resolution grows. Denser overlapping local descriptors introduce more complete coverage of the body and thus help preserve details more faithfully. Since we are not interested in fingers so we do not incorporate the keypoints particularly there.

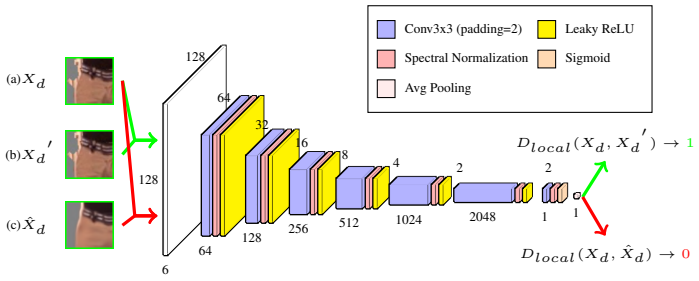

Figure 6: Local Discriminators. X_d, $X_d{}'$ and $\hat{X}_d$ are cropped from the reference image, the target image (GT) and the generated image respectively. The concatenation of X_d and $X_d{}'$ is fed into the local discriminator as a positive example, while the concatenation of $X_d{}'$ and $\hat{X}_d$ is fed as a negative example. Shown here is the architecture of the local discriminator with the last 2x2x1 feature vector being averaged to a scalar as the probability of the input pair being a valid transfer result.

Two kinds of local descriptor are adopted, respectively local perceptual loss and local discriminator. We use local perceptual loss as local descriptors during progressive training and local discriminators at the highest resolution (1024^2).

Specifically, based on the 18 keypoints in $X^{'}$ produced by the pose estimator, a list of N local descriptors is generated, denoted as $d = 1, \cdots, N$. Then two sets of fractional-sized regions centered at the location of each of N local descriptors are cropped from $X^{'}$ and $\hat{X}$ respectively.

Local perceptual loss Perceptual loss is enforced between corresponding local regions. The local loss L_{local} is formulated as the following:

$$L_{local}(X^{'}, \hat{X}) = \sum_{d=1}^{N} \sum_{l} C(\phi_l(X_d^{'}), \phi_l(\hat{X}_d))) \quad (3)$$

where $X_d{}'$ and $\hat{X}_d$ denote the d^{th} region cropped from $X^{'}$ and $\hat{X}$ respectively.

Local discriminators We adopt local discriminators in replacement of the local perceptual loss at the highest resolution (1024^2). Specifically, the local discriminators at the d^{th} region take pairs of inputs where the concatenation of X_d and $X_d{}'$ is considered real and the concatenation of X_d and $\hat{X}_d$ is considered fake, as shown in Figure 6. It is observed that local discriminators can improve the generalization ability of the model and further boost its generation quality.

Self-comparison between the model with and without local descriptors are shown in Figure 7. The significant improvement in image quality demonstrates the efficacy introduced by local descriptors.

3.3. Progressive training of autoencoder

Apart from achieving high-quality image generation, we also aim at producing unprecedentedly high-resolution results (1024^2). However, training the autoencoder in high resolution from scratch does not yield satisfactory results. Inspired by [12] which produces high-resolution results on CelebA-HQ dataset by introducing progressive training to GAN, we adopt a variation of progressive training which fits our setting of autoencoder with skip connections, as shown in Figure 5. Most importantly, instead of fading in a new convolution block to increase resolution using alpha blending, we train the new convolution block with skip connection from scratch, utilizing deeper convolution blocks trained in the previous stage as mature feature extractors. From our observation, this enables faster convergence of newly introduced blocks as well as utilization of skip connections to enhance generation quality. Self-comparison in Figure 7 demonstrates substantial improvement brought by progressive training on autoencoder.

3.4. Implementation details

We use the Adam [18] optimizer with a weight decay of 5×10^{-4}. The initial learning rate is set to 2×10^{-4}. We use $\sigma = 3.2$ to generate the gaussian distribution for pose representation. The autoencoder is trained progressively starting from the resolution of 64^2 with bottleneck shape of 1024×32^2 and ending at the resolution of 1024^2. Within each convolution block, we use two contiguous sets of 3×3 convolution layer followed by batch normalization [10] and leaky Relu with leakiness of 0.2. The number of channels of feature maps is halved as spacial size doubles. We downscale and upscale the feature maps using average pooling and nearest neighbor interpolation respectively. We use 1×1 convolution to project the outermost feature maps into RGB space and vice versa as in RGB back to feature map. We use He's initializer [11] to initialize the autoencoder. A total of 18 local descriptors are used for the resolution of 64^2 and 128^2. For 256^2 and 512^2, we use 31 local descriptors by interpolating between keypoints pairs and 44 local descriptors for 1024^2 through additional interpolations. For each resolution level, we train the network for 700 thousand iterations.

