High-Resolution Pose Transfer via Progressive Training and Pose Disentangling

Anonymous CVPR submission

Paper ID ****

Abstract

This paper proposes a novel High-Resolution Pose Transfer Network (HPN) which transfers an arbitrary target pose with unprecedented image resolution (1024^2) to a reference person, given only an image of the same person with the target poase. Our HPN framework utilizes dense local descriptors to refine local details, which are trained progressively in a coarse-to-fine manner to produce the high-resolution output to faithfully preserve the complex appearance of garment textures and geometry, while transferring seamlessly the target pose including those with self-occlusion. Our progressive encoder-decoder architecture can disentangle pose from appearance inherent the input image at multiple scales. Extensive experimental results on Human3.6M [14], DeepFashion [26] and our dataset collected from YouTube show that our model produces high-quality images, which can be further utilized in useful applications such as high-quality garment transfer between different persons and pose-guided person video generation.

1. Introduction

Learning 3D information inherent in the 2D image domain is a fundamental problem in computer vision. This problem is fundamental in many computer vision tasks such as scene understanding [24], instance segmentation [5, 23] and action recognition [38, 8], to name a few, while remaining a major challenge for deep neural network learning. The challenge lies in the fact that images are 2D projections of the corresponding 3D world where objects can undergo complex deformation and occlusion.

This paper focuses on images of humans, whose different poses introduce complex non-rigid deformation and self-occlusion. Specifically, given a reference image of a person and another image of the same person in a target pose, our method seamlessly transfers the target pose to the reference person 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 Figure 1. Note that the network must not only moves 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. To this end, the network must learn to disentangle the structure and appearance of the reference person from the given image. This is particularly challenging for human images due to the non-rigid nature of 3D human body, and complex texture and geometry distortion on the 3D garment worn by humans.

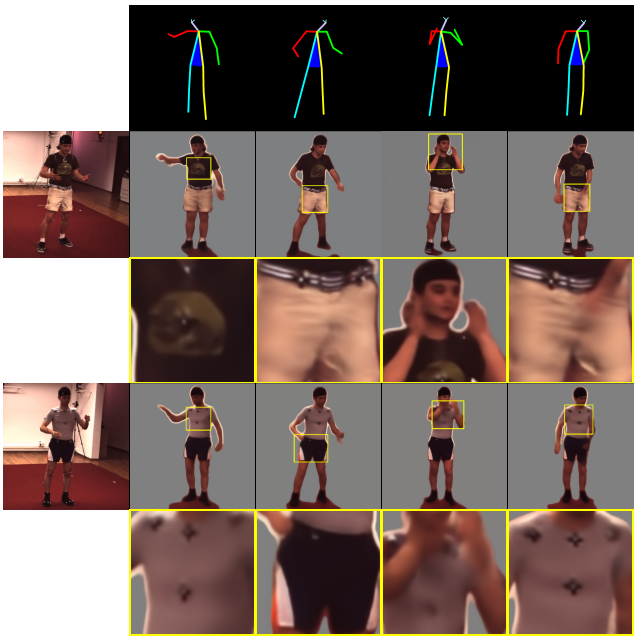


Figure 1. **Transfer results.** Given a reference image (leftmost column) and target poses (top row) as input which contains self-occlusion with complex appearance in texture and geometry, our HPN can transfer the target pose to the reference person in high resolution preserving high level of details. (a), (b) demonstrate the effects of dis-occlusion and (c), (d) demonstrate the effects of transferring to other self-occluded poses with zoom-in shown in detail.

Many recent works seem to provide a plausible solution to our human pose transfer task. Conditional Generative

Adversarial Networks (GANs) [15], for example, have been exploited to effectively solve similar tasks such as generating MNIST [22] digits given labels. In particular, they generate sharp and realistic images based on certain preconditions by minimizing an adversarial loss. However, they can only generate images in accordance to the training distribution, but fail to reconstruct or hallucinate unseen details, and thus are not applicable to our task which may involve dis-occlusion to reveal unseen details. Though the other plausible generative model – the Variational Autoencoders (VAEs) [20] – can generate results complying a given reference image, they may not adequately preserve high-quality details in the reference image, due to the fact that the related method maximizes only a lower bound.

