IEEE Signal Processing Society
IEEE Transactions on Image Processing

PoT-GAN: Pose Transform GAN for Person Image Synthesis

Journal:	*Transactions on Image Processing*
Manuscript ID	TIP-24661-2021
Manuscript Type:	Regular Paper
Date Submitted by the Author:	08-Mar-2021
Complete List of Authors:	Li, Tianjiao; Shandong University, School of Control Science and Engineering Zhang, Wei; Shandong University, School of Control Science and Engineering Song, Ran; Shandong University, School of Control Science and Engineering Li, Zhiheng; Shandong University, School of Control Science and Engineering Liu, Jun; Singapore University of Technology and Design, ISTD Li, Xiaolei; Shandong University, School of Control Science and Engineering Lu, Shijian
EDICS:	36. ARS-SRV Image and Video Synthesis, Rendering, and Visualization < Image and Video Analysis, Synthesis and Retrieval

SCHOLARONE™
Manuscripts

PoT-GAN: Pose Transform GAN for Person Image Synthesis

Tianjiao Li, Wei Zhang*, Ran Song, Zhiheng Li, Jun Liu, Xiaolei Li, and Shijian Lu

***Abstract*—Pose-based person image synthesis aims to generate a new image containing a person with a target pose conditioned on a source image containing a person with a specified pose. It is challenging as the target pose is arbitrary and often significantly differs from the specified source pose, which leads to large appearance discrepancy between the source and the target images. This paper presents the Pose Transform Generative Adversarial Network (PoT-GAN) for person image synthesis where the generator explicitly learns the transform between the two poses by manipulating the corresponding multi-scale feature maps. By incorporating the learned pose transform information into the multi-scale feature maps of the source image in a GAN architecture, our method reliably transfers the appearance of the person in the source image to the target pose with no need for any hard-coded spatial information depicting the change of pose. According to both qualitative and quantitative results, the proposed PoT-GAN demonstrates a state-of-the-art performance on three publicly available datasets for the task of person image synthesis.**

***Index Terms*—Image Synthesis, Pose Transform, Generative Adversarial Network**

I. Introduction

Pose-based person image synthesis intends to generate a high-quality person image by transferring the appearance of the person in the image from a source pose to a target one. It has attracted increasing interest in recent years, due largely to its wide applications in various computer vision tasks such as future video frame prediction [1], [2], video generation from a pose sequence [3], [4], person re-identification [5]–[7], and motion imitation [8], [9].

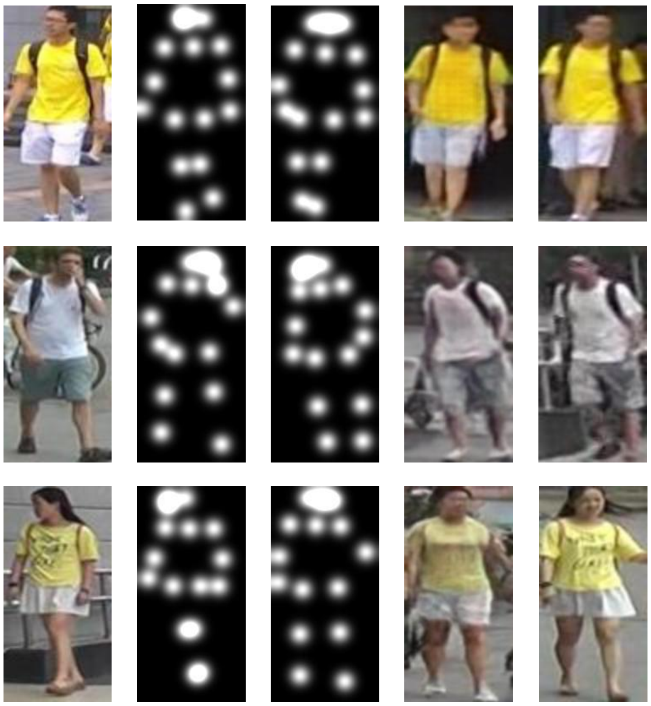

Fig. 1. Pose-based person image synthesis by the proposed PoT-GAN over three source image-pose pairs randomly selected from the Market-1501 dataset. From the leftmost column to the rightmost column: source images, source poses, target poses, synthesized images produced by the PoT-GAN, and target (ground truth) images.

Person image synthesis conditioned on person poses first introduced by [10] is a very challenging task because of the appearance discrepancy between the source and the target person images as shown in Fig. 1. Such a discrepancy mainly results from the non-rigid pose transform and the change of viewpoint between the source and the target images [11]. Although deep neural generative architectures, e.g. generative adversarial network (GAN) [12] and variational autoencoder [13] have been widely explored to synthesize realistic-looking images in various tasks, the intrinsic characteristic of translation invariance of convolutional neural networks hinders their capability of learning large spatial transforms such as the pose transforms in person image synthesis. It is thus found that many generative architectures widely used in the computer vision community cannot correctly discern such pose transform information and response properly to it in the data generation process, leading to poor synthesizing results for the particular task of person image synthesis.

Therefore, researchers have investigated various generative network architectures with specifically designed mechanisms that can capture important features of source images as well as intended poses that may implicitly indicate the knowledge on spatial transforms [10], [11], [14]–[17]. One popular mechanism is the skip connections proposed with the U-Net architecture [18] which empowers network architectures to broadcast local spatial information from encoder to decoder so that it can better preserve features learned from the source images [10]. However, early convolutional layers still cannot discern complex spatial deformations as mentioned in previous works [19], [20], and the concatenation of multi-scale feature maps in the U-Net thus becomes problematic as it just shuttles spatial information without significant transform from the

T. Li, W. Zhang, R. Song, Z. Li and X. Li are with the School of Control Science and Engineering, Shandong University, China.

T. Li and J. Liu are with the Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore.

S. Lu is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

*Corresponding author: Wei Zhang (Email: davidzhang@sdu.edu.cn)

source image to the synthesized one. Another mechanism is to use pre-calculated transform information over multi-scale feature maps [11]. However, this method requires hard-coded spatial transforms over different body parts of the person throughout the training process, which is highly inefficient and thus limits the application of the method. Furthermore, the hard-coded spatial transforms are usually modelled as affine transforms, which inevitably leads to alignment errors as the body parts of a human person are typically subject to non-rigid deformations.

It can be seen that to synthesize realistic-looking person images, the network architecture is supposed to have the capability of describing large and non-rigid spatial transform between the source and the target poses, ideally in an automatic, flexible and learnable manner. We thus propose the Pose Transform Generative Adversarial Network (PoT-GAN) which has a particular focus on the perception of the pose transform between the source and the target person images. The PoT-GAN does not rely on any hard-coded or pre-calculated information but learns a proper pose transform for person image synthesis in a fully end-to-end manner.

The proposed PoT-GAN contains a generator and a discriminator. The generator is composed of three modules: the pose transform module, the appearance encoder module and the pose encoder module. The latter two act as feature extractors while the pose transform module brings the learned knowledge about pose transform into the appearance encoder module to ensure that the pose transform information is taken into account in the process of image synthesis. The pose transform module essentially shifts each pixel of the multi-scale feature maps extracted from the source pose towards its target location of those extracted from the target pose, in accordance with the transformed coordinates calculated by the proposed pose transform module. Then the updated feature maps are passed to the decoder for image synthesis. We also design a discriminator in order to encourage the synthesized image to be indistinguishable from the real target image.

As a result of the network design mentioned above, the PoT-GAN does not rely on the knowledge of fixed affine transforms but learns more accurate representations of the typically non-rigid pose transforms. Moreover, the pose transform module works by manipulating the multi-scale feature maps via a convolutional pipeline, which ensures that the pose transform information with regard to both small and large body parts of a person can be learned. Due in part to this reason, we found that the PoT-GAN performs reliably on various person images with both local and global pose changes.

The contribution of this paper is summarized as follows:

1) We propose a new method, namely PoT-GAN for pose-based person image generation;
2) We design a new generator for the PoT-GAN which can effectively learn the knowledge on pose transform and incorporate it into the process of image synthesis;
3) We evaluate our PoT-GAN on three publicly available datasets including Market-1501, DukeMTMC and DeepFashion, and demonstrate its superiority over the state-of-the-art methods for person image synthesis.

II. Related work

In this section, we investigate the literature from three perspectives as the proposed method for person image synthesis also involves generative models and spatial transform. At the end, we highlight the fundamental differences between our method and other closely related state-of-the-art person image synthesis methods.

A. Pose-based Person Image Synthesis

Ma *et al.* [10] proposed a seminal work for pose-based person image generation where they designed a two-stage architecture to generate realistic-looking person images gradually. This work was further improved by a disentangled person image synthesis network [16] that decomposed a person image into foreground, background and pose and learned the respective latent features separately. Esser *et al.* [14] presented a variational U-Net for shape-guided image generation conditioned on the output of a variational autoencoder. Zhu *et al.* [15] employed a series of progressive pose attention blocks that can steadily attend to different body regions in order to generate person images efficiently and effectively.

