Multistage Adversarial Losses for Pose-Based Human Image Synthesis

Chenyang Si[1,3] Wei Wang[1,3,*] Liang Wang[1,2,3] Tieniu Tan[1,2,3]
[1]Center for Research on Intelligent Perception and Computing (CRIPAC),
National Laboratory of Pattern Recognition (NLPR)
[2]Center for Excellence in Brain Science and Intelligence Technology (CEBSIT),
Institute of Automation, Chinese Academy of Sciences (CASIA)
[3]University of Chinese Academy of Sciences (UCAS)
chenyang.si@cripac.ia.ac.cn, {wangwei, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

Human image synthesis has extensive practical applications e.g. *person re-identification and data augmentation for human pose estimation. However, it is much more challenging than rigid object synthesis,* e.g. *cars and chairs, due to the variability of human posture. In this paper, we propose a pose-based human image synthesis method which can keep the human posture unchanged in novel viewpoints. Furthermore, we adopt multistage adversarial losses separately for the foreground and background generation, which fully exploits the multi-modal characteristics of generative loss to generate more realistic looking images. We perform extensive experiments on the Human3.6M dataset and verify the effectiveness of each stage of our method. The generated human images not only keep the same pose as the input image, but also have clear detailed foreground and background. The quantitative comparison results illustrate that our approach achieves much better results than several state-of-the-art methods.*

1. Introduction

Synthesizing a novel view image from a single image is an important and challenging problem in computer vision. Particularly, human view synthesis which plays important roles in human understanding has extensive practical applications. For example, human view synthesis can effectively solve the cross-view problems, *e.g.* cross-view action recognition [26, 7, 12] and cross-view person re-identification [28, 2]. Multiview human synthesis can be used as a means of data augmentation for human pose estimation in [22].

Novel view synthesis is very challenging due to the need of inferring unseen content from a single image. Geometry-based methods [1, 8, 31] generate novel view images by reconstructing 3D object, while transformation-based methods [19, 27, 32] directly learn to model the transformation between different perspectives of object. Recently, there occurred a lot of work on image generation and synthesis by means of variational autoencoders (VAE) [10] and generative adversarial networks (GANs) [3], which have shown impressive results [21, 20, 33, 11].

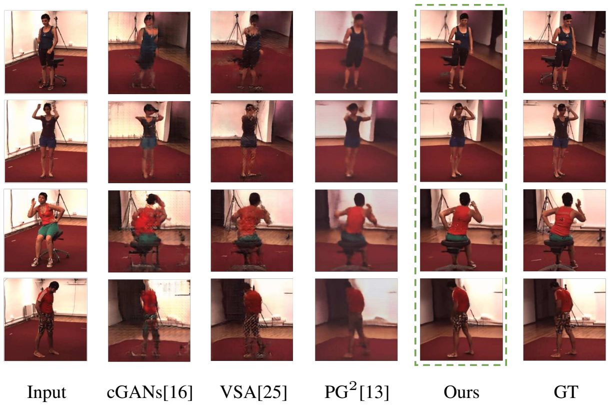

Figure 1. Human image synthesis by our proposed method and three state-of-the-art methods: cGANs [16], VSA [25], PG^2[13]. Our generated images in the green dashed box keep the same pose with the groundtruth images, while the other methods can not always obtain correct poses. Moreover, our method generates much better foreground and background images than the others.

Generally, the above methods are used to synthesize rigid objects, *e.g.* faces, cars and chairs, which have the characteristics of shape invariance and symmetry. Generating human images is much more challenging than synthesizing rigid objects, due to the variability of human posture. Taking [29] for example, although this work can generate multi-view human images from a single image, it can not keep the human posture unchanged.

In this paper, we propose a pose-based human image

[*] Corresponding Author: Wei Wang

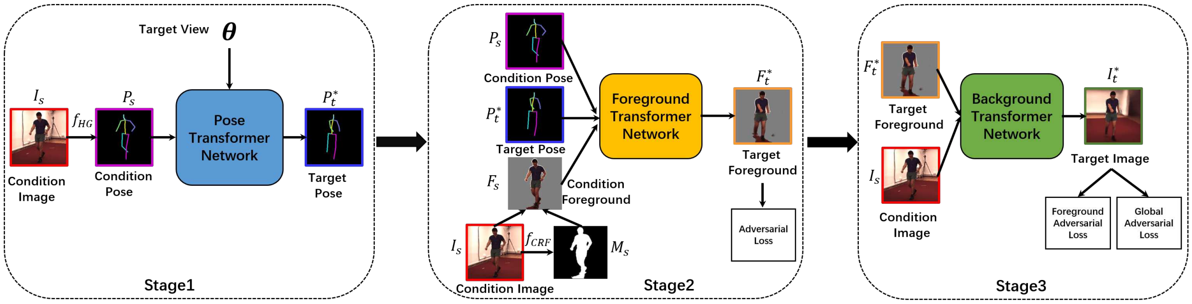

Figure 2. The overall pipeline of our multistage approach which contains three transformer networks for three stages. In the first stage, the pose transformer network synthesizes a novel view 2D pose. Then, the foreground transformer network synthesizes the target foreground image in the second stage. Finally, the background transformer network generates the target image. f_{HG} and f_{CRF} donate the stacked hourglass networks [18] and the CRF-RNN [30] for pose estimation from image and foreground segmentation, respectively.