To address these limitations, we propose the High-resolution Pose Transfer Network (HPN), which is effective in disentangling structure and appearance information inherent in a given reference image, and faithfully transfers the original appearance of the person according to the target pose representation. Figure 2 gives an overview of HPN. Specifically, we inject the target pose representation into the bottleneck of our encoder-decoder architecture for disentangling pose and appearance. Then we adopt local descriptors on image regions to encourage the network to learn more details for enhancing generation quality. Furthermore, progressive growth is employed on both encoder and decoder sides to increase output resolution. During training, perceptual loss [17], globally and locally applied around the regions of local descriptors, is used to compare between the output and the ground truth in the target pose.

To validate our approach, we conduct extensive experiments on Human3.6M [14], DeepFashion [26] and our dataset collected from YouTube, and our results show that the HPN outperforms current state-of-the-art generative models. We also apply HPN to other applications such as high-quality garment transfer and pose-guided person video generation. Our results demonstrate the high potential of HPN in many challenging tasks.

Our contribution is three-fold: 1) to disentangle structure and appearance inherent in a given reference image, we propose a new encoder-decoder architecture that successfully enables seamless human pose transfer; 2) we propose novel local descriptors to enhance the generation quality and local details; 3) we apply progressive training to our autoencoder architecture to achieve outputs of unprecedented high resolution (1024^2). To our knowledge, this is the first progressive, deep autoencoder transfer network that can realistically hallucinate in 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

Conditional image generation Generative models including Variational Autoencoders [20] (VAEs) and GANs[9] had demonstrated great power in image generation.

Human pose transfer is closely related to the problem of conditional image synthesis as it requires a target pose as an output constraint. Zhao *et al.* [42] integrated GANs and other inference models to generate images of persons in various clothing styles from multiple views. Reed *et al.* [33] proposed a conditional generative model that used pose and text as conditions to generate images. Lassner *et al.* [21] also presented a generative model based on human pose that could generate realistic images conditioning on clothing segmentation.

Numerous researchers [36, 31, 40, 41, 45] introduced their respective methods to allow more control on the appearances of the generated images in generative processes by providing different intermediate information such as labels and texts. Models such as ConditionalGAN [15] and CycleGAN [44] also demonstrated their efficacy in image-to-image translation. Yet, compared with our feedforward autoencoder, they and GANs in general are relatively more difficult to train, which often cannot faithfully transfer intrincate patterns and textures from reference images.

In general, 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 reference images is an essential step. Many previous studies [3, 4] attempted to use GANs [9] 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.

Pose transfer There has been much work on pose transfer. Some approaches for pose transfer [6, 28] 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 [30] on transferring the pose of a reference image invariant with its appearance. The PG^2 [27] was a more related work that aims at generating images of a subject in various poses based on an image 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, which added a discriminator 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.* [29] attempted to produce the seamless result by blending the synthesized im-