Our final loss L, which is composed of both global loss L_{global} and local loss L_{local}, is formulated as the following:

$$\begin{aligned} L(X^{'}, \hat{X}) &= L_{global}(X^{'}, \hat{X}) + L_{local}(X^{'}, \hat{X}) \\ &= \sum_{l} C(\phi_l(X^{'}), \phi_l(\hat{X}))) \\ &+ \sum_{d=1}^{N} \sum_{l} C(\phi_l(X_d^{'}), \phi_l(\hat{X}_d)) \end{aligned} \tag{4}$$

4. Experiments

To demonstrate the advantages of our method, we first conduct qualitative and quantitative self-comparisons to validate the effectiveness of different components, namely local descriptors and progressive training on autoencoders. We then demonstrate the network's generalizability by showing the results on various datasets, including the Human3.6M [13] dataset, the DeepFashion [25] dataset and a self-collected dataset from YouTube. We also compare the performance on the DeepFashion dataset with previous work. Lastly, we show the network's potential to be further utilized in real-world applications, such as high-quality garment transfer and pose-guided human video generation.

Table 1: Quantitative self-comparison in different modes.

	Human3.6M		
Model	**SSIM**	**local-SSIM**	**LPIPS**
Baseline	0.909	0.699	0.230
LPL	0.944	0.744	0.205
PT	0.953	0.759	0.164
PT+LPL	0.954	0.772	0.145
PT+LPL+LD	**0.959**	**0.804**	**0.135**
Real Data	1.00	1.00	0.00

4.1. Datasets

Human3.6M We train and test our model primarily on the Human3.6M dataset [13], which includes 11 actors in total with different poses. The dataset provides ground truth 2D human poses, backgrounds and human body bounding boxes. We first subsample the videos at 3 frames per second and obtain image frames with large pose variations. For each image frame, we then subtract the background and retain only the human foreground to reduce training noises. We select 'Posing', 'Greeting' and 'Walking' action classes for training, and 'Directions' class for testing.

YouTube dataset To test the generalizability of our method, we further train and test our network on our self-collected YouTube video datasets. The datasets we collected contains 20 hip-hop dancing videos from World of Dance competition, all of which have large pose variations. We subtract the background of this dataset using human parsing network [23] and subsample the videos at 3 frames per second to produce the training and testing set.

4.2. Self-comparison

Local descriptors Qualitative comparison in Figure 7 demonstrates the effectiveness of local descriptors (local perceptual loss and local discriminators). As shown in column (e), (g) and their corresponding zoom-in views, local perceptual loss results in improvement on local details compared to the baseline. From column (i), (k) and their corresponding zoom-in views, local perceptual loss is still able to bring significant enhancement to the generation quality under progressive training. In particular, the two stars in image (3, *d*) are faithfully preserved in result (3, *l*), but lost in result (3, *j*). In addition, column (k), (m) and their corresponding zoom-in views show the further enhancement on the generation quality of human body and garment textures introduced by the local discriminators applied at the highest resolution (1024^2).

Progressive training The advantages of progressive training is demonstrated through comparisons in Figure 7. Col-