synthesis method (shown in Fig. 2), which keeps human posture unchanged during generating novel view images from a single image. Fig. 2 shows the procedure of human image synthesis, which contains three transformer networks for three stages. (1) In the first stage, we propose a pose transformer network which can synthesize 2D target pose P_t^* of other perspectives from the condition pose P_s corresponding to the condition image I_s. Instead of focusing on 2D pose estimation from image in this work, we adopt the stacked hourglass networks [18] to estimate the condition pose. With the low-dimensional characteristics of human pose, the pose transformation between different perspectives can be easily learned. (2) In the second stage, we extract human foreground F_s from condition image with the segmentation method CRF-RNN [30], and then propose a foreground transformer network to synthesize the target human foreground F_t^* with the 2D poses $\{P_s, P_t^*\}$ and condition human foreground F_s. Here a adversarial loss is used to improve the image quality. 3) In the third stage, a background transformer network is proposed to generate target full image I_t^* with the condition image I_s and the generated foreground image F_t^* as the input. Two adversarial losses separately for the foreground image and the full image, *e.g.* foreground adversarial loss and global adversarial loss, are imposed to generate clear and detailed images.

Fig. 1 shows the comparison results between our method and three state-of-the-art approaches [16, 25, 13]. We can see that our generated images have the same pose with the groundtruth images, while the results of other methods can not always obtain correct poses. Moreover, our images show much better foreground and background images than the other methods.

The main contributions of this paper are summarized as follows:

- We propose a pose-based human image synthesis method which can keep the human posture unchanged in novel viewpoints, which is very difficult for current state-of-the-art methods.
- We propose a multistage adversarial loss approach which generates high-quality foreground and background images in novel viewpoints.
- Our method synthesizes much better novel view human images than several state-of-the-art methods on the Human3.6M dataset.

2. Related Work

In this section, we briefly review the existing literature that closely relates to the proposed method.

View synthesis There have been amounts of work proposed for novel view synthesis, which can be categorized into two broad classes: geometry-based methods and direct transformation methods. Geometry-based methods [1, 8, 31] generate novel view images by reconstructing 3D object, while transformation-based methods [19, 27, 32] directly learn to model the transformation between different perspectives. Ji *et al.* [6] use convolutional neural networks to synthesize a middle view of two images. Tatarchenko *et al.* [24] propose a network to infer 3D representation from an arbitrary viewpoint. Park *et al.* [19] first predict the parts of the geometry visible in both the input and novel views, and then generate a novel image. Zhou *et al.* [32] learn appearance flows for synthesizing novel views. These methods are generally designed to synthesize rigid objects, which do not work well for human image synthesis.

Pose-based human image synthesis To keep human posture correct in image generation and synthesis, modelling human pose is a very natural choice. Villegas *et al.* [25] predict future pose sequence for generating long-term future frames. This method works very well for successive frames due to their small changes, but fails to generate human images of large viewpoint changes. The most similar work to ours is [13] which proposes a pose guided person generation network (PG^2) to synthesize person images in a coarse-to-fine way. It can be seen that our work aims at multiview human synthesis which needs to generate poses in other viewpoints, while [13] uses predetermined pose. Our proposed multistage adversarial losses separately for the foreground and the background achieve much better results than the coarse-to-fine method in [13]. Please refer to Fig. 6 and Table 1 for the comparison results.

Adversarial loss for image generation Adversarial loss is used widely in image generation due to its multi-modal characteristics, which overcomes the average prediction problem caused by mean square error. Mathieu *et al.* [15] use generative adversarial training to predict the sharp frames. Mirza *et al.* [16] propose conditional GANs (cGANs) for image-to-image translation tasks. In this paper, we adopt multistage adversarial losses to optimize the procedure of image generation.

3. Model Architecture

Human image synthesis is very challenging due to the variability of human posture. In this paper, we propose a pose-based human image synthesis method which contains three transformer networks for three stages. Fig. 2 illustrates these three networks: pose transformer network, foreground transformer network and background transformer network. In this section, we will introduce these networks in detail.

3.1. Pose Transformer Network

Inspired by the deep feedforward network proposed for inferring 3D joints from 2D ground truth in [14], we propose a pose transformer network (see Fig. 3). Given 2D pose joints and a rotation angle θ, our goal is to estimate 2D pose of other perspective. The function of the pose transformer network G_p is defined as follows:

$$P_t^* = G_p(P_s, \theta) \tag{1}$$

where P_t^* is the predicted target pose, and P_s is the condition 2D pose of the input image. The pose transformer network G_p has seven linear layers and an embedding layer. The first linear layer encodes the input pose joints into a 1024-dim vector and the embedding layer transforms the rotation angle θ into a 512-dim vector. These two vectors will be concatenated as the input to another two residual blocks.

