Learn a Global Appearance Semi-Supervisedly for Synthesizing Person Images

Zhipeng Ge* Fei Chen* Sidan Du Yao Yu Yu Zhou
Nanjing University

Abstract

We present a novel approach for person images synthesis in this paper, that can generate person images in arbitrary poses, shapes and views. Unlike existing methods just using keypoints' locations in heatmaps format, we propose to render SMPL model to UV maps, which can provide human structural information about poses and shapes. Thus, by varying the parameters of poses, shapes and camera in SMPL model, we can generate different person images with various attributions in a simple way, while in most cases we can only obtain new shapes of people by computer graphics methods. We train an end to end generative adversarial network with unlabeled data. As our SMPL parameters come from a pretrained model, we call our overall network semi-supervised. Our network keeps a global appearance during the fine-tuning stage of the target person, thus we can get a complete appearance of the target person, rather than the inaccurate appearance caused by inferencing without enough information. Experiments on Human3.6M Dataset and a self-collected dataset demonstrate the excellent effectiveness of our approach on person images synthesis for different applications.

1. Introduction

With the development of deep generative networks, image synthesis [15] [4] [33] [43] becomes easier and more realistic. The most common task in person image synthesis is human pose transfer, which generates an image of the source person's pose with the target person's appearance. Many existing methods [2] [8] [20] [21] [25] [26] [40] for human pose transfer adopt an encoder-decoder architecture to learn the appearance of the person in an input image, with the 2D keypoints in a heatmap format representing human structure information as a guidance. However, 2D keypoints only contain skeleton information without shape information. Due to occlusion, 2D keypoints also can be inaccurate, which can lead to mismatching between appearance and human structure, and thus show an abnormal result in generated images.

*These authors contributed equally to this work

There are still many other applications for person images synthesis. For example, we can generate person images with various weight and height for data augmentation. We also can vary the views for a person which can be used in Augmented Reality(AR). However, some difficulties exist such like the pose and the appearance of one person in an image are often difficult to be totally separated and the new pose is also hard to correspond with the unpaired appearance. As inputting a single view image, when transferred to an unseen view, the unseen part about appearance of the original image may also be inferred inaccurate.

In this paper, we propose a two-stages method for person images synthesis. First, we use a pretrained HMR [17] model to get SMPL [24] parameters and then we can separate the human body with pose, shape and camera parameters, thus we can control the variety of the corresponding component of human body. After rendering a 3D model coming from SMPL parameters to UV maps, we concatenate the input images with the UV maps and train a UV-guided appearance encoder. The UV maps with human structure information will guide the latent code to learn proper appearance information, and we also utilize a crossing training between two images to enhance the separation between body structure and appearance. With background images and a decoder we can get our final generated images. All the network is trained in an adversarial way to get a more realistic image. After obtaining a network generalized well in poses, we fine-tune the network to enroll in the appearance information of the target person. In this way, we can keep a global appearance of the target person and can generate images in any unseen views with accurate appearance. As our method needs a set of images of a specific person to enroll the person appearance into the network for robust applications, we adopt Human3.6M [14] dataset which is in video format. We also collect some videos about some specific people by mobile phone to display more results about our method. We show excellent results on Human3.6M [14] and our self-collected dataset which is collected by mobile phone. Our main contributions can be summarized as follows:

i) We propose a novel framework that can separate appearance, pose and shape of a person in an image, thus we

can generate person images in arbitrary poses, shapes, and views.
ii) We propose an end-to-end and semi-supervised network with a UV-guided appearance encoder in a cross-training way to separate body structure and appearance information.
iii) We enroll a global appearance for a specific person in the network, and can generate unseen views for the input image.

2. Related Work

Deep generative networks. Deep generative networks have demonstrated a great progress in image synthesis, such as Generative Adversarial Networks (GAN) [9] and Variational Autoencoder (VAE) [19]. GAN has been widely used due to its capability of generating realistic images with sharp details. And there also develops many variants. CGAN [28] and InfoGAN [5] use labels to enhance the correlation between latent code and generated images. WGAN [1] proposes to use wasserstein distance to gain a stable training process. Works like [32] and [39] leverage text for image generation with GANs. Frameworks like pix2pix [15], CoGAN [22], CycleGAN [44] are targeted at image to image translations, while CycleGAN [44] achieves unsupervised image translation. Works like pix2pixHD [35] use multi-stages to improve generated images' quality in a coarse-to-fine way. As we aim at human pose and shape transfer, which can be also treated as image translation, we can make use of these basic frameworks to generate realistic images.