The aforementioned methods treated images and poses as latent variables and generated person images within a latent space. Recently, Siarohin *et al.* [11] presented a person image synthesis method that applied a set of pre-calculated affine transformations over resolution-variant feature maps. It required hard-coded spatial transformations over different body parts of the person throughout the training process. Similarly, Balakrishnan *et al.* [21] employed directly computed affine transformation while this strategy lost global pose information as it decomposed the whole human body into several rigid parts. By contrast, the proposed PoT-GAN learns the information related to the non-rigid spatial transform between the source and the target poses automatically. In addition, it learns the multi-scale transform and generates person images by moving each pixels of the learned feature maps of the source pose towards the target pose directly.

B. Generative Models

Generative models aim to learn the true distribution of the training data to generate new data with certain variations. With the advances of GANs in recent years, the generation of realistic-looking images has been increasingly attracting interests in deep learning and computer vision communities. The original GAN [12] consists of a generator network and a discriminator network, where the generator learns to map from a latent space to the expected data distribution and produces candidate images whereas the discriminator learns to distinguish whether the generated images are from a fake or real data distribution.

Different GAN variants have been developed since the original GAN. For example, Karras *et al.* [17] proposed a progressive GAN where the generator and discriminator with symmetric structure gradually add new layers to generate higher-resolution features, shifting attention from large-scale structure of the image distribution to features of fine scale.

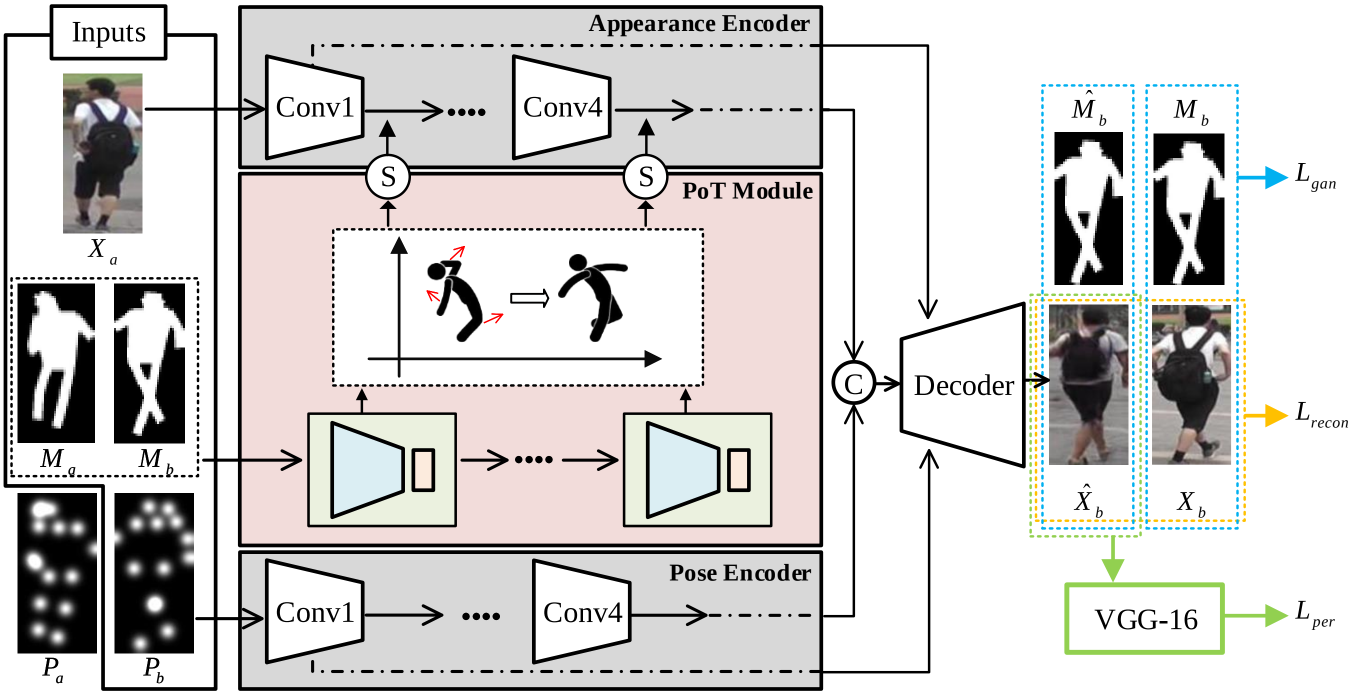

Fig. 2. Illustration of the overall architecture of the proposed PoT-GAN where the "S" operator that connects the appearance encoder and the pose transform module in the generator denotes the grid sampler. The "C" operator represents the concatenation between the shifted image features and the pose features.

By contrast, the generator/decoder in the proposed PoT-GAN does not alter the number of layers according to the pixels of generated images and thus the number of layers is fixed rather than progressive during the training. CGAN [22] learned to generate images with specific characteristics according to certain pre-defined "conditions". Analogously, the image-to-image translation GAN [23] was designed to produce images of different features based on various conditioning inputs, e.g. the translation from a segmentation map to a natural street-scene image. Qian *et al.* proposed the PN-GAN [24] that learns identity-sensitive and view-invariant features to synthesize auxiliary images for person re-identification. In recent years, cycle-consistent adversarial network [25] was introduced. It employed two generators and discriminators, and took two unpaired images conditioned on each other in a cycle-consistent manner. We also noticed that most GAN-based generation models required pixel-level alignments which can hardly be satisfied in pose-based person image synthesis that often involves large spatial discrepancy between source and target poses.

C. Spatial Transform

Spatial Transformer Network (STN) was first proposed by Jaderberg *et al.* [26] for image classification. Spatial transformer module can be incorporated into a common deep neural network to provide spatial transform abilities. The differentiable spatial transformer modules can be trained by final objective functions without any extra supervision. Considering the applications of spatial transform in person image synthesis, Balakrishnan *et al.* [21] employed directly computed affine transform instead of learnable Spatial Transformer Modules. However, this work failed to capture global transforms because the estimation of the person body as a whole is disintegrated into several rigid components.

D. Difference to Closely Related Work

Our method is fundamentally different from Ma *et al.* [10] composed of two disconnected stages for coarse image generation and person refinement. In comparison, the proposed PoT-GAN synthesizes fine-grained person images in an end-to-end manner.

Our method is also different from Siarohin *et al.* [11] although they are both based on the idea of pose transform. However, Siarohin *et al.* [11] pre-calculated a set of hard-coded affine transformations for each body part of the person while we propose a trainable pose transform (PoT) module to model the change of pose. Our method is thus more reliable due to the non-rigid nature of the transform between the source and the target poses.

Our PoT-GAN also differs significantly from Zhu *et al.* [15] where a source pose was transferred through a sequence of intermediate pose representations at a single scale based on an attention mechanism before reaching the target. By contrast, the PoT-GAN learns a one-step pose transform at multiple scales.

Our PoT-GAN is also different from the latest state-of-the-art approach based on attribute encoding proposed by Men [27] which embeds human attributes into a latent space as independent codes to archive person image generation. Instead of learning from an embedding space, our PoT-GAN exploits pose differential information explicitly between source and target human poses.