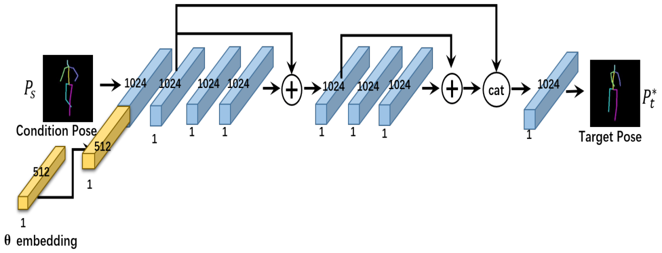

Figure 3. The architecture of the pose transformer network.

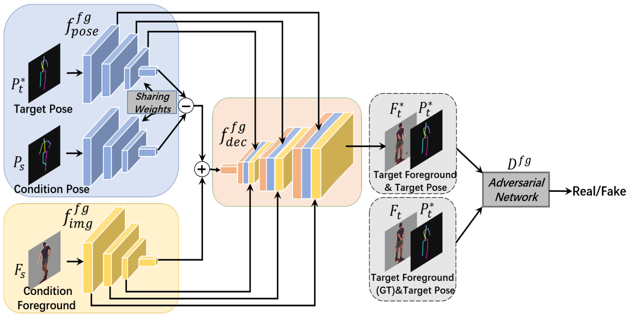

Figure 4. The architecture of the foreground transformer network.

The last layer predicts the target pose. In G_p, all the linear layers have 1024 hidden nodes, which are followed by batch normalization [4] and Rectified Linear Units (RELUs) [17] except the last linear layer. It should be noted that we do not focus on pose estimation from image in this work, and directly use the stacked hourglass networks (f_{HG}) [18] to estimate the condition pose P_s from the image I_s.

We train the pose transformer network with the ℓ_2 regression loss:

$$\mathcal{L}^1 = \sum_i^N \left\| {P_t^*}^i - {P_t}^i \right\|_2^2 \tag{2}$$

where ${P_t}^i$ is the groundtruth of joint i and ${P_t^*}^i$ is the predicted location of joint i. After estimating the target pose, we start to synthesize the human image.

3.2. Foreground Transformer Network

Given the predicted target pose, we need to synthesize the corresponding human image that has the same appearance with the input image. Inspired by [25] and [23], we propose a foreground transformer network for human foreground generation. It comprises of an image encoder f_{img}^{fg}, a pose encoder f_{pose}^{fg} , an image decoder f_{dec}^{fg} and a discriminator D^{fg}, which is shown in Fig. 4. The network synthesizes the target human foreground F_t^* by inferring the transformation from the condition pose P_s to the target pose P_t^* and transferring the condition foreground image F_s to target foreground based on this pose transformation:

$$F_t^* = f_{dec}^{fg}(f_{pose}^{fg}(P_t^*) - f_{pose}^{fg}(P_s) + f_{img}^{fg}(F_s)) \tag{3}$$

where F_s is the segmented condition human foreground image which is predicted by CRF-RNN [30].

The foreground transformer network can generate multi-view foreground images that have the same appearance and posture with the input image. In this network , the condition pose P_s and target pose P_t^* are encoded into the pose features by the encoder f_{pose}^{fg}. A subtraction operation between the target pose feature and the condition pose feature is used to model the pose transformation. The image encoder f_{img}^{fg}

extracts the appearance feature. Finally, the sum of image feature and the pose transformation feature is decoded to the target foreground F_t^* by the image decoder f_{dec}^{fg}. The skip connections between the pose encoder f_{pose}^{fg} and the image decoder f_{dec}^{fg} propagate the pose features, which can ensure the synthesized human image to have the same pose with the condition input image. It should be mentioned that although the target pose and the condition pose are encoded by the same encoder f_{pose}^{fg}, there are no skip connections when encoding the condition pose (see Fig. 4). We also add skip connections between image encoder and image decoder to make sure that the generated foreground has the same appearance with the input image.

In foreground transformer network, we use two-dimensional skeleton image as the input to the pose encoder instead of 2D pose joints. Particularly, we assign different values to different parts in this skeleton image, *e.g.* the value of the left leg is (0, 255, 255) and the right leg is (255, 0, 255), which can represent the high-level structure of the human pose well. All the encoders and decoder in this network adopt convolutional networks. Due to the multi-modal characteristics of generative adversarial networks [3], we use the adversarial loss to improve the quality of generated foreground images. The training loss of this network is defined as follows:

$$\mathcal{L}^2 = \alpha_f \mathcal{L}^2_{fg} + \beta_f \mathcal{L}^2_{bg} + \mathcal{L}^2_{gen} \tag{4}$$

where $\mathcal{L}^2_{fg}$ and $\mathcal{L}^2_{bg}$ are the ℓ_1 loss for the foreground and the background of the synthesized image F_t^*, and $\mathcal{L}^2_{gen}$ is the term in the adversarial loss that makes the model to generate real images. α_f, β_f are the weighting coefficients. In our experiments, β_f is smaller than α_f due to the small changes in the background.