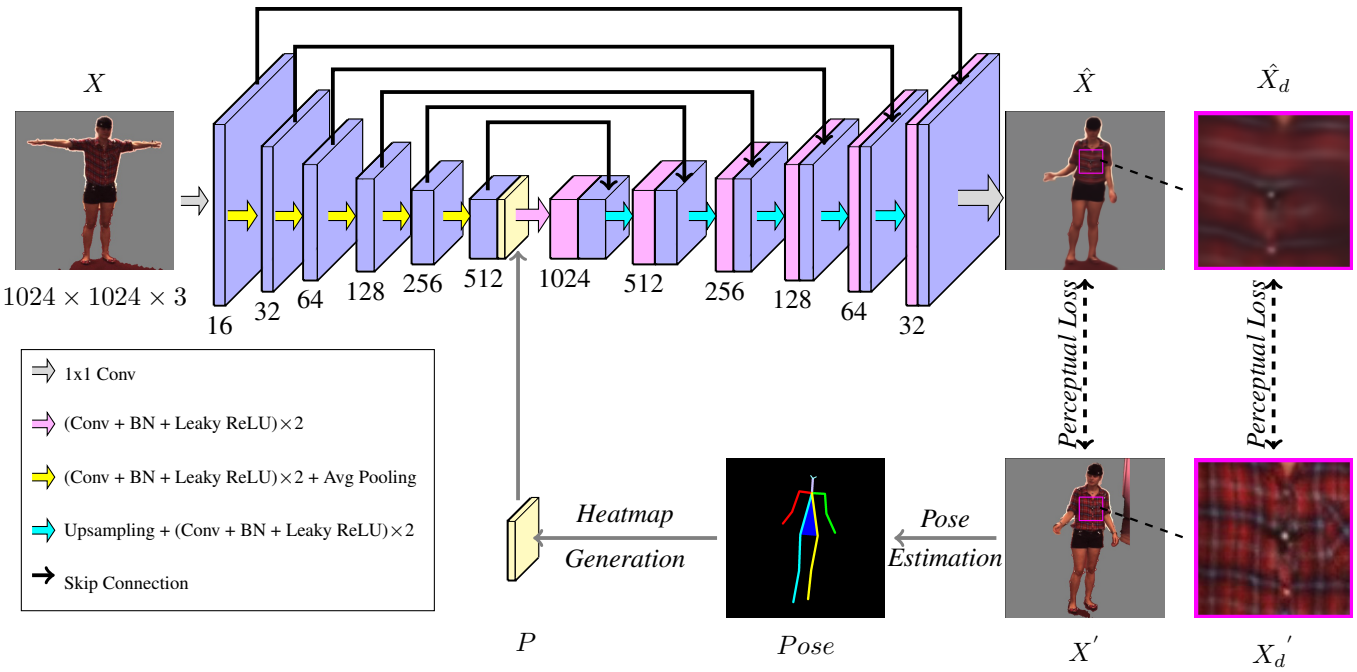


Figure 2. **Network Architectures.** The reference image X is first passed through an encoder to generate a latent representation. In the lower branch, 18 keypoints are estimated from the ground truth image $X^{'}$ to produce an explicit pose representation P. P is then concatenated with the latent representation, which is further decoded into the output $\hat{X}$. Global perceptual loss is enforced between $X^{'}$ and $\hat{X}$. To improve local details, local perceptual loss is also enforced on the corresponding local regions ($X_d{}^{'}$, $\hat{X}_d$), indicated by the bounding boxes.

age 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.* [16] actually demonstrated acceptable results on pose transferring. This is achieved by using a simple encoder-decoder network with the learning landmarks concatenated in an intermediate representation.

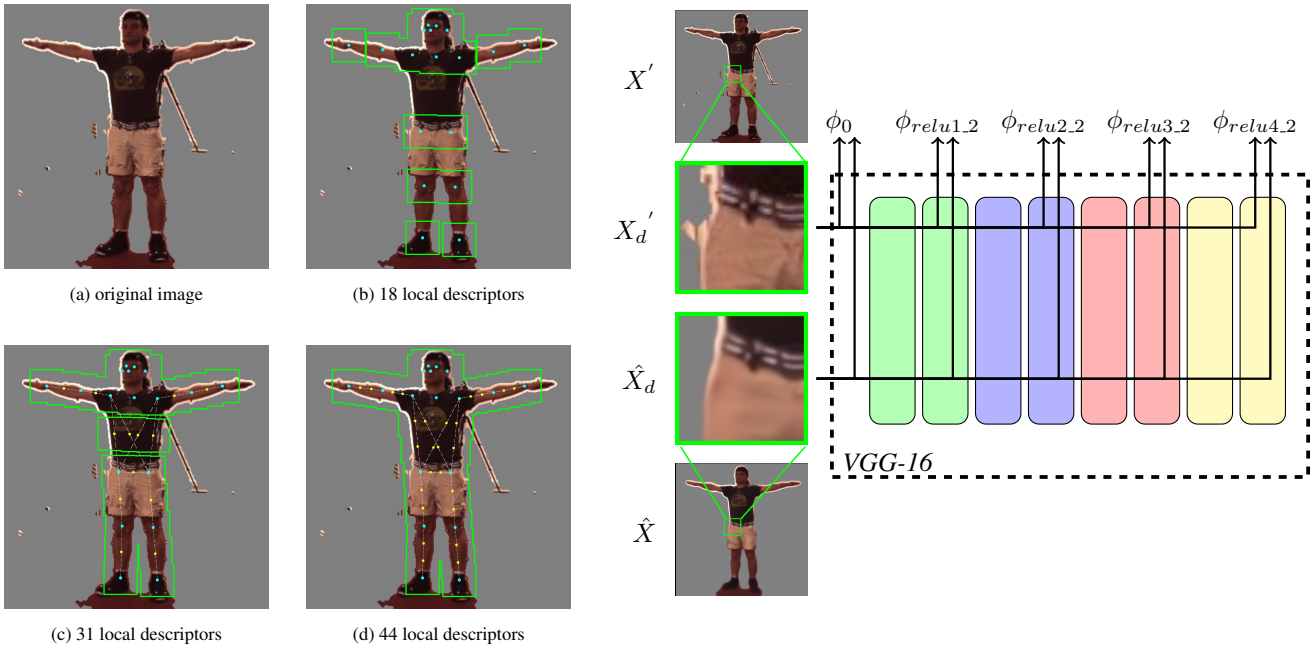
In general, comparing to [27, 34], our method does not use sophisticated GANs which may introduce unstabilizing factors to the training process. More importantly, although [27, 34, 29, 16] performed well on changing pose at low-resolution (128 × 128) reference images while keeping their rough identity, they could not preserve but significantly blur complex textures after pose transfer.