Human Pose Transfer. Human pose transfer usually obtains appearance from the target person images and can generate people in clothing [20] or new action sequences with the same appearance [3] after we give source person's poses. Ma et al. first demonstrates human pose transfer with the given 18 keypoints in heatmaps format. They concatenate the keypoints' heatmaps with source images and train a CNN network in an adversarial way. Zhao et al. proceed a coarse-to-fine process to get detailed output images. Balakrishnan et al. separates appearance into person and background, while Pumarola et al. proposes a fully unsupervised strategy to render images with new poses. Esser et al. trains a conditional U-Net for pose-guided image generation, conditioned on the output of a variational autoencoder for appearance. All of these works utilize only poses of people, but we introduce SMPL model to control not only poses but also shapes of the person in the images, thus the generated images with the transferred person can be in various shapes. Neverova et al. also leverage SMPL model, but they don't generate images with various human shapes.

Appearance Generation. To obtain the accurate appearance texture, the source pose needs to match the pose in the target image. Zanfir et al. proposes to fit a 3D body model to capture the body deformations and match the source and the target pose. The drawback of fitting a 3D model derives from high computational cost and an inaccurate result. As optical flow [13] demonstrates pixel-to-pixel correspondence between two adjacent images, appearance flow [42] shares the same point that two images exist dense correspondence with different view-points. Li et al. use appearance flow to perform feature warping on the input image. Neverova et al. render the 3D human model into meaningful UV-coordinate maps based on DensePose [10] and warped surface interpolation and inpainting explicitly back to the image space. Differently, we use UV-coordinate maps rendered from SMPL model to guide the image warping implicitly in a simpler way without high computational cost. Works like [26] [31] input only single image and try to separate appearance from human body structure with weak supervision. For lack of global appearance information, when generating unseen views the appearance will be inferred from the distribution of the datasets and cause wrong appearance. In our work, we fine-tune the network to preserve the appearance information completely and even if to generate unseen views, the network will also perform an accurate result same to the real image.

3. Method

3.1. Overall Network Architecture

Figure 1 illustrates our end-to-end semi-supervised person images synthesis framework. There are two steps to realize the synthesis. First, We train a network primarily generalize on various poses. The network consists of a UV-guided appearance encoder extracting latent code representing appearance from input images, and a decoder concatenating appearance latent code with human UV maps which are rendered from SMPL parameters that are estimated from HMR model [17]. To make the generated images more realistic and more clear, a discriminator like patchGAN [6] which better deals with high frequency features are also applied. Making assumption that we have dataset contains multiple images of the same person which has the same appearance, we can leverage cross training to compel the appearance latent code independent from human motion structure. This detail is illustrated in Section3.3. In second step, we fine-tune the network with images of a specific person so as to extract a global appearance during cross training. As we have a pretrained network which is generalized well in poses, we just need to collect a few set of images of one person that contain all views. We show that we need about 500 images, or a video with 500 frames of a specific person with all views to achieve the accurate results in Section 4. At the test stage, we can provide a new pose or new shape parameters and realize the person images synthesis for dif-