The $\mathcal{L}^2_{fg}$and $\mathcal{L}^2_{bg}$ are defined as follows:

$$\begin{aligned}\mathcal{L}^2_{fg} &= \|F_t \odot M_t - F_t^* \odot M_t\|_1 \\ &= \frac{1}{\sum\limits_{M_t^{i,j}=1} M_t^{i,j}} \sum_{i,j} \left|(F_t{}^{i,j} - F_t^*{}^{i,j}) \times M_t{}^{i,j}\right|\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}^2_{bg} &= \|F_t \odot (1 - M_t) - F_t^* \odot (1 - M_t)\|_1 \\ &= \frac{1}{\sum\limits_{M_t^{i,j}=0} (1 - M_t^{i,j})} \sum_{i,j} \left|(F_t{}^{i,j} - F_t^*{}^{i,j}) \times (1 - M_t{}^{i,j})\right|\end{aligned} \tag{6}$$

where F_t is the groundtruth of the target foreground image, and M_t is the foreground mask which is predicted from the groundtruth image I_t by CRF-RNN [30]. The mask M_t is only needed during training. Here $\odot$ denotes element multiplication.

The adversarial term $\mathcal{L}^2_{gen}$ is defined by:

$$\mathcal{L}^2_{gen} = -\log(D^{fg}([F_t^*, P_t^*])) \tag{7}$$

where F_t^*, P_t^* are the predicted foreground image and 2D target pose, and $D^{fg}(.)$ is the discriminator network in adversarial loss. The discriminator loss is defined as follows:

$$\begin{aligned}\mathcal{L}^2_D = &-\log(D^{fg}([F_t, P_t^*])) \\ &-\log(1 - D^{fg}([F_t^*, P_t^*]))\end{aligned} \tag{8}$$

3.3. Background Transformer Network

We have synthesized the foreground image, and there is no clear background for the synthesized image. In this section, we propose a background transformer network to generate the target image with clear background.

The background transformer network is illustrated in Fig. 5, which consists of a foreground encoder f_{fg}^{bg}, a condition image encoder f_{img}^{bg} and a image decoder f_{dec}^{bg}. The input of this network is the condition image I_s and the synthesized foreground image F_t^*. The condition image contains both the background information and the characteristics of human appearance, *e.g.* color, texture, which are extracted by the condition image encoder f_{img}^{bg}. The foreground encoder f_{fg}^{bg} maps the target foreground F_s^* to the target human feature. Then the concatenation of the outputs of the image encoder f_{img}^{bg} and the foreground encoder f_{fg}^{bg} is fed into the image decoder f_{dec}^{bg}. Similar to the foreground transformer network, we utilize the skip connection between the image encoder, the foreground encoder and the image decoder, which can help to recover more details.

Due to the complexity of image background, generating a high quality image becomes more difficult and challenging. So we adopt two adversarial losses to allow our model to generate realistic looking images. They are a foreground adversarial loss and a global adversarial loss, which have a foreground discriminator network D_{fg}^{bg} and a global discriminator network D^{bg}, respectively. The D_{fg}^{bg} takes the foreground of the generated target image I_t^* and the target pose P_t^* as the inputs, which focuses on the foreground generation. The D^{bg} takes the generated image I_t^* as input, which optimizes the quality of the full image.

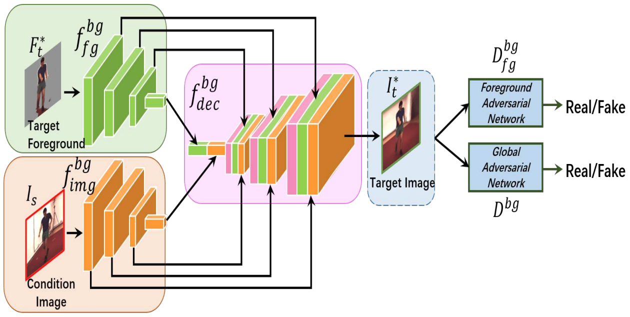

Figure 5. The architecture of the background transformer network.

This network is trained with the following loss function:

$$\mathcal{L}^3 = \alpha_b \mathcal{L}^3_{fg} + \beta_b \mathcal{L}^3_{bg} + \mathcal{L}^3_{gen_{fg}} + \mathcal{L}^3_{gen} \tag{9}$$

where $\mathcal{L}^3_{fg}$ and $\mathcal{L}^3_{bg}$ are the ℓ_1 loss for the foreground and the background of the generated target image I^*_t, and α_b, β_b are the weighting terms. The $\mathcal{L}^3_{fg}$ and $\mathcal{L}^3_{bg}$ have the same formula as the Eqn. 5 and Eqn. 6, except replacing F_t, F^*_t with I_t, I^*_t. The two adversarial terms $\mathcal{L}^3_{gen_{fg}}$ and $\mathcal{L}^3_{gen}$ are defined by:

$$\mathcal{L}^3_{gen_{fg}} = -\log(D^{bg}_{fg}([I^*_t \odot M_t, P^*_t])) \tag{10}$$

$$\mathcal{L}^3_{gen} = -\log(D^{bg}(I^*_t)) \tag{11}$$

and the discriminator losses of D^{bg}_{fg} and D^{bg} are defined as follows:

$$\begin{aligned}\mathcal{L}^3_{D_{fg}} &= -\log(D^{bg}_{fg}([I_t \odot M_t, P^*_t])) \\ &\quad - \log(1 - D^{bg}_{fg}([I^*_t \odot M_t, P^*_t]))\end{aligned} \tag{12}$$

$$\mathcal{L}^3_D = -\log(D^{bg}(I_t)) - \log(1 - D^{bg}(I^*_t)) \tag{13}$$

4. Experiments

4.1. Experimental Settings

To verify the effectiveness of our multistage approach, we perform extensive experiments on the Human3.6M dataset [5]. This dataset is collected from 4 cameras simultaneously, which contains the images and the poses of 11 subjects. Each subject performs 15 kinds of actions. In our experiments, 3 actions of each subject are used as the test dataset and the rest of the data are used to train the model. The size of the input image is set to 224 × 224 × 3. For the pose transformer network, we train 15 epochs with a minibatch of size 500. The initial learning rate is set to 0.001. For foreground transformer network and background transformer network, we train 10 epochs with a minibatch of size 40 and an initial learning rate of 0.0001. We set $\alpha_f = \alpha_b = 100$ and $\beta_f = \beta_b = 50$. All the networks are optimized using the ADAM optimizer [9]. For the baseline models below, the training parameters are same as our proposed model.

We visualize the synthesized images, and compare our model with the other three state-of-the-art methods [16, 25, 13]. In quantitative comparison, we adopt the structural similarity index measure (SSIM) and the peak signal to noise ratio (PSNR) to measure the quality of the generated image. SSIM measures image similarity from three aspects of brightness, contrast and structure. The range of SSIM is [0, 1], and large SSIM value indicates high structural similarity. PSNR is one of the most widely used index of image objective evaluation. The higher the PSNR value,the better the image quality.

Methods	SSIM	PSNR
Mirza *et al.* [16] (cGANs)	0.52	17.05
Villegas *et al.* [25] (VSA)	0.54	17.52
Ma *et al.* [13] (PG^2)	0.60	19.19
Ours	**0.72**	**20.62**

Table 1. The comparison results between our method and the other state-of-the-art methods.

4.2. Experimental Results

We compare our proposed method with the other three state-of-the-art methods, *e.g.* Mirza *et al.* [16] (cGANs), Villegas *et al.* [25] (VSA) and Ma *et al.* [13] (PG^2), and report their SSIM and PSNR values in Table 1. We can see that our proposed multistage approach achieves the best performance with SSIM (0.72) and PSNR (20.62).

We also visualize the generated images of the compared methods and our method in Fig. 6. It can be seen that although VSA [25] and PG^2 [13] are two pose-based image generation methods, they still can not keep the pose correct in novel viewpoint. Particularly, all the compared methods can not recover the arm pose. Due to the large viewpoint changes, they all generate distorted background. Our method achieves the best results in the seventh column, which keeps human pose the same as the input image and generates clear foreground and background images. These results demonstrate the effectiveness of our multistage adversarial losses approach.

4.3. Model Analysis

We analyze the proposed foreground network and background network by comparing them with several variants. The comparison results demonstrate their effectiveness. Moreover, we perform multiview human image synthesis to show the generalization ability of our model. We also explore the role of human pose in novel view synthesis.

4.3.1 Analysis of Foreground Network

There are three key ingredients in the proposed foreground transformer network: skip connection (Unet), pose encoder (Pose) and generative loss (GAN). To analyze the role of each component, we compare it with several combinations of these components, *e.g.* Unet+GAN and Unet+Pose.

Unet This generator only has an U-net architecture, which consists of an image encoder f^{fg}_{img}, an image decoder f^{fg}_{dev} and several skip connections between them.

Unet+GAN It adds a discriminator network D^{fg} to the Unet. This discriminator only takes images as input. The architecture of Unet+GAN is similar to Fig. 4 but without the pose encoder.