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.* [18] proposed a progressive training methodology for generative adversarial networks to generate high-quality results. They started training from low resolution and added layers to the model progressively to obtain satisfactory high-resolution results. Ari *et al.* [13] also introduced an progressive architecture for autoencoder to encode and reconstruct high-quality images (up to 256^2). They focused on how to train the autoencoder progressively for image reconstruction and image generation from the random sample while our goal is conditional image generation for even higher resolution output (1024^2).

3. Method

Our goal is to transfer the pose of a reference person to a given target pose with high quality. This task is achieved by the disentanglement of appearance and pose of the person in the reference image through an autoencoder architecture. Our network architecture is shown in Figure 2.

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 (Sec 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 $\hat{X}$ to enforce pose transfer learning (Section 3.1). To improve generation quality, we adopt novel local descriptors to refine output details. Local



(a) original image (b) 18 local descriptors (c) 31 local descriptors (d) 44 local descriptors

Figure 3. **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:** it demonstrates the mechanism of local 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. Figure is best viewed online.

descriptors are applied under the guidance of keypoint locations from the pose estimator (Section 3.2). The same reconstruction loss is enforced locally at the corresponding regions described by local descriptors between $X^{'}$ and $\hat{X}$. To generate images in an unprecedented high resolution (1024^2), the encoder and decoder are grown progressively as training proceeds (Section 3.3).

3.1. High-resolution Pose Transfer Network

Pose representation To represent human pose information in an explicit manner, we employ a state-of-the-art pose estimator [2], 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:

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

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 2. Since P contains no appearance information, the network is forced to utilize the appearance information in X. Furthermore, we add skip connections similar to that in a U-Net [30] to enable smoother gradient flow along the autoencoder. Then we adopt 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 pose and appearance disentanglement. Inspired by recent excellent practices [17], we adopt perceptual loss as the reconstruction loss between $X^{'}$ and $\hat{X}$. Apart from comparing only the raw pixel values, perceptual loss passes the output and the ground truth im-