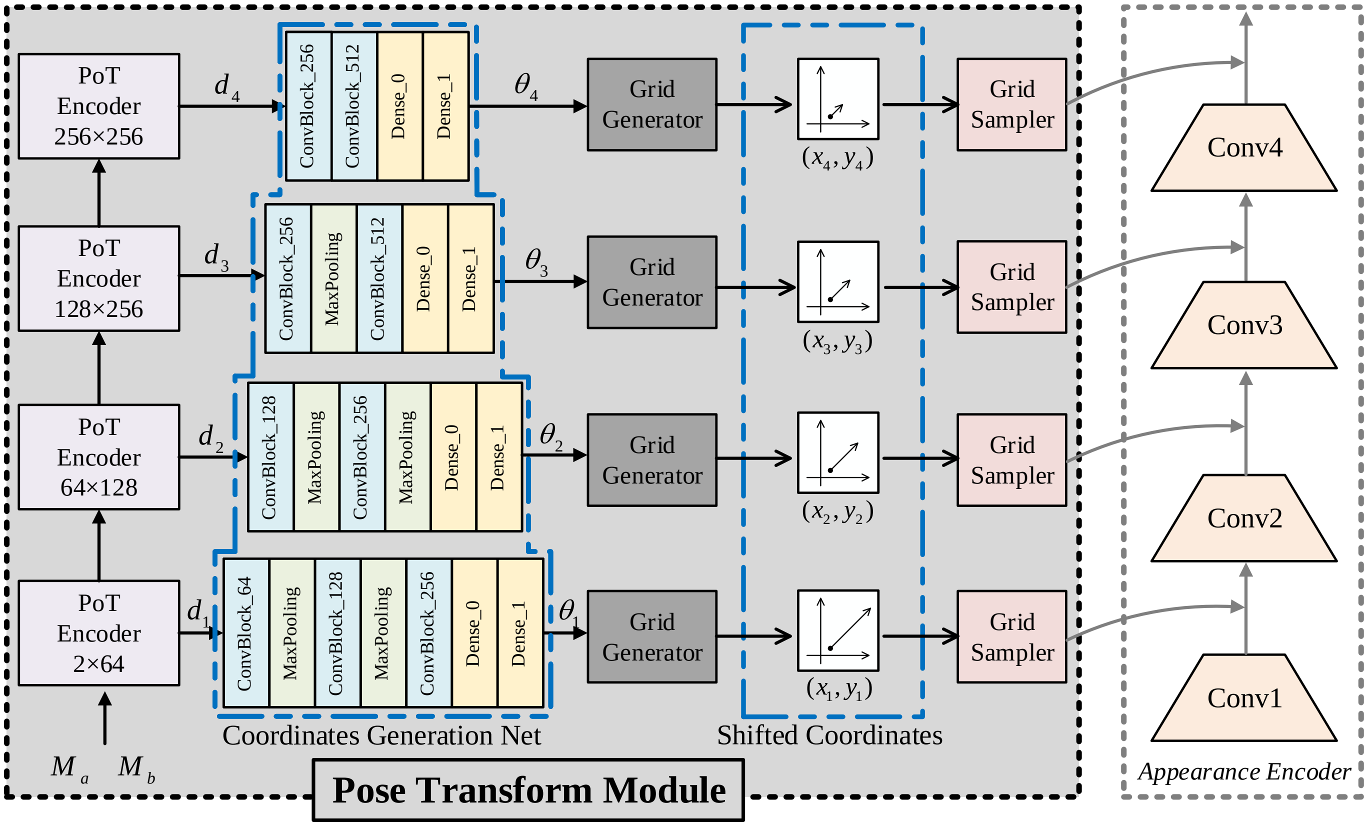

Fig. 3. Illustration of the proposed Pose Transform module. The global difference between M_a and M_b is progressively discerned at four different scales, each of which is in correspondence with a certain scale of latent feature maps extracted by the appearance encoder. The concatenated pose masks M_a and M_b are first sent into the multi-scale encoders for feature extraction. Then the extracted features provide pose transform information to calculate the transformed coordinates which explicitly represent the deformations between the source and the target poses.

III. Method

In this section, we first provide the problem statement for person image synthesis. Then we elaborate the proposed PoT-GAN with a specific focus on the pose transform module. We also give the details for training the PoT-GAN.

A. Problem Statement

Given two person images X_a and X_b, they contain the same person with different poses represented by two sets of pose landmarks P_a and P_b respectively. Person image synthesis aims to generate X_b based on X_a, P_a and P_b. It is noteworthy that following existing works [10], [11], [15], [16], the robust Human Pose Estimator (HPE) [28] is used to extract 18 human joints from person images. The pose landmarks are thus represented in the form of a matrix of 18 channels, each of which is a heat map of a certain human joint. Also, by simply connecting the pose landmarks, the pose masks M_a and M_b can be obtained. In the training phase, the required data include X_a, X_b, P_b, M_a and M_b while in the inference phase, the trained network takes X_a and P_b as inputs and synthesizes X_b.

B. Pose Transform Generative Adversarial Network

Overview. The overall architecture of the proposed Pose Transform Generative Adversarial Network (PoT-GAN) is shown in Fig. 2. Large spatial deformations can be observed between the source and the target person images due to the change of pose. Consequently, person image synthesis often suffers from insufficient differential information due in part to unaligned body parts between the source and the target person images. Therefore we propose a generative network to synthesize the target person image by discerning such differential information between the source and the target poses.

As illustrated in Fig. 2, in order to produce realistic-looking person images, the PoT-GAN is composed of a person image generator which perceives large spatial difference from the source pose M_a and the target pose M_b, and a discriminator which compares the synthesized and the target image-pose pairs. The generator of the PoT-GAN includes a pose transform module, an appearance encoder module, a pose encoder module and a decoder module. The pose transform module takes the concatenation of two pose masks M_a and M_b as input, and produces pose transform information that is further incorporated into the appearance encoder through the grid sampler which generates transformed feature maps at multiple scales.

The appearance encoder module contains N ($N = 4$ throughout the paper) convolutional blocks down-sampled one by one. The source image X_a is fed as input into the encoder, and intermediate features are extracted from each convolutional block. As discussed above, immediate concatenation of these untouched latent features causes problems for geometric alignments required by pose transform. Therefore, the output of the pose transform module is manipulated by the grid sampler to directly "move" each pixel in the feature maps corresponding to the source image accordingly.

The pose encoder module converts the target pose to a deep feature so that it can be concatenated with the output of the grid sampler which essentially integrates pose transform information with the appearance of the source image. The concatenated deep feature encodes all information of the person image to be synthesized and is then transferred to a visible person image by the decoder module subject to a U-Net backbone. Together with the target pose, the synthesized image forms an image-pose pair that intends to fool the discriminator so that it cannot correctly distinguish whether the image-pose pair contains a real target image or a synthesized one.

Pose transform module. To exploit the large differential information between source and target poses, we design a learnable pose transform module which delivers a coordinate transform for the feature maps of the source person image output by the appearance encoder module at various scales as illustrated in Fig. 3. Based on the transformed coordinates, each pixel in the feature maps is moved towards its corresponding position with regard to the target pose, which effectively conducts a pose transform. It can also be seen that the manipulated feature maps for synthesizing the intended person image are derived from the grid sampler at multiple scales to ensure that multi-scale local details of the person are well preserved during the pose transform.

Since the pose transform module is trainable, our network can exploit large non-rigid spatial deformations between source and target poses flexibly. As later demonstrated in Section IV, this significantly improves the reliability of the proposed method and enables it to produce realistic-looking person images even if the change of pose is fairly large.

As we can see in Fig. 3, first, the pose masks M_a and M_b generated by connecting the pose landmarks P_a and P_b are concatenated. Then, the concatenation is fed into the PoT encoders to produce four differential feature maps d_j:

$$d_j = \Phi_j^{\omega}(M_a, M_b), j \in \{1, 2, 3, 4\} \tag{1}$$

where ω represents the weights of the PoT encoders to be learned and j denotes the scale index. Each PoT Encoder denoted by $\Phi_i^{m\times n}$ sequentially consists of 1 convolutional layer with m input channels, 1 instance normalization layer, 1 Leaky ReLU layer, 1 convolutional layer with n output channels and 1 instance normalization layer.

As shown in Fig. 3, each of the four pathways of the pose transform module consists of a PoT encoder, a coordinates generation net F^{CGN}, a grid generator and a grid Sampler. The pose differential feature d_j in Eq. (2) is sent into a coordinates generation net F_j^{CGN} to produce the target control points θ_j at each of the 4 scales. Each coordinates generation net contains a specific number of convolutional blocks which further encodes the transform information between source and target poses, followed by a coordinates regression network to generate transformed coordinates as follows:

$$\theta_j = F_j^{CGN}(f_j), j \in \{1, 2, 3, 4\}. \tag{2}$$

The regression network is composed of two fully connected layers followed by a $Tanh$ activation function for generating the controlling points. Once the controlling points are attained, thin plate spline (TPS) [29] interpolation is employed to produce a transformed coordinates map which provides the shifted coordinates x_i and y_i. Note that the transformed coordinates map has the same size as its manipulated image feature maps. After obtaining the transformed coordinates maps and the corresponding feature maps, a re-sampling procedure is invoked to "move" every pixel in the feature maps of the source image across all of the 4 pathways.

In our method, the pose mask is treated as a whole, which means that we do not decompose the pose mask into several parts for manipulation. As a result, the overall information of the spatial pose transform can be well encoded and shuttled to the decoder.

Pose encoder module. A pose landmark encoder is deployed to better preserve the intended spatial information. Either P_a or P_b is provided in the form of a matrix with 18 channels, each of which is a human joint heat map. However, only target pose landmark P_b is available in practical training phase because redundant spatial information could be harmful to the pose encoding features empirically. Compared with sparse pose landmarks composed of several joint channels, pose masks have integrated global information in the form of stickman structure. Therefore, following the original STN [26], to attain valid global deformation information between source and target poses, single-channel greyscale pose masks are taken into account. As a result, P_b is fed into the pose encoder and transformed into four pose feature maps with variant resolutions to mitigate pose deformation at different scales.

U-Net backbone. To better transfer shifted appearance features from encoder to decoder and to preserve target spatial information, U-Net [18] structure serves as the backbone of our generator. U-Net architecture provides skip connections between encoder and decoder. Compared to the simple concatenation of the outputs of the structure and the appearance encoders like DG-NET proposed by Zheng *et al.* [30], the skip connection structure of U-Net benefits the preservation of local features for image synthesis. However, due to the large deformations between the source and target poses, untouched latent features lack geometric information of the target person images. Therefore, instead of the skip connections of untouched source image features, shifted appearance feature maps re-sampled by the calculated coordinates from the pose transform module are utilized for U-Net connections. Consequently, benefited from the skip connections between encoder and decoder, the shifted appearance information can be well preserved to produce realistic-looking target person images. During the training procedure, the concatenated features are fed into N up-sampling blocks to generate person images.