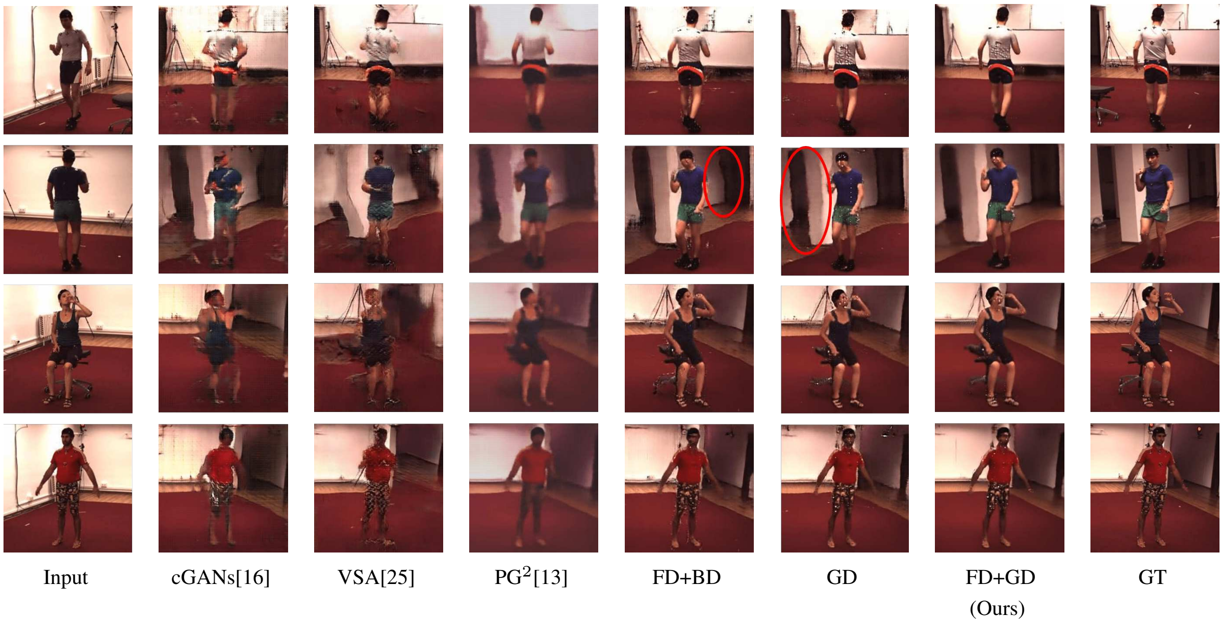

Figure 6. Visualization of the synthesized images from three state-of-the-art methods, two baselines and our model. Our method achieves the best results with clear foreground and background.

Unet+Pose It dose not consist of the discriminator network D^{fg} compared to Fig. 4.

EnDe+Pose+GAN EnDe denotes the encoder-decoder architecture. This model is similar to Fig. 4 but without skip connections between the encoders and the decoder.

Unet+Pose+GAN (Ours) It denotes our proposed foreground transformer network.

We compare our foreground transformer network with the above baseline networks. Table 2 shows the SSIM and PSNR values and Fig. 7 visualizes the generated images. Compared with Unet, Unet+GAN and EnDe+Pose+GAN, our model Unet+Pose+GAN achieves better SSIM and PSNR values, which illustrates the importance of pose encoder and skip connections in foreground transformer network. The generated images of Unet, Unet+GAN and EnDe+Pose+GAN in Fig. 7, which can not keep the pose correct in novel viewpoint and particularly can not recover the arm pose, verify the importance of each component in our model again. It should be noted that although Unet+Pose performs better than our model in terms of PSNR and SSIM (0.82 and 22.57 vs. 0.81 and 22.10), it can not recover clear details in the generated images. We can see that Unet+Pose generally generates blurred foreground and losses details in human back in Fig. 7. It is mainly caused by the missing of adversarial loss. Our foreground transformer network with Unet+Pose+GAN achieves the best visual results compared with the groundtruth in Fig. 7.

Methods	SSIM	PSNR
Foreground		
Unet	0.77	20.49
Unet+GAN	0.73	19.39
Unet+Pose	**0.82**	**22.57**
EnDe+Pose+GAN	0.75	19.81
Unet+Pose+GAN (Ours)	**0.81**	**22.10**
Background		
FD+BD	0.67	19.70
GD	0.65	19.45
FD+GD (Ours)	**0.72**	**20.62**

Table 2. The comparison results between several variants and our transformer networks for the foreground and background.

4.3.2 Analysis of Background Network

In the proposed background transformer network, we adopt two discriminator networks for the foreground and the full image, respectively. To verify the effectiveness of this architecture, we compare it with another two methods.

FD+BD It denotes that we use two discriminator networks for the foreground and the background, respectively. We name them as foreground discriminator (FD) and background discriminator (BD).

GD It only uses a discriminator network for the full image, which is called global discriminator (GD).

FD+GD (Ours) It is our model that is illustrated in Section 3.3.

Table 2 shows the SSIM and PSNR values of the two baselines and our model. The SSIM and PSNR of GD are