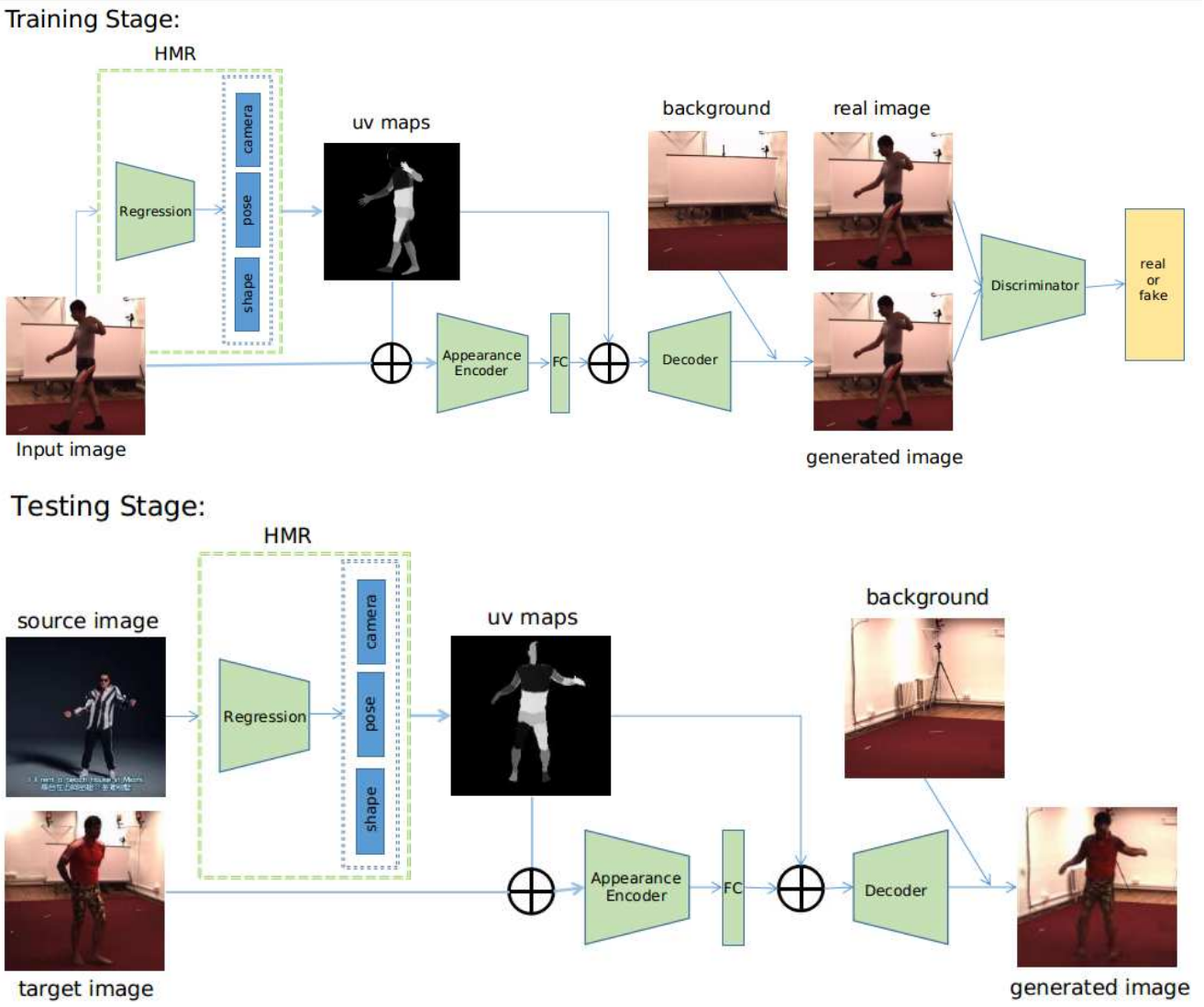


Figure 1. Network Architecture. $\bigoplus$ represents concatenate operation. The top image is the training process. The bottom image is the testing process. At test stage, we vary the camera, pose, shape parameters to synthesize images for different applications.

ferent applications with the appearance saved in the second step.

3.2. Human Structure Representation

In this paper we use SMPL [24] body model for human structure representation since it has high realism and low parameter space. SMPL parameterizes a triangulated mesh with N = 6890 vertices with pose parameters $\theta \in R^{72}$ and shape parameters $\beta \in R^{10}$. Shape $B_s(\beta)$ and pose dependent deformations $B_p(\theta)$ are first applied to a base template T_μ; then the mesh is posed by rotating each body part around skeleton joints $J(\beta)$ using a skinning function W:

$$M(\beta,\theta) = M(T(\beta,\theta), J(\beta), \theta, W), \quad (1)$$

$$T(\beta,\theta) = T_\mu + B_s(\beta) + B_p(\theta), \quad (2)$$

where $M(\beta,\theta)$ is the SMPL function, and $T(\beta,\theta)$ outputs an intermediate mesh in a T-pose after pose and shape deformations are applied. SMPL produces realistic results using relatively simple mathematical operations which are fully differentiable with respect to pose and shape. We leverage these operations, including the ones that determine the projected points of a parameterized 3D body to reconstruct the human body, to be a part of our network. To better provide a proper representation for the network, We leverage UV maps which are a surface-to-image representation and are often used to render textures to represent human structure information. In detail, We divide the SMPL template into 20 parts and use Multi-Dimensional Scaling [37] to unfold surface of each part. We place all part's surface on the common 2D space which is repre-

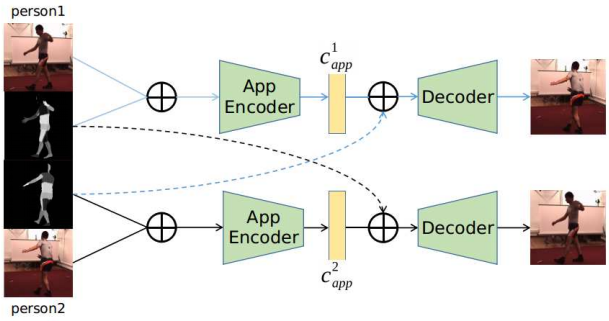


Figure 2. The appearance encoder utilizes UV maps as a guidance and we exchange UV maps between two different images to enhance the separation of appearance and pose.

sented as UV coordinates. As we assign the UV coordinate on each vertex of the template body model, we render the UV coordinate of 3D body model which we call them UV maps. we use DIRT[12] as render tool. This process is similar to that mentioned by DensePose[10]. Compared to some traditional representation for 3D body model, like point cloud, voxel, parameters and so on, UV maps avoid costly computation for 3D model and the neighborhood information about structure in UV maps can be better captured with convolution layers.

To achieve an semi-supervised training, we use a pretrained HMR [17] model to estimate SMPL pose parameters and shape parameters, rather than using groundtruth 3D body model, which may be more accurate but will cause higher computation and introduce more restrictions for input data.