Guided Filter. Guided filter is an edge-preserving filter proposed by He *et al.* [31]. We employ it to better preserve the local features for the synthesized image. An additional guidance image is required by the guided filter to boost image quality near the guided edges. And it is natural to utilize segmentation heat maps as the demanded guidance images for human geometry preservation. Therefore, we propose to employ pose masks M as the guidance for the derived person images to sharpen the areas near the profiles of the generated person.

Discriminator. The discriminator of the PoT-GAN comprises convolutional layers and a dense layer which intends to find out if an image-pose pair contains a real target image or a fake one (i.e. the person image synthesized by the generator). Importantly, the discriminator takes as input image-pose pairs rather than person images alone so that it does not only inspect the appearance of the synthesized image, but also ensures that the relation between the synthesized image and the target pose is consistent with that between the real target image and the target pose.

C. Training Details

Let G denote the generator. The pose-based person image $\hat{X_b}$ can thus be derived by the following function:

$$\hat{X_b} = G(X_a, M_a, M_b, P_b) \tag{3}$$

where the pose masks M_a and M_b depend on the pose landmarks P_a and P_b respectively as mentioned in Section III-A. The pixel-wise reconstruction error between the synthesized image $\hat{X_b}$ and the target image X_b can then be measured by using the $L1$ loss:

$$L^G_{recon} = \|X_b - \hat{X_b}\|_1. \tag{4}$$

On the other hand, the use of the $L1$ loss alone as pixel-wise error measurement is often problematic as the $L1$ norm quintessentially leads a blurring effect and might lose high frequency information of the image [21], [23]. To alleviate this issue, we also include a perceptual loss to boost reconstruction performance. The perceptual loss was first introduced for style transfer [20] and image super-resolution [32]. It then becomes more popular in image synthesis [11], [15], [21]. The perceptual loss is a feature reconstruction loss which penalizes the output image that deviates in content from the target image:

$$L^G_{per} = \frac{1}{CHW}\|\psi(\hat{X_b}) - \psi(X_b)\|_1 \tag{5}$$

where ψ denotes the VGG-16 network [33] pre-trained on the ImageNet dataset [34]. We select $relu1\_2$ of the VGG-16 network to help regularize spatial information of the produced images. This is because a host of works [19], [20] have revealed that early layers of convolutional networks tend to obtain low level information such as edges and local shape geometries important for achieving pose-based person image synthesis of high quality.

In addition, we propose a discriminator to distinguish the synthesized image-pose pair and the target image-pose pair by using a typical adversarial loss as illustrated in Fig. 2, in order to produce realistic-looking person images with plausible geometric correspondence of the full body of the person.

$$L^D_{gan} = \mathbb{E}[\log D(X_b, M_b)] + \mathbb{E}[\log(1 - D(\hat{X_b}, M_b))]. \tag{6}$$

Algorithm 1: Learning procedure of the PoT-GAN

```
Input: Source person image ($X_a$),
       Source pose mask ($M_a$),
       Target pose mask ($M_b$),
       and Target pose landmarks ($P_b$)

Freeze $VGG-16$ for perceptual loss;
while not converge do
    Extract person image features $f_i$ from $X_a$;
    Extract pose features $p_i$ from $P_b$;
    Extract pose differential features $d_i$ from $M_a$ and
      $M_b$;
    Generate target control points $\theta_i$ from $d_i$;
    Generate shifted coordinates $x_i$ and $y_i$;
    for $i = 1;\ i <= 4;\ i = i + 1$ do
        Calculate re-sampled image feature $f_i'$ by $x_i$
          and $y_i$;
    end
    Generate transferred person image $\hat{X_b}$;
    Calculate reconstruction loss by using $\hat{X_b}$ and $X_b$;
    Calculate adversarial loss by using $\hat{X_b}$, $X_b$ and $M_b$;
    Calculate perceptual loss by using $\hat{X_b}$;
    Update generator $G$ (appearance encoder, pose
      encoder and PoT module);
    Update discriminator $D$
end
```

Meanwhile, the synthesized person image $\hat{X_b}$ and the target pose M_b are sent into the discriminator and assigned a true label to update the generator.

$$L^G_{gan} = \mathbb{E}[\log D(\hat{X_b}, M_b)]. \tag{7}$$

As a result, the full loss function for training the PoT-GAN can be formulated as:

$$L_{all} = \lambda_1 L^G_{recon} + \lambda_2 L^G_{per} + \lambda_3 L^G_{gan}. \tag{8}$$

We train the generator and the discriminator for around 90K iterations by using the Adam optimizer with β_1 = 0.5 and β_2 = 0.999. Learning rate is initially set to 2×10^{-4}. λ_1, λ_2 and λ_3 are set to 10, 1 and 0.5, respectively. And the detailed training procedure is reported in Algorithm 1.

IV. Experimental results

In this section, we evaluate the proposed PoT-GAN on three publicly available person image datasets including Market-1501 [35], DukeMTMC [36] and DeepFashion [37]. We present both quantitative and qualitative results to prove the efficacy of the pose transform module.

A. Datasets

The Market-1501 dataset contains 12,936 images for training and 19,732 images for testing. The images capture 1501 different persons from 6 different surveillance cameras. All images in this dataset are in the resolution of 128×64, and all captured human persons have various viewpoints, poses, backgrounds and illuminations.

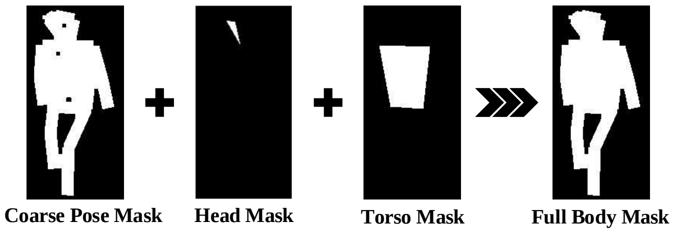

Fig. 4. Visualization of the refined full body mask.

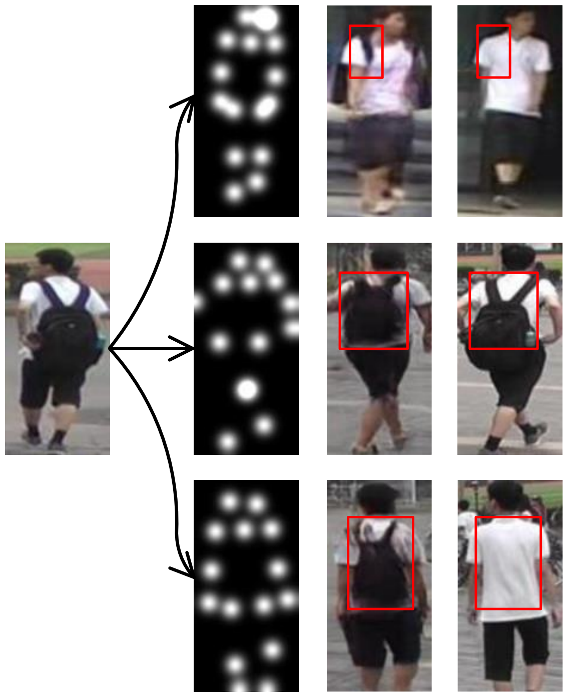

Fig. 5. A sample of the transferred results on the Market-1501 dataset. From the leftmost column to the rightmost column: the source image, the target poses, the synthesized images produced by the PoT-GAN and the target (ground truth) images. The backpack in the source image is successfully transferred to the synthesized images even if it is absent in the target image (ground truth) as shown in the first and third rows.

The DukeMTMC dataset contains 16,522 images for training and 17,661 images for testing. The images are captured by 8 cameras and vary in different resolutions. Since the sizes of person images in DukeMTMC are not fixed, the transferring task from the source to the target pose is more challenging.

The DeepFashion (*In-shop Clothes Retrieval Benchmark*) dataset contains 52,712 in-shop clothing images. All person images have the same size of 256 × 256. The in-shop clothing retrieval benchmark contains approximately 200,000 cross-pose or cross-scale pairs and 7,982 clothing images.

B. Evaluation Metrics

Evaluation of methods for pose-based person synthesis remains an open question. We adopt Structural Similarity (SSIM) [38], Inception Score (IS) [39] and Fréchet Inception Distance (FID) [40] for quantitative evaulations. For Market-1501 and DukeMTMC dataset, we also compute masked scores, mask-SSIM and mask-IS.

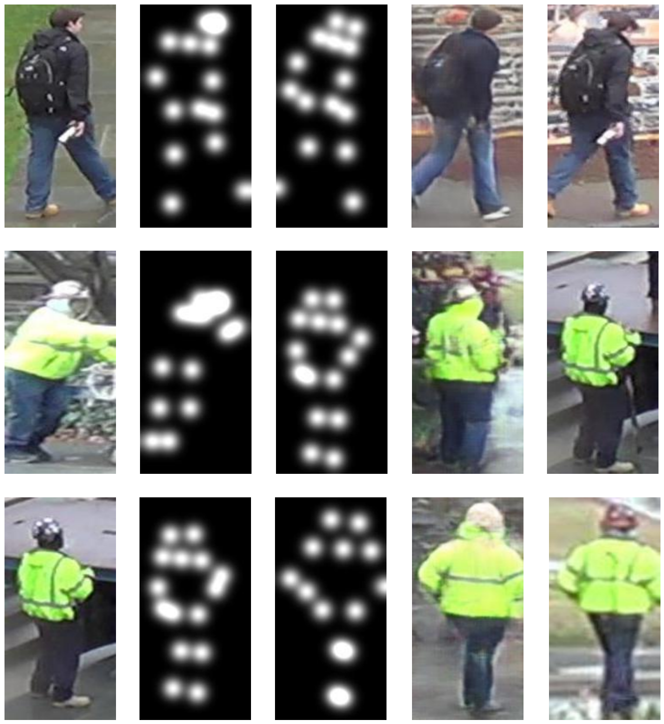

Fig. 6. Pose-based person image synthesis by the PoT-GAN over three sample images from the DukeMTMC dataset. From the leftmost column to the rightmost column: source images, source poses, target poses, synthesized images produced by the PoT-GAN, and target (ground truth) images.