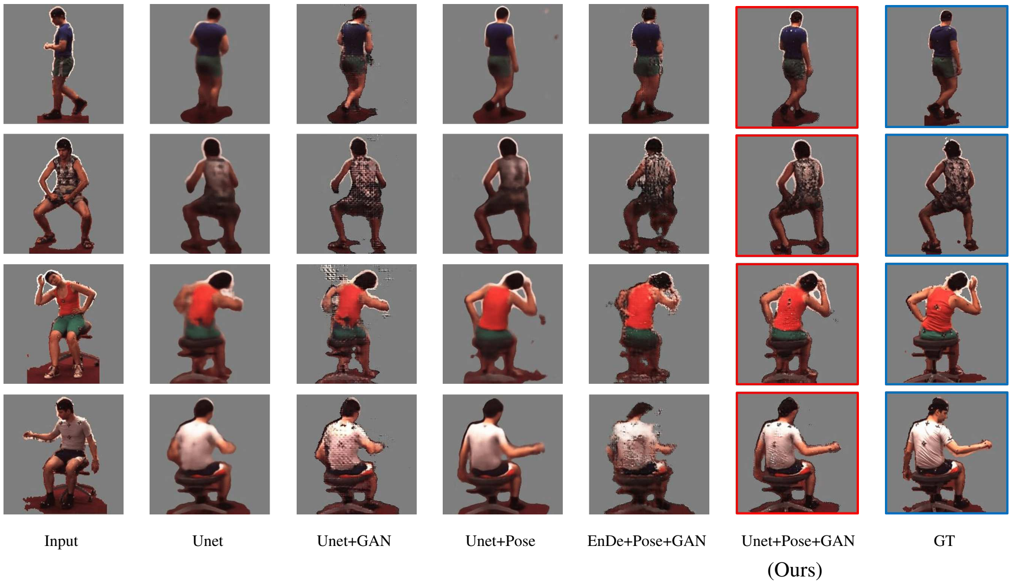

Figure 7. Visualization of the synthesized images from four foreground baselines and our foreground transformer network.

0.65 and 19.45. With two discriminator networks, FD+BD increases the SSIM and PSNR to 0.67 and 19.70. Our model (FD+GD) achieves the best performance with SSIM(0.72) and PSNR(20.62).

Fig. 6 shows the generated images of GD, FD+BD and FD+GD. We can see that the backgrounds generated by FD+BD and GD are easily distorted, *e.g.* the red circle areas. Our proposed network achieves the best results which have clear foreground and background with less noises.

4.3.3 Pose Analysis

We explore the role of 2D human pose in our model. Instead of modelling the complete human pose, we input several parts of human pose into the proposed networks, *e.g.* arms, legs and their combinations. The generated images are showed in Fig. 8. We can see that our model can generate human body appearances corresponding to these pose parts, which demonstrates that modelling pose in our model is vital for human image generation.

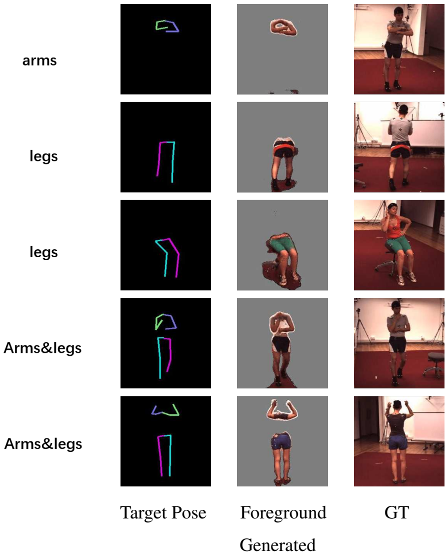

Figure 8. Images generated by inputting the parts of human pose into our proposed networks, *e.g.* arms, legs and their combinations.

4.3.4 Multiview Synthesis

We train our networks on the Human3.6M dataset which is collected from 4 cameras/viewpoints. However, our proposed model can synthesize images of more perspectives. In the experiments, we generate multiview images from a single image at intervals of 45 degrees. The results are showed in Fig. 9. We can see that our proposed method can generate high-quality multiview human images. More importantly, these images all keep correct poses, which is very difficult for the other state-of-the-art methods. We can see that the same backgrounds are shown for some different views. So there is a limitation for the background transformer network. It cannot synthesize a background that is not visible in the original image.

4.3.5 Failure Case

There are several failure cases in our experiments, which are shown in Fig. 10. One case is the occlusion between body parts, which makes our model very difficult to recover

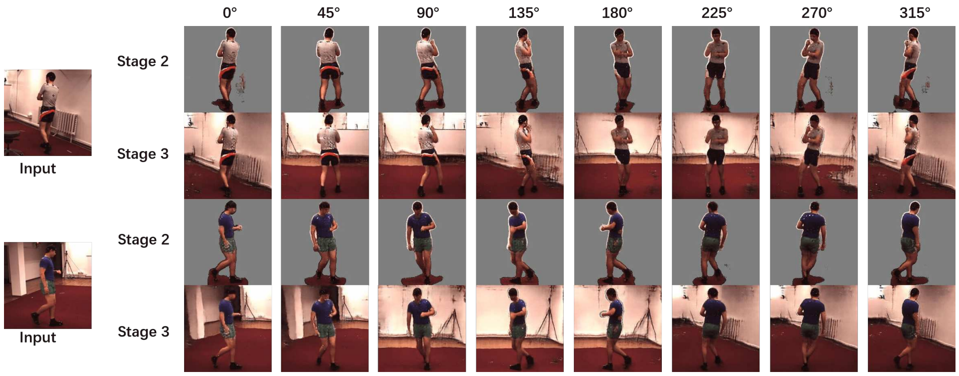

Figure 9. Multiview human images generated by our model.