3.3. UV-guided Appearance Encoder

Figure 2. demonstrates our UV-guided appearance encoder and the cross training process. The encoder E_{app} generates a latent code c_{app} representing appearance information with a person image and its corresponding UV maps as inputs. The mapping function can be defined as:

$$c^{app} = E_{app}(x_i, uv_i; \theta) \tag{3}$$

where x_i is the i-th input image, and uv_i is the corresponding i-th UV maps with 2 channels respectively representing u map and v map and has the same resolution with x_i. We utilize a ResNet50 [11] network to learn E_{app} and θ is the learned parameters. In the encoder, as we supply corresponding UV maps, the learned latent appearance code will be enhanced to consider the relationships between the UV maps and the image. When supplying another UV maps, the network will implicitly align with a proper appearance. [29] also concatenates the input image with its corresponding UV maps, but they explicitly operate a spatial transformer network for texture mapping, which is a complicated and costly process. To obtain more information about the whole person, they also add an inpainting autoencoder that needs multi-view images for supervision, while we don't need multi-view data for inpainting as we have multi frames for fine-tuning. To enforce a latent code independent from human pose, we also adopt a cross training. We train two different frames x_i and x_j about the same person at the same time. The two images separately pass by HMR model and generate uv_i and uv_j, then c_i^{app} and c_j^{app} are generated by the appearance encoder. The encoding process is formulated as:

$$\begin{aligned} uv_i &= E_{HMR}(x_i), \qquad uv_j = E_{HMR}(x_j) \\ c_i^{app} &= E_{app}(x_i, uv_i), \qquad c_j^{app} = E_{app}(x_j, uv_j) \end{aligned} \tag{4}$$

Then we exchange uv and c^{app} for two images and utilize a decoder to resynthesize the final images $\hat{x_i}$ and $\hat{x_j}$. As the person's appearance does not vary across the two images, and the differences between the two images are just caused by human pose information, the c^{app} tends to learn the common features that represent appearance information after exchanging. After obtaining appearance latent code, we apply fully connected layer and reshape it to the same size with the UV maps, and then concatenate the appearance latent code and UV maps in channels as the decoder's input. Simultaneously, to exclude background from appearance latent code, we supply the background for each input image. The decoder outputs the foreground and a mask in binary value, then we add the background to foreground according the mask to get the final image. The final generated images x_i and x_j are described as below:

$$\begin{aligned} fg_i, mask_i &= D(uv_i, c_j^{app}) \\ \hat{x_i} &= (1 - mask_i) * bg_i + mask_i * fg_i \\ fg_j, mask_j &= D(uv_j, c_i^{app}) \\ \hat{x_j} &= (1 - mask_j) * bg_j + mask_j * fg_j \end{aligned} \tag{5}$$

3.4. Global Appearance Enroll

To achieve an accurate person images synthesis, we adopt fine-tune methods to enroll a global appearance into the network. After the first training stage where we use data with abundant poses, the network can generalize well in various poses. Then we fine-tune the network with a few of images of a specific person containing complete views. During training, the network parameters will try to fit all the appearance appeared in the training dataset, thus the appearance encoder will capture a global appearance about the specific person. In the test stage, we input a single view image and even if synthesising an unseen view we can still obtain a correct appearance because of our enrolled global appearance. As pretrained network contains pose information, we can realize a short-time training and can still generalize well for various human pose and shape synthesis.

3.5. Loss Functions

To obtain a clear and realistic image, we use a combination of three loss functions, a L1 reconstruction loss called L_{L1}, an adversarial loss called L_{adv}, a perceptual loss called $L_{perceptual}$.

L1 Reconstruction Loss. We compute pixel-wise L1 distance between the generated image and the groundtruth image. It guides the network to capture the low frequency details of images. As we find during experiments, compared to L2 loss, L1 loss provides faster convergence and better stability. It is defined as:

$$L_{L1} = ||\hat{x} - x||_1 \tag{6}$$

Adversarial Loss. GAN loss are added to capture the high frequency details of images, to deal with images blurry. It is defined as:

$$L_{adv} = \mathbb{E}_x[logD(x)] + \mathbb{E}_{x,uv}[log(1 - D(G(x, uv)))] \tag{7}$$

Perceptual Loss. We also apply perceptual loss, as the L2 distance of multi-scale features extracted by a pretrained CNN about two images can encourage image structure similarity, which is illustrated in [16]. We adopt a ResNet18 [11] pretrained on ImageNet [7] to extract the multi-scale features φ_i of totally n scales. It is defined as:

$$L_{perceptual} = \sum_{i=1}^{n} ||\varphi_i(\hat{x}) - \varphi_i(x)||_2^2 \tag{8}$$