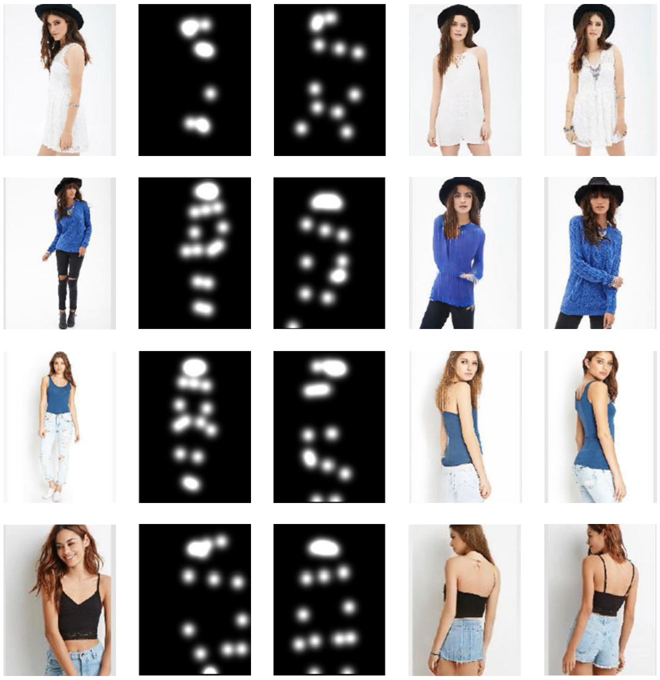

Fig. 7. Pose-based person image synthesis by the PoT-GAN over four images from the DeepFashion dataset. From the leftmost column to the rightmost column: source images, source poses, target poses, synthesized images produced by the PoT-GAN, and target (ground truth) images.

SSIM utilizes global co-variance and means in order to compute consistency score between reference person images and generated person images. IS score is essentially a pre-

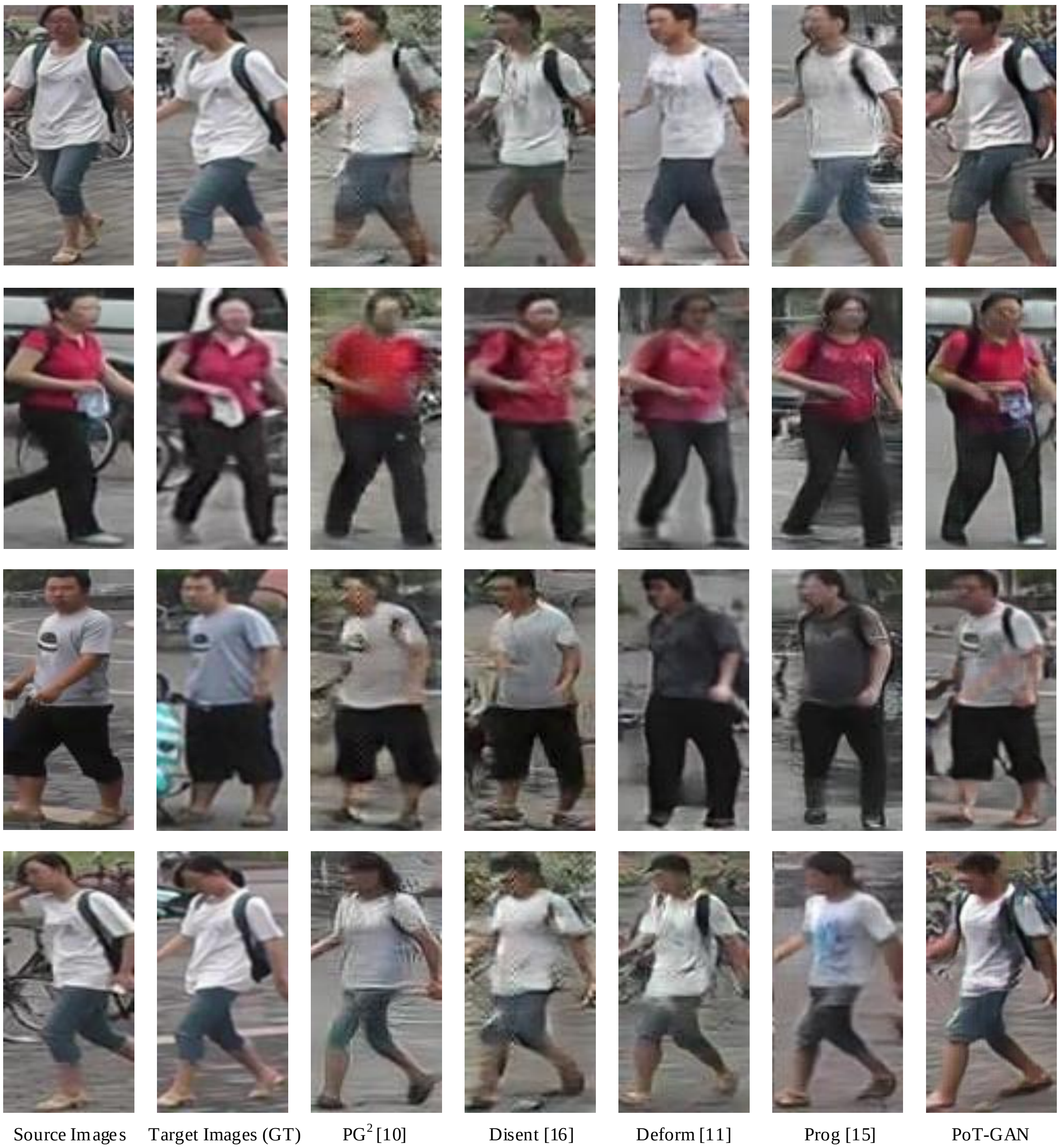

Fig. 8. Qualitative comparisons using sample images from the Market-1501 dataset between the proposed PoT-GAN and some state-of-the-art approaches including PG^2 [10], Disent [16], Deform [11] and Prog [15].

trained Inception Net [41], deploying image classifier to assess the qulity of derived images. FID uses the Inception network to extract features of two groups of images, and then calculates the Fréchet distance between two Gaussians fitted to feature representations. Note that IS scores are derived only by generated person images, and thus it is not capable of measuring the similarities between source and target poses.

However, Ma *et al.* [10] brought up that pose transfer task is rather challenging on the Market-1501 dataset because of its tangled background information. In order to reduce the influence of the background, the authors proposed variants of SSIM and IS, named mask-SSIM and mask-IS. We conduct pixel-wise multiplication between the pose mask (i.e. a binary map) and the synthesized or reference person images.

C. Pose Representations

Thanks to the development of human pose estimation methods that are able to provide quite reliable estimation results [28], [42], we follow the works in [10], [11], [14]–[16], and employ Human Pose Estimator (HPE) [28] to produce pose estimations for all the three datasets. The estimated poses are represented by 18 human pose landmarks respectively. Ma *et al.* [10] introduced two different pose representation methods, namely coordinate embedding (CE) and heat map embedding (HME). As described in their paper, the CE method increases the training complexity because CE requires one more step to