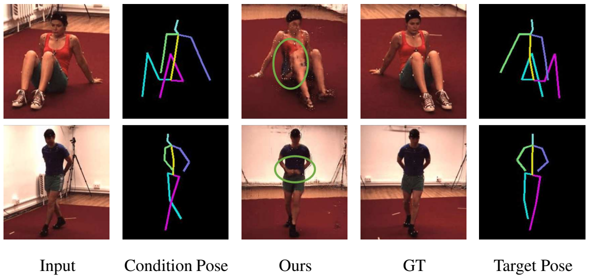

Figure 10. Two failure cases of our model.

the occluded appearance. The first row in Fig. 10 shows this case, and the appearance of the right waist has some errors. The other case is the depth missing of 2D human pose, which leads to depth disorder of body parts. The second row in Fig. 10 shows this case. The hands should be behind the back, while our model puts the hands in front.

4.4. Detailed Network Architecture

The detailed architectures of encoders f^{fg}_{pose}, f^{fg}_{img}, f^{bg}_{fg} and f^{bg}_{img} are provided in Table 3. Each convolution layer is followed by Batch Normalization and RELUs in these encoders. Table 4 shows the architectures of decoders f^{fg}_{dec} and f^{bg}_{dec}, in which each layer is followed by Batch Normalization and RELUs except the last layer. We use the Tanh units as the nonlinearity activation for the last layer.

Layer	Out channels	Kernel size	Stride	Padding
conv1	64	4 × 4	2	1
conv2	128	4 × 4	2	1
conv3	256	4 × 4	2	1
conv4	512	4 × 4	2	1
conv5	512	4 × 4	2	1
conv6	512	3 × 3	2	0
conv7	512	3 × 3	2	0

Table 3. The architectures of encoders f^{fg}_{pose}, f^{fg}_{img}, f^{bg}_{fg} and f^{bg}_{img}.

Layer	Out channels	Kernel size	Stride	Padding
deconv1	512	3 × 3	2	0
deconv2	512	3 × 3	2	0
deconv3	512	4 × 4	2	1
deconv4	256	4 × 4	2	1
deconv5	128	4 × 4	2	1
deconv6	64	4 × 4	2	1
deconv7	64	4 × 4	2	1
conv8	64	3 × 3	1	1
conv9	3	3 × 3	1	1

Table 4. The architectures of decoders f^{fg}_{dec} and f^{bg}_{dec}.

5. Conclusion and Future Work

In this paper, we propose a pose-based human image synthesis method which can keep the pose unchanged in novel viewpoints. We also propose multistage adversarial losses during model training, which contribute a lot to generate rich image details. With extensive experiments on the Human3.6M dataset, we verify the effectiveness of our model.

As can be seen, in this paper we focus on human image synthesis and do not apply the results on other visual tasks. In the future, we will further improve the image quality and apply the generated images on various visual tasks, *e.g.* cross-view gait recognition and person re-identification.

6. Acknowledgements

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61721004, 61420106015, 61572504).

References

[1] T. Chen, Z. Zhu, A. Shamir, S.-M. Hu, and D. Cohen-Or. 3-sweep: Extracting editable objects from a single photo. *ACM Transactions on Graphics (TOG)*, 32(6):195, 2013.

[2] Y.-C. Chen, W.-S. Zheng, J.-H. Lai, and P. C. Yuen. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(8):1661–1675, 2017.

[3] 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.

[4] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[5] 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(7):1325–1339, 2014.

[6] D. Ji, J. Kwon, M. McFarland, and S. Savarese. Deep view morphing. In *CVPR*, 2017.

[7] I. N. Junejo, E. Dexter, I. Laptev, and P. Pérez. Cross-view action recognition from temporal self-similarities. In *ECCV*, 2008.

[8] N. Kholgade, T. Simon, A. Efros, and Y. Sheikh. 3d object manipulation in a single photograph using stock 3d models. *ACM Transactions on Graphics (TOG)*, 33(4):127, 2014.

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

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

[11] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016.

[12] J. Liu, M. Shah, B. Kuipers, and S. Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR*, 2011.

[13] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NIPS*, 2017.

[14] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[15] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.

[16] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[17] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[18] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[19] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*, 2017.

[20] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[21] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[22] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *NIPS*, 2016.

[23] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[24] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3d models from single images with a convolutional network. In *ECCV*, 2016.

[25] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017.

[26] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, 2014.

[27] X. Yan, J. Yang, E. Yumer, Y. Guo, and H. Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *NIPS*, 2016.

[28] H.-X. Yu, A. Wu, and W.-S. Zheng. Cross-view asymmetric metric learning for unsupervised person re-identification. In *ICCV*, 2017.

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

[30] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.

[31] Y. Zheng, X. Chen, M.-M. Cheng, K. Zhou, S.-M. Hu, and N. J. Mitra. Interactive images: cuboid proxies for smart image manipulation. *ACM Trans. Graph.*, 31(4):99–1, 2012.

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

[33] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.