Joint Loss. The three loss functions are governed by a coefficient λ and the joint loss is defined as:

$$L = \lambda_1 L_{L1} + \lambda_2 L_{adv} + \lambda_3 L_{perceptual} \tag{9}$$

4. Experiments

4.1. Datasets and Training Details

We pretrain our network on Human 3.6M dataset [14], which has 5 subjects for training and 2 subjects for testing. The subjects perform typical activities such as smoking, sitting and totally have 15 activities. During our pretrained stage, to achieve the generalization in poses, we use all 15 activities for 5 subjects. We leverage 80 per cent of data as training set and the rest 20 per cent as validation set to determine the most suitable hyper parameters. When we finetune the network to enroll the person appearance, we only use one activity that contain all views, as one activity has already contain enough frames for enrollment. This finetune stage use the same hyper parameter pretrained. Details about choosing number of frames are showed in subsection 4.6. Here we use the Photoing activity to fine-tune. We use the remaining 14 activities for testing. The images in this dataset are cropped according to their bounding box. We also collect a few of images of different people by mobile phone for fine-tune. Each person has a video about 50 seconds and totally about 1500 frames. Note that we can not use DeepFashion [23] and Market1501 [41] datasets for human pose tranfer, as we need a set of images of a specific person to enroll the person appearance into the network for more robust applications, while these datasets just have a single image for one person.

In our experiments, all images are resized to 256×256 as input. We load pretrained weight from HMR [17] model and freeze the weight parameters. We adopt a batch size of 32 and use the Adam optimizer [18] with parameters β_1 = 0.5 and β_2 = 0.9, also with a learning rate of 1e-4. The loss weights are set to $\lambda_1 = 80$ for reconstruction loss, $\lambda_2 = 1$ for adversarial loss and $\lambda_3 = 150$ for perceptual loss after validation.

4.2. Human Pose Transfer

When we fix shape and camera parameters, we show the qualitative results of human pose transfer in Figure 3 which use self-collected dataset. As the network enrolls the person appearance in, so the pose in unseen views also generate images with true appearance.

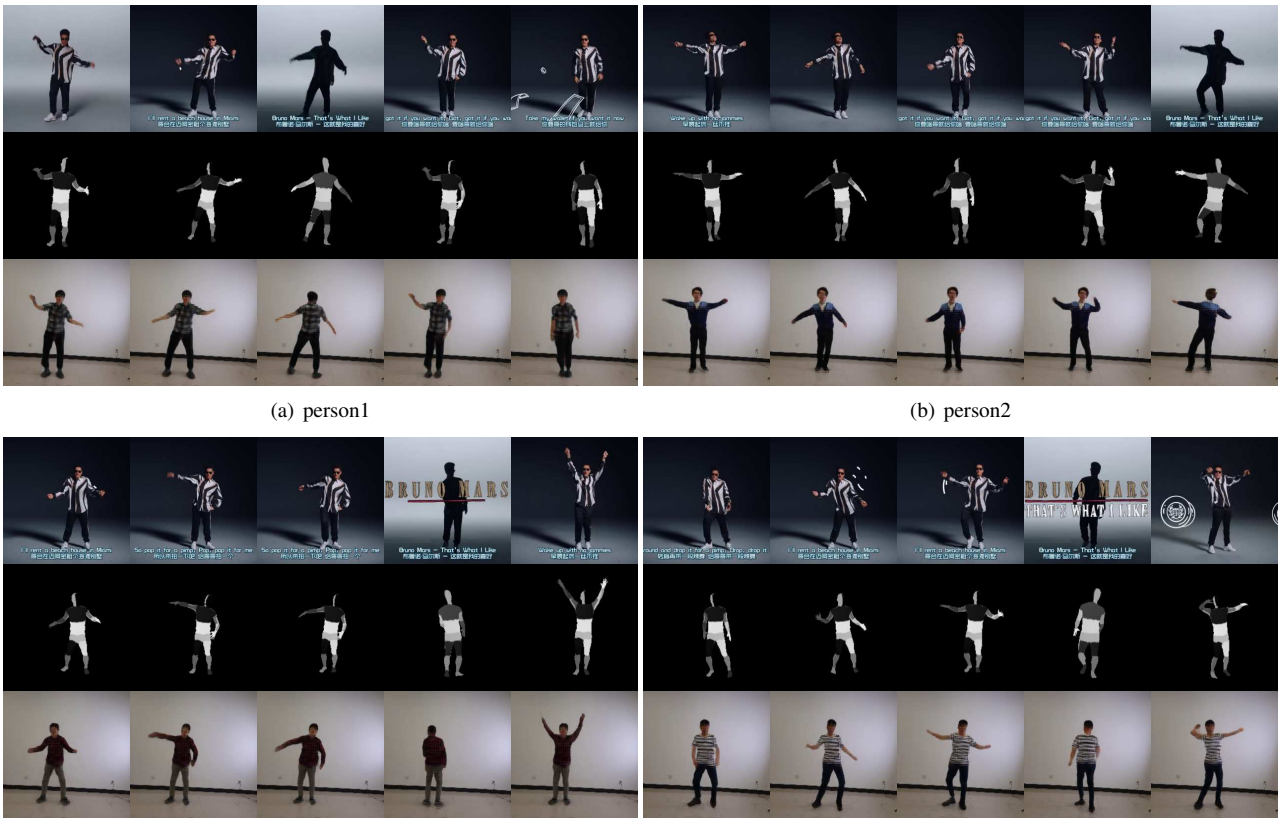
Figure 4 shows the results in Human3.6M test set with our method and Dance [3] which is most similar to us. Additionally, Table 1 calculates Inception Scores(IS) [34] and Structural Similarities(SSIM) [36]. As our method completely separates human appearance, pose and shape, and the UV-guided encoder better correspond the uv maps with the input image for extracting appearance, we can achieve a better result quantitatively and qualitatively. Our pretrained network which generalized well on the poses also brings an improvement in the end of arms and legs, while [3] is more blurry at these positions. Note that in [3] GAN for face is additionally trained, but our work in this paper is focus on human body shape and pose.