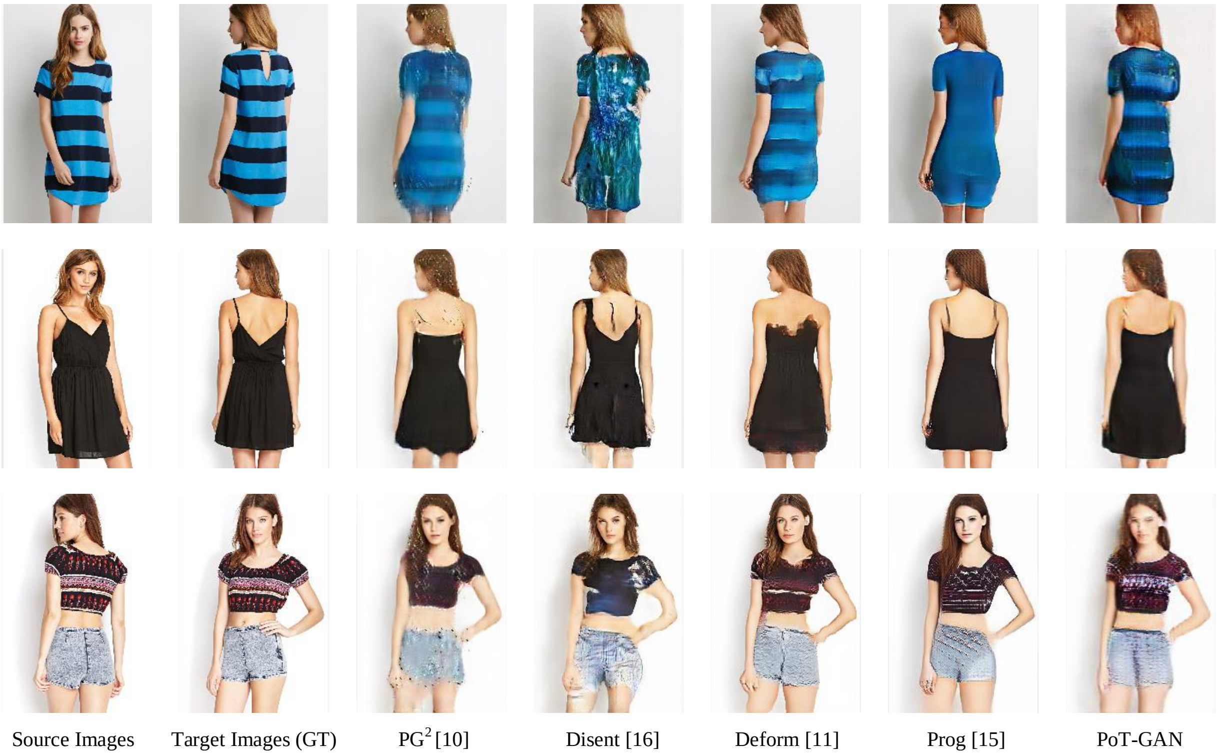

Fig. 9. Qualitative comparisons using sample images from the DeepFashion dataset between the proposed PoT-GAN and some state-of-the-art approaches including PG^2 [10], Disent [16], Deform [11] and Prog [15].

relocate coordinate embedding features back to their geometric locations. Therefore, we adopt the HME as our human pose landmark representation.

Ma *et al*. [10] also introduced pose masks to reinforce the $L1$ loss by pixel-wise multiplication between pose masks and synthesized or reference images. However, after pose key-points connection and morphological operations, large cavities can be found within head and torso areas. In our experiments, those unmasked areas would mislead spatial transform as those regions provide no useful information. To alleviate this problem, we seek to fill the cavities by depicting head and torso areas as polygons, as illustrated in Fig. 4. After that, we add up torso mask, head mask and morphological pose mask to attain the full pose mask.

D. Qualitative Results

Fig. 5 shows that the person in the source image is carrying a backpack whereas only one of the three reference images in the rightmost column contains the backpack. However, the synthesized person images in the first and third rows show that even without sufficient information in reference images, the transferred persons can still preserve the key information from the source image. It is obvious that the synthesized person in the first row is carrying a backpack as the black shoulder straps can be clearly observed. This observation is consistent with the person in the source image but conflicted to the one in the target image who does not carry a backpack. This demonstrates the pose transform module indeed has the ability to transfer appearance from source images to target poses. Moreover, Fig. 6 shows some visual results of the synthesized person images produced by the PoT-GAN over the DukeMTMC dataset.

Note that our model is trained with the mean absolute error (MAE), also known as the $L1$ loss. Although it is widely deployed as a criterion to train the generative models to encourage visual similarities between source and target images at pixel level, $L1$ loss can sometimes give rise to blurring and ghosting artifacts, or even unexpected image patches. Nonetheless, in our cases, the derived images shown in Fig. 5 provide solid evidence that despite the training with MAE, the PoT-GAN still demonstrates the ability to preserve appearance information as well as semantic information, and further indicates that it is capable of modeling global deformation.

Fig. 7 shows the results of the PoT-GAN over the DeepFashion dataset. As shown in the first two rows in Fig. 7, the hat of the model is successfully transferred from the source images to the synthesized images without large spatial distortions, which validates that the PoT-GAN has the potential to shuttle spatial information and preserve geometric invariance between source and target poses.

In addition, we qualitatively compare the proposed PoT-GAN with the state-of-the-art approaches and show some visual results in Figs. 8 and 9. It can be seen that our PoT-GAN produces the most realistic-looking person images compared to other methods including PG^2 [10] based on latent features in the third column and [15] based on Progressive Attention Network in the sixth column. In some testing images from the Market-1501 and the DeepFashion datasets, the

TABLE I
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON MARKET-1501 AND DEEPFASHION DATASETS WHERE THE SCORES CORRESPONDING TO THE TOP AND THE SECOND BEST PERFORMANCES ARE BOLDED AND UNDERLINED RESPECTIVELY.

	Market-1501				DeepFashion	
Models	SSIM	IS	mask-SSIM	mask-IS	SSIM	IS
Ma *et al*. [10]	0.253	3.460	0.792	3.453	0.762	3.090
Siarohin *et al*. [11]	0.290	3.185	0.805	3.502	0.756	**3.439**
Esser *et al*. [14]	0.353	3.241	-	-	**0.786**	3.087
Ma *et al*. [16]	0.099	3.483	0.614	3.491	0.614	3.228
Zhu *et al*. [15]	0.311	3.323	0.811	**3.773**	0.773	3.209
Men *et al*. [27]	-	-	-	-	0.772	3.364
w/o Pose transform	0.356	3.741	0.816	2.731	0.730	3.316
w/ Pose transform	**0.394**	**3.946**	**0.828**	2.775	0.775	3.365
Real data	1.000	4.100	1.000	3.471	1.000	4.088

competing methods suffer from image blurring and ghosting artefacts, due in part to the lack of the sufficient exploitation of geometric differential information between source and target poses. Moreover, the high-level latent features used by PG^2 and [10] merely convey the abstract information about some uninterpretable features rather than the concrete pose discrepancies exploited by our PoT-GAN that encourages it to learn how to "move" each body part from a source person to a target person. Therefore, our PoT-GAN can preserve more details in the original person images. For instance, the logo of the T-shirt in the source person image is successfully transferred to the person image produced by our method as shown in the third row of Fig. 8. Compared to DeformableGAN [11] that pre-calculates a set of hard-coded affine transformations, the proposed PoT-GAN is capable of automatically learning the shifted coordinates at different scales from low-level features to high-level features, and explicitly "moving" each pixel in the source feature maps to its corresponding target position. To some extent, our PoT-GAN exhibits the capability of dynamically comprehending pose differential information between source and target person images.

E. Quantitative Results

TABLE I shows quantitative comparisons with the state-of-the-art pose-based person image synthesis methods [10], [11], [14]–[16] where the results are directly collected from their released works. As shown in TABLE I, our method clearly outperforms these methods on the Market-1501 dataset in terms of most of the metrics. It can be observed that the trends of IS and mask-IS are not fully consistent with those of SSIM and mask-SSIM. The higher the SSIM scores, the more reasonable the person images. Intuitively, when the background information is masked out from the generated images, a significant improvement is expected. Moreover, extra quantitative comparisons on the DukeMTMC dataset are reported in TABLE II. Similar to Market-1501, DukeMTMC is also a Person-ReID dataset. However, more intricate backgrounds are expected for the DukeMTMC dataset. Therefore, as shown in TABLE II, when the background is masked out, the improvement for the Market-1501 dataset is less than that for the DukeMTMC dataset in terms of the SSIM.

TABLE II
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON DUKEMTMC DATASET WHERE THE SCORES CORRESPONDING TO THE TOP AND THE SECOND BEST PERFORMANCES ARE BOLDED AND UNDERLINED RESPECTIVELY.

	DukeMTMC			
Models	SSIM	IS	mask-SSIM	mask-IS
Ma *et al*. [10]	0.356	3.379	0.849	2.161
Ma *et al*. [16]	0.283	3.579	0.716	2.635
Zhu *et al*. [15]	0.361	3.096	0.831	**2.810**
w/o Pose transform	0.349	3.625	0.849	2.287
w/ Pose transform	**0.380**	**3.931**	**0.902**	2.716
Real data	1.000	4.093	1.000	3.258

TABLE III presents the quantitative comparisons with [10], [11], [15], [16] in terms of FID. We calculated the FID score between source and target images. The results demonstrate that our method makes a good balance between image authenticity and similarity. As shown in TABLE III, our method achieves the lowest FID score on DeepFashion dataset but is outperformed by Zhu *et al*. [15] on Market-1501 and DukeMTMC datasets. Presumably, this is because [15] is more reliable to complex background due in part to the attention scheme it contains.

On the other hand, the PoT-GAN performs comparably well with the competing state-of-the-art methods in terms of most metrics on the DeepFashion dataset. This is because the DeepFashion dataset consists of a large amount of samples that have both full-body and half-body person images sharing the same appearance. When transferring from half-body images to full-body images, the extracted appearance information is usually insufficient to predict a full-body person image due to the missing of the information about the lower half body. It is noteworthy that the pose transform module is essentially a "manipulator", instead of a "creator", which means that it cannot impaint what is missing from the source images. Thus, the pose transform module can well preserve what the source image contains, but is not good at creating large image areas that does not exist in it.