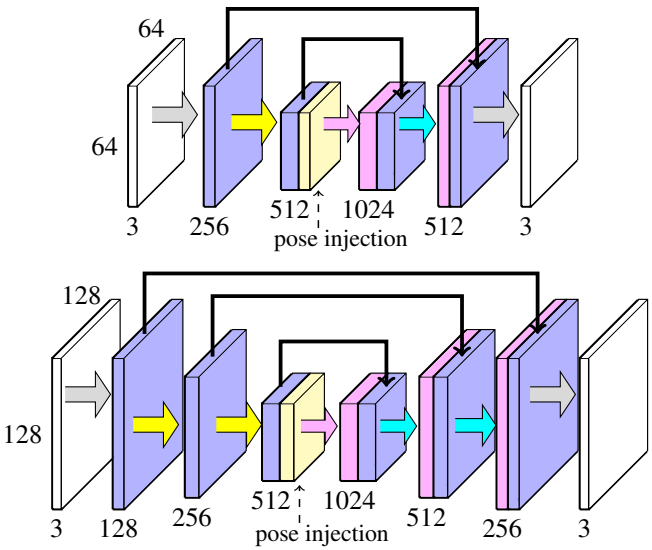


Figure 4. **Progressive training.** The bottleneck size of the autoencoder is 32×32. We start from a low spacial resolution of 64×64 pixels and incrementally add layers to encoder and decoder as training proceeds until we reach the ultimate resolution of 1024×1024. All existing layers remain trainable 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. 1×1 convolution layer used to project RGB channels to/from feature space is also replaced by a new one that fits the network.

ages individually through a pre-trained deep network and compares the activations extracted from multiple layers inside the network. This process enables the network to better learn the disentanglement of appearance and pose and alleviates overfitting. Specifically, we define perceptual loss as:

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

where $\phi(x)$ is a pre-trained network, such as VGG-16 [35], and ϕ_l denotes the activation of the lth 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

The adoption of perceptual loss does not enforce sufficient preservation of local details. It is observed that sharp garment textures cannot be well preserved under the restriction of global perceptual loss only, as shown in Figure 3. To address this limitation, we introduce novel local descriptors which enable 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. The same reconstruction loss is applied locally between the corresponding regions in $X^{'}$ and $\hat{X}$.

Figure 3 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.

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

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

where $X_d^{'}$ and $\hat{X}_d$ denote the dth region cropped from $X^{'}$ and $\hat{X}$ respectively. Self-comparison between the model with and without local descriptors are shown in Figure 3. The significant improvement in image quality demonstrates promising enhancement 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 [13] 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 4. 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 5 demonstrates substantial improvement brought by progressive training on autoencoder.

3.4. Implementation details

We use the Adam [19] 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 [11] 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 [12] 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 prove the advantages of the above proposed method, we first conduct qualitative and quantitative self-comparisons to validate the effectiveness of different components of HPN, respectively local descriptors and progressive training on autoencoders. We then demonstrate our generalizability by showing the results produced by HPN on various datasets, including human 3.6M [14], DeepFashion [26] and a self-collected dataset from YouTube. We also compare our performance on DeepFashion dataset with previous work. Lastly, we show our potential to be further utilized in real-world applications like high-quality garment transfer between different persons and pose-guided person video generation.

Datasets We train and test our model mainly on the Human3.6M dataset [14], which has 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 sample 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' action classes for training, and 'Directions' class for testing.

To test the generality of our method, we further train and test on our self-collected youtube video datasets. The datasets we collected has 10 dancing videos in total. All of them have large pose variations. We subtract the background of this dataset using JPPNet [25] and subsample the videos at 3 frames per second for training set as well as testing set.

4.1. Self-comparison

Local descriptors Qualitative comparison in Figure 5 demonstrates the effectiveness of local descriptors. From column (e), (g) and their corresponding zoom-in views, local descriptors introduce improvement both in global coherency and local details compared to the baseline. And from column (i), (k) and their corresponding zoom-in views, local descriptors are still able to bring significant enhancement to 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*).

Figure **??** provides the validation of local descriptors from another perspective, where the number of local descriptors adopted by each model steadily increases. Specifically, the four models are trained using 0, 18, 31, 44 local descriptors respectively. Subtle but evident improvement can be identified in the process of increasing the number of overlapping local descriptors.

Progressive training The advantages of progressive training is also demonstrated through comparisons in Figure 5. Column (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 for the models 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*). Even though local descriptors help enhance local details, vanishing gradient problem still persists in deep networks necessary for high-resolution image generation. Progressive training enables separate and progressive convergence of different layers in a deep network.