Method	SSIM	IS	
		mean	std
Real Image	1.000	2.360	0.033
Dance [3]	0.310	2.017	0.100
Ours	**0.334**	**2.232**	0.109

Table 1. Inception scores(IS) and structured similarities(SSIM) comparison of reconstructed test images on Human3.6M dataset.

4.3. Human Shape Change

We change various body shapes as shown in Figure 5. We vary the weight and the height respectively for the test



(a) person1 (b) person2

(c) person3 (d) person4

Figure 3. Human pose transfer results in self-collected datasets. The source poses come from Mars. In each subfigure, the first row is the source person's poses, the second row is the corresponding UV maps with poses, and the third row is the corresponding generated images with target person's appearance. We show results on four target people.

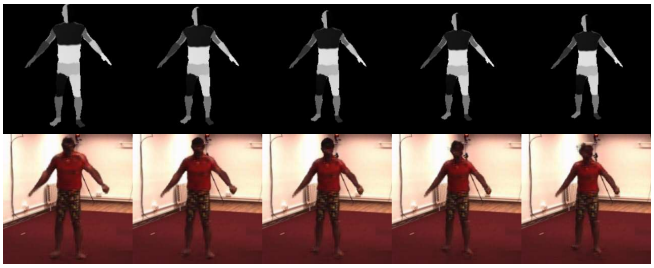
Figure 4. Qualitative comparison between our method and previous work [3].

set in Human3.6M. As we completely separate the components of human images, we can achieve an excellent results with different body shapes. In previous work, when we want to get different human shape, the most common method is to adjust 3D human model which are complex and costly, and then project to 2D images. Our method for changing human shape is easier to realize.

For those applications that need to keep the target person's shape, we can get the shape parameters from HMR model and change the source person's shape parameters to

the target's shape.

(a) Change Weight



(b) Change Height

Figure 5. Human shape change results in Human3.6M dataset. The first row is the UV maps, and the second row is the corresponding generated images with different shapes.

4.4. Novel Views Synthesis

When we fix shape and pose parameters, we can rotate the camera parameters from 0° to 360° and synthesis images in arbitrary views. The qualitative results on Human3.6M dataset are shown in Figure 6. The input is a single image when testing, but as the fine-tune process targeting at a specific person enrolls the appearance into the network, if even turn to an unseen view, we can still synthesis an accurate image with right appearance. Note that for a better visualization about the separation of the poses, shapes and camera parameters, we control the variable parameters for changing only one. In fact, we can vary the poses, shapes and camera parameters in the same time to get more various images.

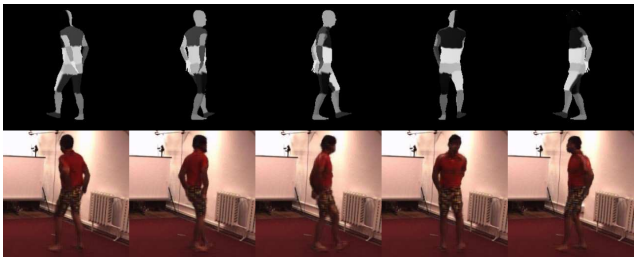


Figure 6. Novel views synthesis results in Human3.6M dataset. The first row is the UV maps, and the second row is the corresponding generated images in arbitrary views.

4.5. Ablation Study

Figure 7 shows the qualitative results and Table 2 shows the quantitative results of ablation study, which demonstrate the effect of individual components of our network. With the same UV maps showed in the first row, the second row is the result without UV guided, that means we do not concatenate UV maps with the target image. When no UV maps guide the input image to extract the appearance information for a person, the appearance may mix something that not belongs to the human structure. The second row is the result without cross-training strategy. In this case the human body structure and appearance may not separate well, and when we input a different pose, the generated images may be a little blurry and lost end joints to compromise the conflict. In contrast, our UV-guided appearance encoder and the cross-training strategy can lead to a clear and realistic result.