TABLE III
COMPARISONS WITH THE STATE-OF-THE-ART METHODS ON 3 DATASETS IN TERMS OF THE FID SCORES. LOWER SCORES MEANS BETTER RESULTS.

	Market-1501		DukeMTMC		DeepFashion
Models	FID	mask-FID	FID	mask-FID	FID
Ma *et al.* [10]	131.950	28.220	218.478	75.326	75.780
Siarohin *et al.* [11]	73.821	-	-	-	23.219
Ma *et al.* [16]	42.609	19.039	177.153	70.825	37.263
Zhu *et al.* [15]	38.316	11.970	95.631	34.466	19.749
Ours	54.561	36.091	106.210	56.875	18.529

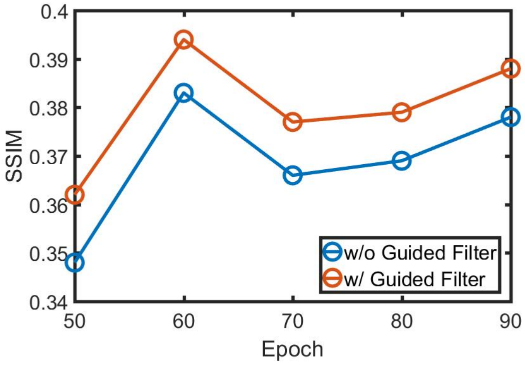

Fig. 10. Ablation study for the guided filter over the Market-1501 dataset

F. Ablation Studies

Fig. 10 carries out an ablation study for the guided filter using the Market-1501 dataset. We select five SSIM scores varying between 50 and 90 training epochs to validate the impact of the filter guided by the pose mask on the synthesized person images. The overall improvement in terms of SSIM demonstrates that the guided filter conditioned on the fine-grained full body mask indeed works for synthesizing the person images of higher quality.

TABLE IV shows another ablation study for the loss L_{recon} on 3 datasets. We introduce L_{recon} to guide pixel-wise image reconstruction. The SSIM and IS scores shown in TABLE IV reflects the effect of L_{recon} on the generation of complex pixels during the training process. It can be seen that L_{recon} significantly improves the performance on all 3 datasets. Since the background of DeepFashion is completely white, the performance gain caused by L_{recon} is proportionally smaller than that of the other two datasets.

TABLE IV
ABLATION STUDY FOR L_{recon} OVER 3 DATASETS. FROM TOP TO BOTTOM, THE THREE ROWS REPRESENT THE RESULTS ON MARKET-1501, DUKEMTMC AND DEEPFASHION DATASET, RESPECTIVELY

Models	SSIM	IS	mask-SSIM	mask-IS
w/o L_{recon}	0.234	2.355	0.724	2.086
w/ L_{recon}	0.394	3.946	0.828	2.775
w/o L_{recon}	0.297	3.631	0.774	2.473
w/ L_{recon}	0.380	3.931	0.902	2.716
w/o L_{recon}	0.680	2.031	-	-
w/ L_{recon}	0.775	3.365	-	-

V. CONCLUSIONS

In this paper, we propose a pose-based person image synthesis method based on a GAN architecture. The proposed PoT-GAN is capable of transferring appearance information from the source pose to the target one by learning a pose transform mechanism. This is achieved by the specifically designed pose transform (PoT) module that can directly manipulate multi-scale feature maps in order to solve the non-rigid transform between the source and the target poses. Compared with previous methods where the pose transform is usually modeled by fixed representations, the PoT-GAN estimates it through learnable convolutional neural networks, leading to more accurate person image synthesis.

Future work will investigate the design of a creator module that can provide the desired functionality of impainting to endue the network with the capability to make predictions on some unseen body parts of a person. This could further improve the reliability of person image synthesis particularly when the target image contains large areas unseen in the source image.

REFERENCES

[1] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 3560–3569.

[2] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 204–219.

[3] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *Proceedings of the ACM International Conference on Multimedia*, 2017, pp. 199–207.

[4] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3352–3361.

[5] W. Zhang, X. He, X. Yu, W. Lu, Z. Zha, and Q. Tian, "A multi-scale spatial-temporal attention model for person re-identification in videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3365–3373, 2020.

[6] W. Zhang, X. He, W. Lu, H. Qiao, and Y. Li, "Feature aggregation with reinforcement learning for video-based person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 12, pp. 3847–3852, 2019.

[7] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3774–3782.

[8] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE International Conference on Computer Vision*, October 2019.

[9] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro, "Video-to-video synthesis," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[10] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2017, pp. 406–416.

[11] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3408–3416.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the International Conference on Neural Information Processing Systems*, 2014, pp. 2672–2680.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes." in *Proceedings of the International Conference on Learning Representations*, 2014.

[14] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8857–8866.

[15] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2347–2356.

[16] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, 10 2015, pp. 234–241.

[19] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 5188–5196.

[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2016.

[21] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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

[23] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967–5976.

[24] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 650–667.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[26] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2017–2025.

[27] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[28] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[29] F. L. Bookstein, "Principal warps: thin-plate splines and the decomposition of deformations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.

[30] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2138–2147.

[31] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2013.

[32] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 105–114.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[35] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[36] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 17–35.

[37] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[38] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, 2004.

[39] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems 29*, 2016, pp. 2234–2242.

[40] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.

[41] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[42] J. Liu, H. Ding, A. Shahroudy, L. Duan, X. Jiang, G. Wang, and A. C. Kot, "Feature boosting network for 3d pose estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 494–501, 2020.

PoT-GAN: Pose Transform GAN for Person Image Synthesis

Revision Report of TIP-23292-2020

Dear Associate Editor,

We regret that our paper was rejected in the previous round of review. Although the comments from the reviewers are much appreciated, we believe that the contribution of our work is underestimated. Thus here we resubmit a revised version of the manuscript where all critical comments have been taken carefully into consideration. Without them, the quality of the resubmitted paper could not have been improved.

Below are our responses to each individual reviewer. To make it self-contained, each critical comment made by the reviewers is presented in *Italic* font, followed by our response to it.

To Reviewer #1:

1. *The model overfits the target dataset.*

- *How can the model learn the target background, given the target pose only?*

- *The results in the first row of Figure 1 and Figure 6 are unreasonable. I guess that the model overfits the datasets, which is not desirable.*

Response: This is a misunderstanding as we provide not only the target pose, but also the source and target images to train the network. Our network learns the target background by minimizing the difference between synthesized and target images through training. This is principally implemented by utilizing the L_1 loss to measure the pixel-wise reconstruction error between the synthesized and the target images as elaborated in Section III-C.

In the first row of Figure 1, since the background of the target image is mostly simply black, our PoT-GAN can easily learn such a common type of background. For the target image in the first row of Figure 6, the DukeMTMC dataset contains many images with a similar background texture (i.e. bricks that form a wall). Thus the network is capable of learning such a background for image synthesis through training. However, it can be seen that although the network has reconstructed the texture of the background in the target image well, the dominant tone of the reconstructed background is still biased towards the source image due to incomplete conversion at pixel-level. In other words, some attributes of the source background are still retained in the reconstructed image.

It is worth mentioning that we also showed quite a few examples where the backgrounds in the reconstructed images are more similar to the source images rather than the target images. For example, in the first row of Figure 8, the background of the reconstructed image contains features such as bicycles which exist in the original image. In the third row of Figure 8, the reconstructed background is more consistent with that of the source image, and particularly, it does not contain the prominent light blue object at the bottom left like the target image. These results indicate that our method does not overfit the datasets.

To make our statement more convincing, in the new version of the paper, we have evaluated the PoT-GAN on the test sets of three public datasets including Market-1501, DukeMTMC and DeepFashion. The quantitative results suggest that our method does not suffer from overfitting.

2. *The main idea is not new.*

- The main pipeline of the model is close to [Ma 2017] and [a]. The losses are similar as well.

[a] Qian X, Fu Y, Xiang T, et al. Pose-normalized image generation for person re-identification[C]//Proceedings of the European conference on computer vision (ECCV). 2018:650-667.

- The encoder with multi-scale information is not new, which is close to Progressive GAN [b] and [15].

- The decoder structure is similar to DG-Net [c].

[b] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation[J]. arXiv preprint arXiv:1710.10196, 2017.

[c] Zheng, Zhedong, et al. "Joint discriminative and generative learning for person re-identification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.

Response: Thanks for the suggestion of the related papers. In the resubmitted paper, we have discussed the differences between our PoT-GAN and all of these papers, summarized as below.

The difference between our method and [Ma 2017] has been explained in the first paragraph of Section II-D. Compared to their structure composed of two disconnected stages for coarse image generation and person refinement, our proposed PoT-GAN synthesizes fine-grained person images in an end-to-end manner.

The PN-GAN proposed by Qian *et al.* does not contain a component for pose transfer. This is because it is designed to address the problem of person re-identification by learning identity-sensitive and view-invariant features in the process of large pose variations as mentioned in the second paragraph of Section II-B in the resubmitted paper. In comparison, the pose transfer component, the PoT module is the main contribution of our work which imposes pixels to shift according to pose changes.