Quantitative comparison Image generation quality can be hard to assess due to various standards. Here we adopt Structural Similarity (SSIM) [39] as our main evaluation metric. Due to the limitations of SSIM such as insensitivity and distortion under-estimation near hard edge [32], we also adopt a variation of SSIM, local-SSIM, to more effectively evaluate local details. Instead of global evalu-

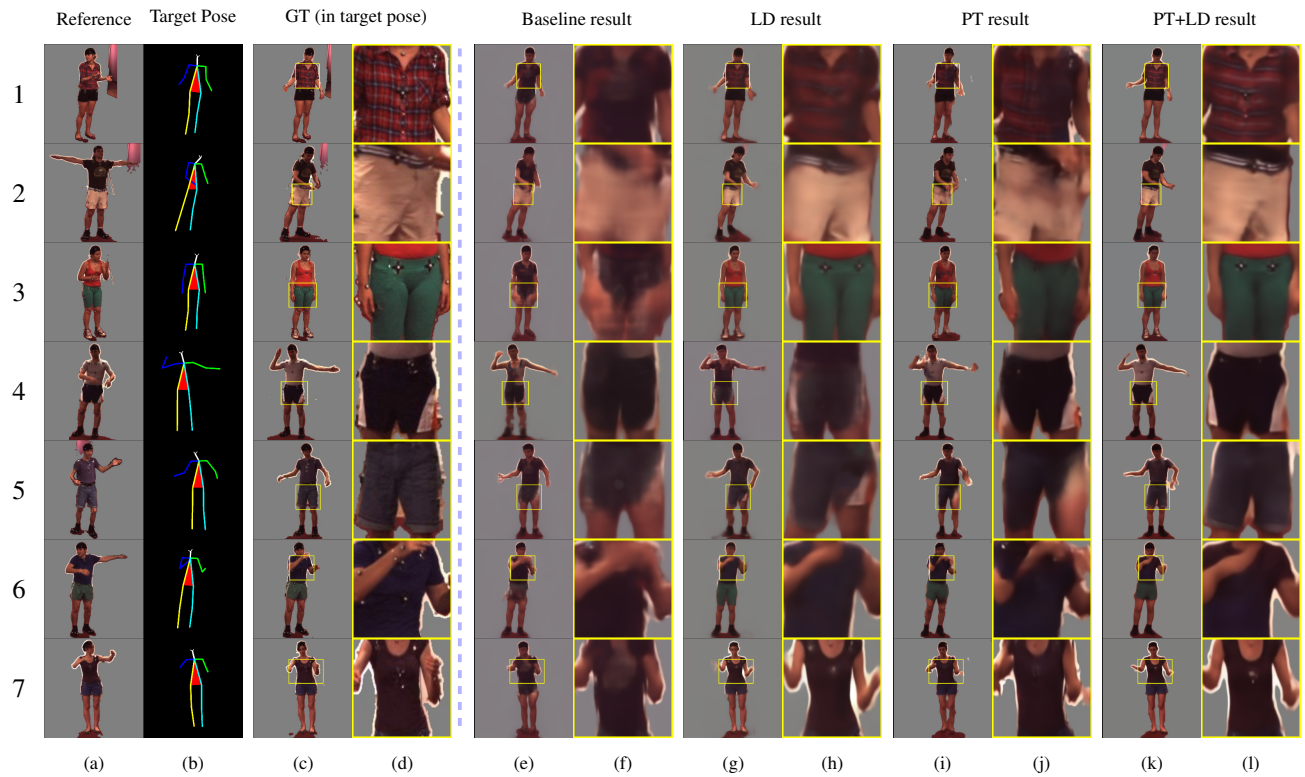


Figure 5. **Self-comparison results.** Test results on human 3.6M generated by Baseline (no local descriptors or progressive training), LD (with local descriptors), PT (with progressive training), LD + PT (with both local descriptors and progressive training) and their corresponding zoom-in views are provided. Local descriptors and progressive training 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 Human 3.6M dataset.

ation performed by SSIM, local-SSIM operates on 44 corresponding local regions between the generated image and 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. Either local descriptors or progressive training brings considerable enhancement in generation quality, with combination of the other further boost the result. Local SSIM more evidently reflects the improvement in the quality of local regions.

Table 1. Quantative self-comparison between different modes of our model.