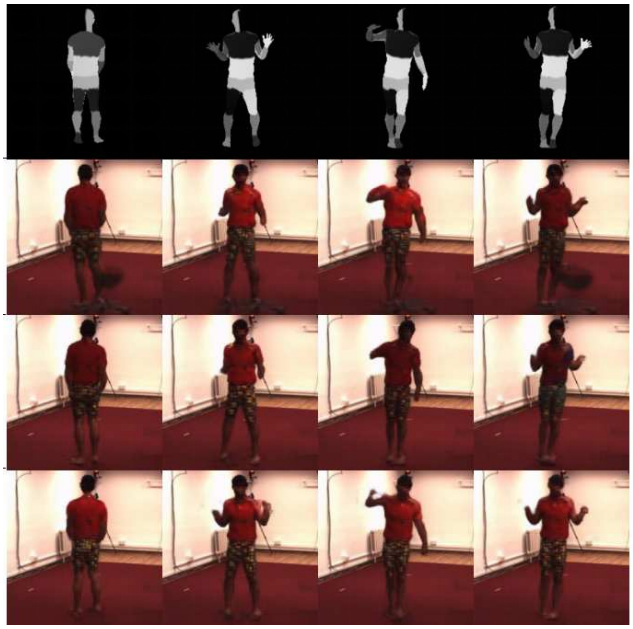


Figure 7. Qualitative comparison of ablation study on Human3.6M dataset. From top to bottom, each row shows the results of uv maps, our method without UV-guided encoder and without cross-training, our method without cross-training, and our full method.

Method	SSIM	IS	
		mean	std
Real Image	1.000	2.360	0.033
w/o uv& w/o cross	0.311	2.082	0.056
w/o cross	0.313	2.011	0.108
full	**0.334**	**2.232**	0.109

Table 2. Quantitative comparison of ablation study on Human3.6M dataset. Each row is the corresponding inception scores(IS) and structured similarities(SSIM) results of different methods.

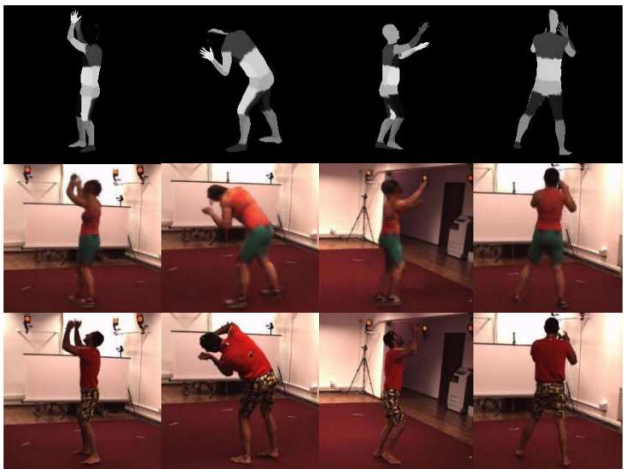


Figure 8. Results with no appearance enrollment. The first row is the supplied uv maps, the second row is the generated images, and the third row is the corresponding ground-truth images.

4.6. Enrollment Amounts Study

To determine the proper number of images for accurate appearance enrollment, we compare the results with different amounts of images for fine-tuning. Figure 8 shows the results without fine-tuning. While the training set contains only 5 subjects, the network may be overfitting. Thus, when we test with a new person, the generated images keep the most similar appearance of the training set rather than test person's. When we add appearance enrollment, while amount of images for fine-tuning is increasing, the results of the generated images is better. But when it turns to about 500 or more, quality of generated images show no obvious promotion, as shown in Figure 9. Figure 10 shows qualitative results for different amounts of images, respectively containing 5, 100 and 500 images. When the images are too less, the network may be overfitting in poses, thus may cause blurry in generated images. When images become more, the generated images will be more clear. We also use only front images to fine-tune the network. Figure 11 shows the results. As the front appearance is seen by the network, the result is clear. But when we test with back images, the network hasn't enrolled the back appearance, causing the generated images blurry.

5. Conclusion

In this paper we propose a semi-supervised framework for synthesising person images, that can generate person images in arbitrary poses, shapes and views for different applications. We also adopt a UV-guided appearance encoder and a cross-training strategy for a better separation, and fine-tune is adopted to enroll the appearance into the network. Results on Human3.6M dataset and self-collected datasets show that the framework can be utilized in many applications and can achieve an excellent performance. In the future, we will further development face generation for a better visualization and improve our method to generate clearer images, thus the easy semi-supervised person images synthesis method can be applied to more fields.

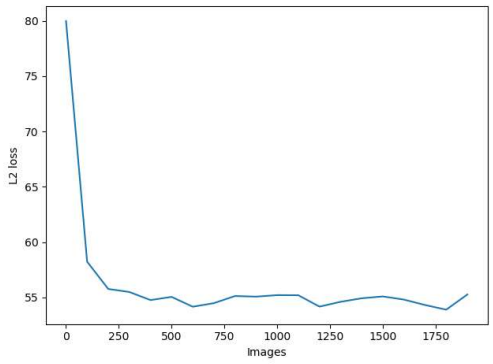


Figure 9. The relationship between the number of images and L2 loss. L2 loss is computed between generated images and corresponding ground-truth.