In terms of loss function, in addition to the basic loss of generative adversarial network and the pixel-wise L_1 loss for image reconstruction, we incorporate the perceptual loss to alleviate the negative impact caused by L_1 norm.

The difference between our method and [15] has been explained in the third paragraph of Section II-D. [15] proposed Pose-Attentional Transfer Network to transfer the source pose through a sequence of intermediate pose representations at a single scale based on an attention mechanism to generate target images. In contrast, our PoT Module transforms the source pose at multiple scales in one step rather than progressively.

As mentioned in the second paragraph of Section II-B in the resubmitted paper, Karras *et al.* proposed a progressive training methodology for generative adversarial networks. The generator and discriminator with symmetric structure gradually add new layers to generate higher-resolution features, shifting attention from large-scale structure of the image distribution to features of fine scale. By contrast, the generator/decoder in our PoT-GAN does not alter the number of layers according to the pixels of generated images and thus the number of layers is fixed rather than progressive during the training.

As mentioned in the "U-Net backbone" section in Section III-B, the backbone of our PoT-GAN is based on U-Net with proposed skip connection instead of simply concatenating the outputs of the structure and the appearance encoders like DG-NET proposed by Zheng *et al.* Our decoder generates target images in accordance with the concatenated features and feature maps output by previous convolutional layers of the appearance and the pose encoders.

3. *The results do not convince me.*

- *It would be provide the FID performance, which is more stable and widely adapted in terms of image quality.*

- *The ssim of real data is not 1.00. It would be good to report the intra-dataset ssim score.*

- *How to generate the background? (The first row in Fig.6) The model overfits the target dataset.*

Response: Thanks for the suggestion. We have provided the FID performance in TABLE III of the resubmitted manuscript, listed below. It can be seen that our method achieves the lowest FID score on DeepFashion dataset but is outperformed by Zhu *et al.* [15] on Market-1501 and DukeMTMC datasets. Presumably, this is because [15] is more reliable to complex background due in part to the attention scheme it contains.

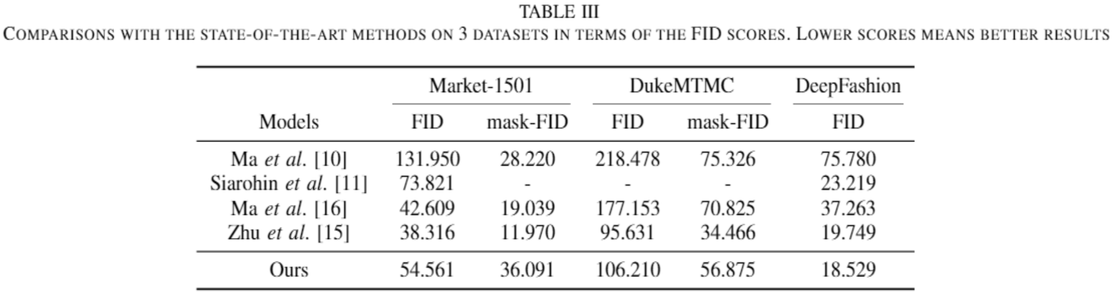

TABLE III
Comparisons with the state-of-the-art methods on 3 datasets in terms of the FID scores. Lower scores means better results

	Market-1501		DukeMTMC		DeepFashion
Models	FID	mask-FID	FID	mask-FID	FID
Ma *et al*. [10]	131.950	28.220	218.478	75.326	75.780
Siarohin *et al*. [11]	73.821	-	-	-	23.219
Ma *et al*. [16]	42.609	19.039	177.153	70.825	37.263
Zhu *et al*. [15]	38.316	11.970	95.631	34.466	19.749
Ours	54.561	36.091	106.210	56.875	18.529

The SSIM of real data is 1.00 as reported in [15]. And we cannot find any paper that defines or uses the intra-dataset SSIM score in the literature.

We use pose masks, source images and target images for training. There are quite a few images with backgrounds similar to the that of the images in the first row of Fig.6 in DukeMTMC dataset. During the training, the L_1 loss measures the pixel-wise reconstruction error to implement the conversion of background pixels, and thus the tone of reconstructed image in the first row of Fig.6 still retains the dark gray tone of the source image. There also exist many reconstructed backgrounds more similar to the source images, such as those shown in the first and third rows of Fig.8.

The PoT-GAN has demonstrated superior performance on three different public datasets, which indicates that it generalizes well for different datasets and is less likely to overfit them. In addition, The updated TABLE II (see below) in the resubmitted paper shows the evaluation results of the competing methods on the DukeMTMC dataset.

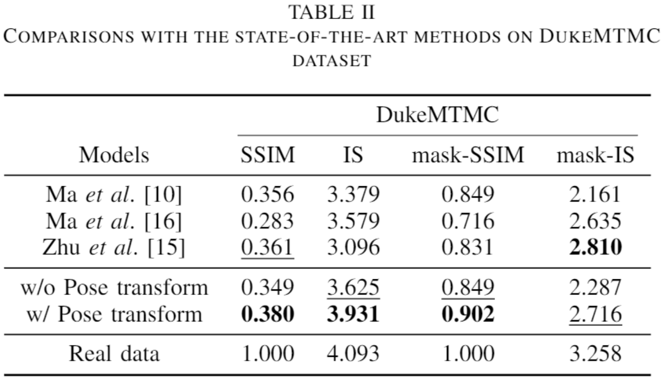

TABLE II
Comparisons with the state-of-the-art methods on DukeMTMC dataset

	DukeMTMC			
Models	SSIM	IS	mask-SSIM	mask-IS
Ma *et al*. [10]	0.356	3.379	0.849	2.161
Ma *et al*. [16]	0.283	3.579	0.716	2.635
Zhu *et al*. [15]	0.361	3.096	0.831	**2.810**
w/o Pose transform	0.349	3.625	0.849	2.287
w/ Pose transform	**0.380**	**3.931**	**0.902**	2.716
Real data	1.000	4.093	1.000	3.258

To Reviewer #2:

1. *The novelty of this paper is very limited.*

Response: As summarized in Section 1, the main novelty of this paper is the new generator of the proposed PoT-GAN which can effectively learn the knowledge on pose transform and incorporate it into the process of pose-based person image synthesis. We evaluate it on three publicly available datasets demonstrate its superiority over the state-ofthe-art methods.

In the resubmitted paper, we have discussed the differences between our method and the closely related work in Section II-D.

2. *Experiments are not sufficient to demonstrate the effectiveness of each part of the proposed method. Authors only provide experimental results for the final comparison. The lack of ablation experiments and parametric discussions will influence the judgment on the effectiveness of the proposed method. For examples,*

- the effect of L_{recog} on the final performance

- the effect of adding the guided filter in the proposed framework

Response: In this paper, we introduce the L_{recon} loss to guide low-frequency pixel-wise image reconstruction, including human body and background information. As suggested by the reviewer, we have conducted the ablation experiments of L_{recon} and reported the results in TABLE IV (see below) of the resubmitted paper.

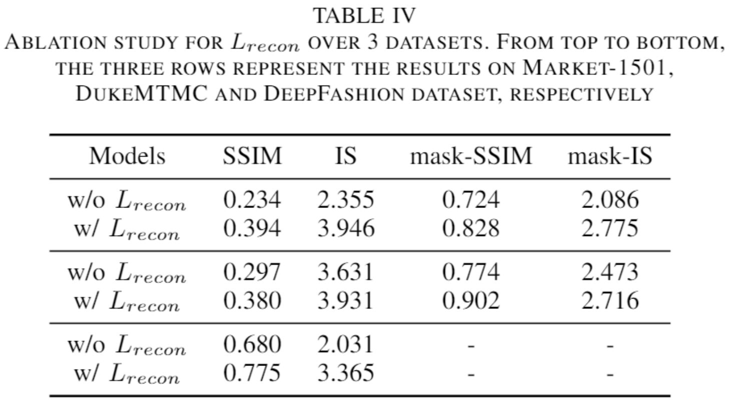

TABLE IV
ABLATION STUDY FOR L_{recon} OVER 3 DATASETS. FROM TOP TO BOTTOM, THE THREE ROWS REPRESENT THE RESULTS ON MARKET-1501, DUKEMTMC AND DEEPFASHION DATASET, RESPECTIVELY

Models	SSIM	IS	mask-SSIM	mask-IS
w/o L_{recon}	0.234	2.355	0.724	2.086
w/ L_{recon}	0.394	3.946	0.828	2.775
w/o L_{recon}	0.297	3.631	0.774	2.473
w/ L_{recon}	0.380	3.931	0.902	2.716
w/o L_{recon}	0.680	2.031	-	-
w/ L_{recon}	0.775	3.365	-	-

It can be seen that L_{recon} significantly improves the performance on the three datasets. It is worth noting that since the background of the DeepFashion dataset is completely white in both training and testing sets, the performance gain caused by L_{recon} is proportionally smaller than that of the other two datasets.

Moreover, the effect of adding the guided filter in the proposed framework has already been investigated in the paper (See Fig. 10 of the revised and resubmitted paper).