	Human3.6M	
Model	**SSIM**	**local-SSIM**
Baseline	0.909	0.699
LD	0.944	0.744
PT	0.953	0.759
LD+PT	**0.954**	**0.772**
Real Data	1.00	1.00

4.2. Youtube dataset results

4.3. Comparison with previous work

We compare our results with the current state-of-the-art method (**DSC**) on DeepFashion dataset. From the Figure 7

Table 2. Quantative comparison witb previous work.

	DeepFashion
Model	**SSIM**
DSC	0.776
Ours	**0.806**
Real Data	1.00

4.4. Futher application

Virtual try-on Virtual try-on has seen great application potentials due to its convenience and reduction in cost. This problem involves the transfer of any garment with detailed and complex texture. While a recent approach [10] successfully preserve garment details and shapes, there still exist artifacts due to self-occlusions. With our HPN, we can tackle this problem with two steps. First, we transfer the

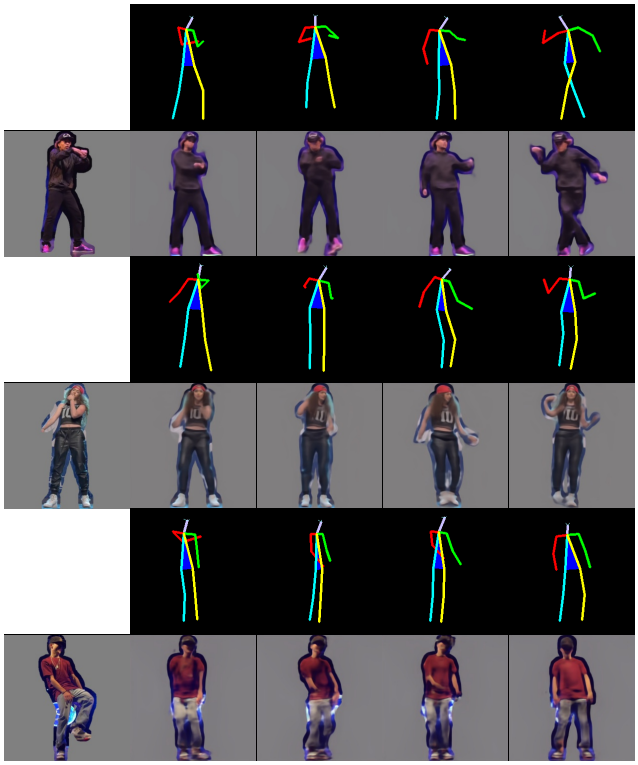


Figure 6. **Youtube results.**

image of the target person (with self-occlusion) into a predefined frontal pose (without occlusion). Then, we apply our appearance flow network to transfer the garment to that person.

Zhou *et al.* [43] first proposed the idea of appearance flows for view synthesis. Appearance flows are 2-D coordinate vectors describing how pixels in the input image could be used to reconstruct the image from the target viewpoint. They found that the images of the same target from different perspectives have high correlations with each other. Garment transfer is similar to view synthesis since the same garment piece in different human pose is also highly correlated. Inspired by their work, we propose an encoder-decoder framework to predict appearance flows for garment transfer. We first apply the JPPNet [25] to extract the mask of the garment in reference and target poses. The autoencoder will take the garment image, reference pose mask and target pose mask as input. Then the decoder outputs the appearance flows and yield the garment synthesized image through a bilinear sampling layer. We adopt perceptual loss between ground-truth image and synthesized image and update the layer's weight through backpropagation.

Video generation Since our network can transfer the target pose to the reference person in high quality, it can be applied to human video generation given a sequence of target poses.

Figure 7. **Comparison with previous work.**

5. Conclusion

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

[3] 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.

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

[5] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[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] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. In *NIPS*, 2015.

[9] 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.

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

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

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

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

[14] 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.

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

[16] 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.

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

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

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

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

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

[22] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

[23] G. Li, Y. Xie, L. Lin, and Y. Yu. Instance-level salient object segmentation. In *CVPR*, 2017.

[24] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.

[25] 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.

[26] 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.

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

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

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

[30] P. F. O. Ronneberger and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *Springer International Publishing, Cham*, pages 234–241, 2015.

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

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

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

[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] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, 2015.

[39] 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.

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

[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] 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.

[43] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros. View synthesis by appearance flow. 2016.

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

[45] 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.