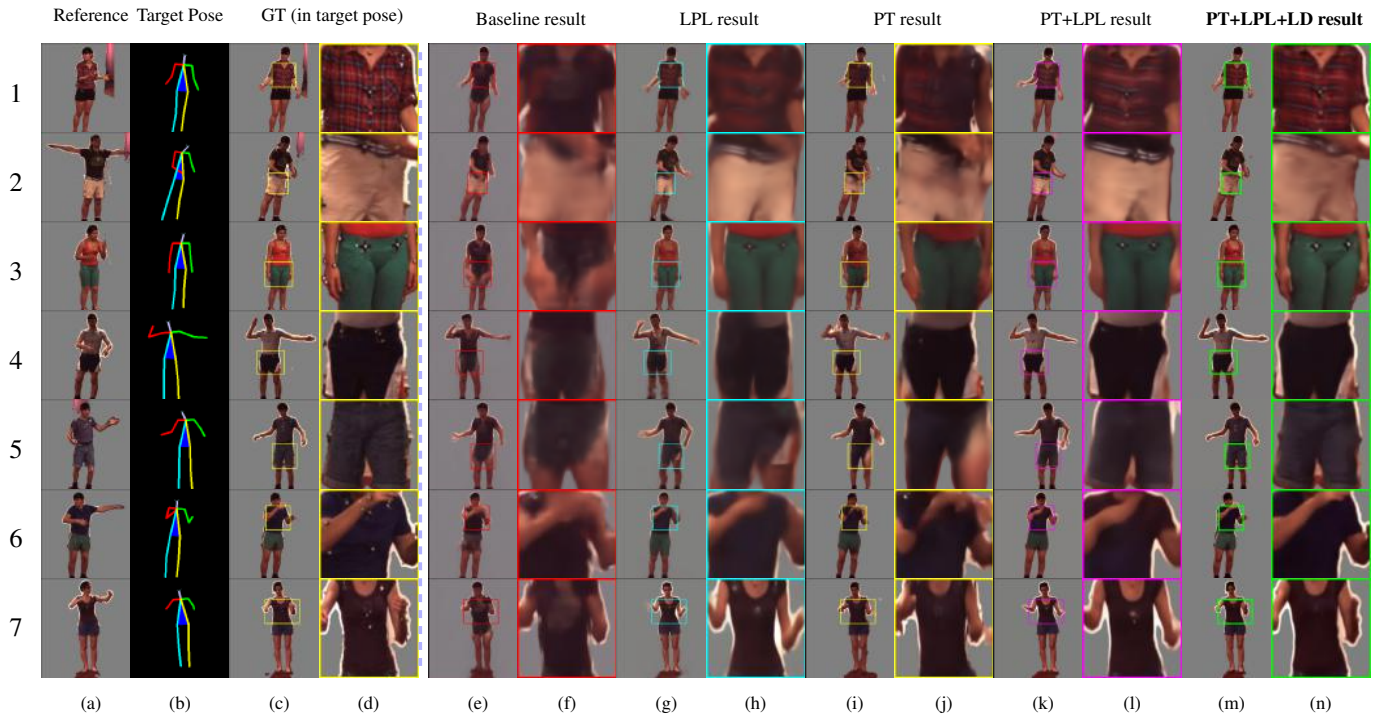

Figure 7: Self-comparison results. Test results on the Human3.6M generated by Baseline (no local descriptors or progressive training), LPL (with local perceptual loss only), PT (with progressive training), PT+LPL (with progressive training and local perceptual loss), PT+LPL+LD (with progressive trainning, local perceptual loss and local discriminators) and their corresponding zoom-in views are provided. Progressive training and local descriptors (local perceptual loss + local discriminator) each introduces considerable improvement in generation quality and produce the best result when combined. Our model also demonstrates robustness to the segmentation error introduced by the Human3.6M dataset. Figure is best viewed online.

umn (g) and (k) with their corresponding zoom-in views show the improvement for the models with local descriptors, while column (e) and (i) with zoom-in views show the improvement without local descriptors. In particular, the garment texture in image (1, *d*) is faithfully preserved in result (1, *l*), but lost in result (1, *h*). Therefore, while local descriptors enable local detail enhancement, the entire network still suffers from the vanishing gradient problem. To alleviate this problem, progressive training enables separate and progressive convergence of deep network layers, thus reduce the effects of vanishing gradient.

Quantitative comparison Image generation quality can be difficult to assess due to various standards. Here we adopt Structural Similarity (SSIM) [38] and Learned Perceptual Image Patch Similarity (LPIPS) [42] as our main evaluation metric. Due to the limitations of SSIM such as insensitivity and distortion under-estimation near hard edges [30], we also adopt a variation of SSIM, local-SSIM, to more effectively evaluate local details. Instead of global evaluation performed by SSIM, local-SSIM operates on 44 corresponding local regions between the generated image and the reference image. The 44 local regions correspond to the areas described by 44 local descriptors, where the highest coverage of human body is achieved.

Quantitative comparison between models under different settings are shown in Table 1. Both local descriptors and progressive training bring considerable enhancement in generation quality, with a combination of the two further boosting the result. Local-SSIM more evidently reflects the improvement in the quality at local regions.

4.3. YouTube dataset

Since appearance variation of the Human3.6M dataset is extremely limited (*i.e.* many actors wear garment with plain texture), we train and test our model on our YouTube dataset to validate its generalizability. Test results are shown in Figure 1. Our method clearly demonstrates its power in detail preservation and pose manipulation even under hard situations where large pose variations introduce different self-occlusions and thus dis-occlusion.