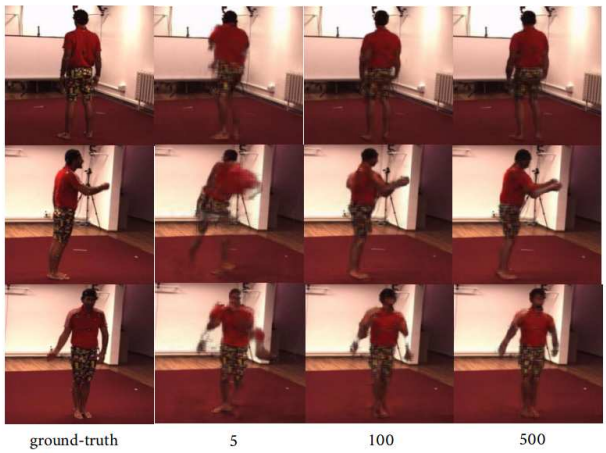


Figure 10. Qualitative results for different number of images used to fine-tune.

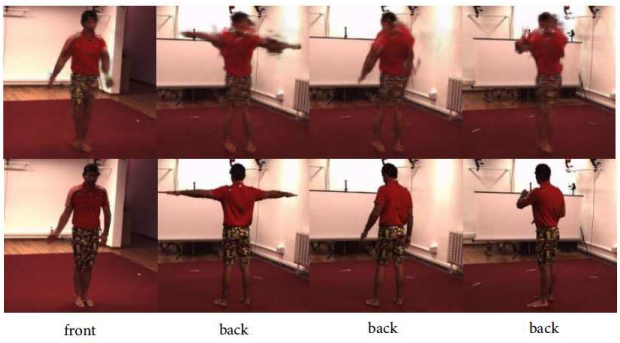


Figure 11. First row is generated images trained only with only front view, and the second row is ground-truth images.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

[2] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John V. Guttag. Synthesizing images of humans in unseen poses. *CoRR*, abs/1804.07739, 2018. URL http://arxiv.org/abs/1804.07739.

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. *CoRR*, abs/1808.07371, 2018. URL http://arxiv.org/abs/1808.07371.

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, pages 1520–1529. IEEE Computer Society, 2017.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2172–2180. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6399-infogan-interpretable-representation-learning-by-information-maximizing-generative-adversarial-pdf.

[6] Ugur Demir and Gözde B. Ünal. Patch-based image inpainting with generative adversarial networks. *CoRR*, abs/1803.07422, 2018. URL http://arxiv.org/abs/1803.07422.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (short Oral)*, 2018. URL https://compvis.github.io/vunet/.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[10] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Paul Henderson and Vittorio Ferrari. Learning to generate and reconstruct 3d meshes with only 2d supervision. In *British Machine Vision Conference (BMVC)*, 2018.

[13] Berthold Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 08 1981. doi: 10.1016/0004-3702(81)90024-2.

[14] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016. URL http://arxiv.org/abs/1611.07004.

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

[17] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018.

[18] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL http://arxiv.org/abs/1312.6114. cite arxiv:1312.6114.

[20] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model of people in clothing. *CoRR*, abs/1705.04098, 2017. URL http://arxiv.org/abs/1705.04098.

[21] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[22] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *CoRR*, abs/1606.07536, 2016. URL http://arxiv.org/abs/1606.07536.

[23] 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)*, 6 2016.

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–248:16, 2015.

[25] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *CoRR*, abs/1705.09368, 2017. URL http://arxiv.org/abs/1705.09368.

[26] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] Bruno Mars. That's what i like official video. https://www.youtube.com/watch?v=PMivT7MJ41M.

[28] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[29] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[30] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer. Unsupervised Person Image Synthesis in Arbitrary Poses. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[31] Amit Raj, Patsorn Sangkloy, Huiwen Chang, Jingwan Lu, Duygu Ceylan, and James Hays. Swapnet: Garment transfer in single view images. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[32] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/reed16.html.

[33] Jose C. Rubio, Angela Eigenstetter, and Björn Ommer. Generative regularization with latent topics for discriminative object recognition. *Pattern Recogn.*, 48(12):3871–3880, December 2015. ISSN 0031-3203. doi: 10.1016/j.patcog.2015.06.013. URL http://dx.doi.org/10.1016/j.patcog.2015.06.013.

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016.

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[36] 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.

[37] Wikipedia contributors. Multidimensional scaling — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Multidimensional_scaling&oldid=887960174, 2019. [Online; accessed 15-April-2019].

[38] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1612.03242, 2016. URL http://arxiv.org/abs/1612.03242.

[40] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 383–391, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5665-7. doi: 10.1145/3240508.3240536. URL http://doi.acm.org/10.1145/3240508.3240536.

[41] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015.

[42] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A. Efros. View synthesis by appearance flow. In *ECCV*, 2016.

[43] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

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