4.4. Comparison with previous work

Here we compare our method with three state-of-the-art approaches, DSC proposed by Siarohin *et al.*, Dense Intrinsic Appearance Flow (DIAF) proposed by Li *et al.* and Liquid Warping GAN (LWGAN) proposed by Liu *et al.* They reported their results on the market-1501 [44] dataset and the DeepFashion [25] dataset. Since our method focuses on high-resolution pose transfer, we only compare with them on the DeepFashion dataset, which has relatively high resolution (256^2). Figure 8 shows a qualitative comparison. Our method successfully transfers sharp appearance details to

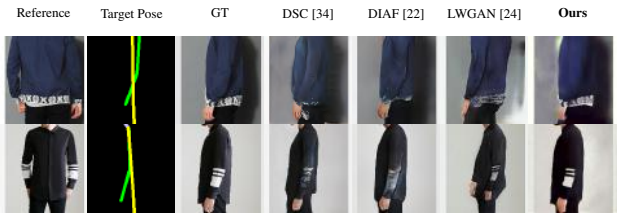

Figure 8: Comparison with previous work. Comparing with previous work, given the large pose differences our network captures more details despite that some parts may not be sharp enough.

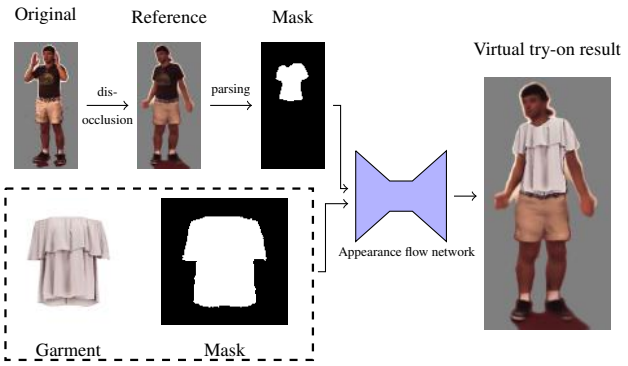

Figure 9: Garment Transfer Network. We first use our Our method to dis-occlude the original image and apply human parsing network [23] to extract the mask of the garment in reference and target poses. The appearance flow network will then take the garment image, reference pose mask and target pose mask as input and outputs the appearance flows, which yield the synthesized garment image through a bilinear sampling layer.

large pose variations and faithfully hallucinates previously occluded parts. Table 2 shows our better quantitative performance compared with their methods.

Table 2: Quantitative comparison with previous work.

DeepFashion			
Model	**SSIM**	**MS-SSIM**	**LPIPS**
DSC	0.776	0.792	0.345
DIAF	0.778	0.798	0.252
LWGAN	0.781	0.788	0.227
Ours	**0.806**	**0.814**	**0.198**
Real Data	1.00	1.00	0.00

4.5. Further Applications

Virtual try-on Virtual try-on has demonstrated great application potential. This task requires the transfer of any garment with detailed texture. While a recent approach [9] successfully preserves garment details and shapes, there still exists artifacts due to self-occlusions. We can tackle this problem with two steps. First, we transfer the image of the target person (with self-occlusion) into a pre-defined frontal pose (without occlusion). Then, we apply our appearance flow network to transfer the garment to that person. Details are in Figure 9.

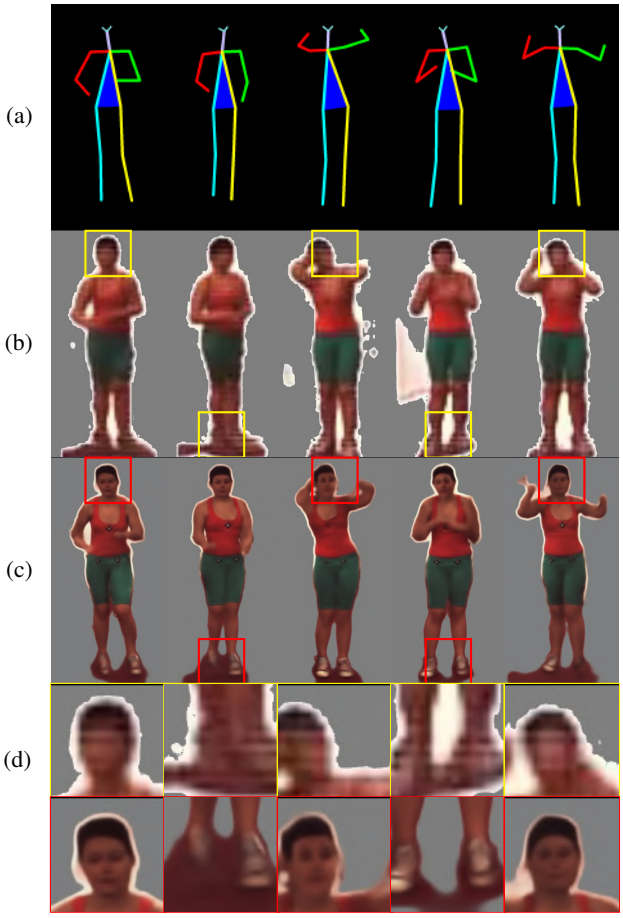

Figure 10: Pose-guided human video generation. five sample frames of (a) input target poses, (b) generated video by [2], (c) our generated video, (d) zoom-in views of the two methods.

Pose-guided human video generation Our method can be applied to generate human action *videos* in 1024^2 under a pose guidance. Specifically, given a reference image and a video sequence of target poses, a high-resolution video of the reference person complying the target pose frames is generated. Although we generate the video frame-by-frame and do not consider temporal conherency as done in [2] on human video generation, their video resolution was only 128^2. Figure 10 shows a comparison on 5 subsequent frames, where our result shows the facial features as well as other details of the garment and shoes. Our method demonstrates a great potential in high-resolution human video generation.

5. Conclusion

In this paper we present a solution for pose-guided high-resolution appearance transfer between images, where the proposed local descriptors (local perceptual loss + local discriminators) and progressive training on autoencoder are shown to be effective in generating plausible and photorealistic images of human at target poses. Self-comparisons have clearly validated the advantages of different compo-

nents. We have also demonstrated our method's high generalizability in our extensive experiments. We have also shown important applications in high-quality garment transfer and pose-guided human video generation.

References

[1] P. Baldi. Autoencoders, unsupervised learning and deep architectures. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, pages 37–50. JMLR.org, 2011.

[2] H. Cai, C. Bai, Y.-W. Tai, and C.-K. Tang. Deep video generation, prediction and completion of human action sequences. *ECCV*, 2018.

[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[4] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.

[5] B. Cheung, J. Livezey, A. Bansal, and B. Olshausen. Discovering hidden factors of variation in deep networks. In *ICLR workshop*, 2015.

[6] R. de Bem, A. Ghosh, T. Ajanthan, O. Miksik, N. Siddharth, and P. H. Torr. DGpose: Disentangled semi-supervised deep generative models for human body analysis. In *arXiv preprint arXiv:1804.06364*, 2018.

[7] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[9] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

[12] A. Heljakka, A. Solin, and J. Kannala. Pioneer networks: Progressively growing generative autoencoder. In *ACCV*, 2018.

[13] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:1325–1339, 2014.

[14] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[15] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Conditional image generation for learning the structure of visual objects. In *arXiv preprint arXiv:1806.07823*, 2018.

[16] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[19] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[20] C. Lassner, G. Pons-Moll, and P. V. Gehler. A generative model of people in clothing. In *ICCV*, 2017.

[21] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

[22] Y. Li, C. Huang, and C. C. Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, 2019.

[23] X. Liang, K. Gong, X. Shen, and L. Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[24] W. Liu, W. L. L. M. Zhixin Piao, Min Jie, and S. Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[25] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. *CVPR*, pages 1096–1104, 2016.

[26] L. Ma, X. Jia, Q. Sun, B. Schiele, and L. V. G. Tinne Tuytelaars. Pose guided person image generation. In *NIPS*, 2017.

[27] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *CVPR*, 2018.

[28] N. Neverova, R. A. Guler, and I. Kokkinos. Dense pose transfer. In *ECCV*, 2018.

[29] A. Odena, C. Olah, , and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *arXiv preprint-arXiv:1610.09585*, 2017.

[30] J. F. Pambrun and R. Noumeir. Limitations of the ssim quality metric in the context of diagnostic imaging. *ICIP*, pages 2960–2963, 2015.

[31] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *NIPS*, 2016.

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. 2015.

[33] T. Rott Shaham, T. Dekel, and T. Michaeli. Singan: Learning a generative model from a single natural image. In *Computer Vision (ICCV), IEEE International Conference on*, 2019.

[34] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe. Deformable gans for posebased human image generation. In *CVPR*, 2018.

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[36] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, pages 3483–3491, 2015.

[37] L. Tran, X. Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, 2017.

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

[39] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, 2016.

[40] M. Zanfir, A.-I. Popa, A. Zanfir, and C. Sminchisescu. Human appearance transfer. In *CVPR*, 2018.

[41] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[43] B. Zhao, X. Wu, Z. Cheng, H. Liu, and J. Feng. Multi-view image generation from a single-view. In *arXiv preprint arXiv:1704.04886*, 2017.

[44] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. *ICCV*, pages 1116–1124, 2015.

[45] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[46] R. S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. Chen. Be your own prada: Fashion synthesis with structural coherence analysis. In *ICCV*, 2017.