Loss Functions for Person Image Generation

Haoyue Shi[1]
shyern@stu.xjtu.edu.cn
Le Wang ✉[1]
lewang@xjtu.edu.cn
Wei Tang[2]
tangw@uic.edu
Nanning Zheng[1]
nnzheng@xjtu.edu.cn
Gang Hua[3]
ganghua@gmail.com

[1] Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University
Xi'an, Shaanxi, China

[2] Department of Computer Science
University of Illinois at Chicago
Chicago, Illinois, USA

[3] Wormpex AI Research
Bellevue, Washington, USA

Abstract

Pose guided person image generation aims to transform a source person image to a target pose. It is an ill-posed problem as we often need to generate pixels that are invisible in the source image. Recent works focus on designing new architectures of deep neural networks and have shown promising results. However, they simply adopt the loss functions commonly used for generic image synthesis and restoration, e.g., L1-norm loss, adversarial loss, and perceptual loss. This can be suboptimal due to the unique appearance and structure patterns of person images. In this paper, we first have a comprehensive study of these prior loss functions for person image generation. We also consider the structural similarity (SSIM) index as a loss function since it is widely used as the evaluation metric and can capture the perceptual quality of generated images. Moreover, motivated by the observation that a person can be divided into part regions with homogeneous pixel values or textures, we extend the SSIM into a novel part-based SSIM loss to explicitly account for the articulated body structure. Quantitative and qualitative results indicate that (1) using different loss functions significantly impacts the generated person images and (2) the proposed part-based SSIM loss is complementary to the prior losses and helps improve the performance.

1 Introduction

Pose guided person image generation means to transform a person image from a source pose to a target pose while retaining the appearance details. It has many valuable applications such as movie making, image editing and data augmentation for person re-identification and action recognition. This task is very challenging especially in the case of human body occlusion, large pose transfer and complex textures.

Recent approaches are built on deep neural networks and have demonstrated encouraging performance. Ma *et al.* [□] propose a two-stage framework based on the generative adversarial network (GAN). It first generates an initial but coarse image, which is then refined

Corresponding author: Le Wang

in an adversarial way. Zhu *et al.* [28] introduce a simple pose-attentional transfer architecture, which can generate person images progressively. Recently, Ren *et al.* [16] design a differentiable global-flow local-attention framework to reassemble the input to a target pose.

As the essential component of deep learning methods, the loss function guides the neural network to produce the desired output. However, all prior work in person image generation focuses on designing new network architectures and simply adopts the loss functions commonly used for generic image synthesis and restoration, e.g., L1-norm loss, adversarial loss, and perceptual loss. They have several limitations. Both the L1-norm loss and the perceptual loss compute the elementwise difference of images or feature maps. But they are extremely sensitive to the spatial misalignment. The adversarial loss [5] discriminates generated samples from real ones but often ignores the detailed texture. Although the combination of different losses leads to promising results, it is still unclear how each individual loss impacts the generated person images. Moreover, there is surprisingly no work on designing loss functions to account for the unique appearance and structure patterns of person.

The goal of this paper is to study the loss functions for person image generation, which is largely ignored by existing work. We first compare the strengths and weaknesses of prior losses widely used in this task. Extensive ablation study is performed to demonstrate the impact of different loss functions on generated person images. We also consider the structural similarity (SSIM) index [25] as a loss function since it is widely used as the evaluation metric and can capture the perceptual quality of generated images. By calculating the statistics on the patch level, the SSIM value compares the local textures of two images and is invariant to small spatial misalignments. Moreover, motivated by the observation that a person can be divided into part regions with homogeneous pixel values or textures, we extend the SSIM into a novel part-based SSIM loss to explicitly account for the articulated body structure. Quantitative and qualitative results on two benchmark datasets indicate that (1) using different loss functions significantly impacts the generated person images and (2) the proposed part-based SSIM loss is complementary to the prior losses and helps improve the quality of generated images. As shown in Fig. 1, the adversarial loss (GAN) makes sharper images but neglects the detailed texture; the part-based SSIM loss (pSSIM) captures more detailed texture but often results in blurred images. Then the combination of both two losses (GAN+pSSIM) preserves the detailed texture and makes a clear border of persons. Further including the perceptual loss (GAN+Per+pSSIM) makes the generated images more realistic.

2 Related Work

Pose guided person image generation. Ma *et al.* [11] are the first to study the problem of pose guided person image generation. They propose a two-stage generation approach with adversarial training. Ma *et al.* [12] improve their method by disentangling person image into three types of embedding features then re-compose them back to the desired image. Esser *et al.* [3] disentangle the appearance and pose of a person image using a variational autoencoder combined with the conditional U-Net [17]. [14, 22] utilize a bidirectional strategy to synthesize person images in an unsupervised manner. To better handle the non-rigid body deformation in large pose transfer, Siarohin *et al.* [20] propose deformable skip connections warping local image feature according a set of local affine transformations. Li *et al.* [8] use the 3D appearance flow between the source and target poses calculated by an additional 3D human model to warp features of the input image. The pose transfer model proposed by Zhu *et al.* [28] draws a great attention in recent years, which introduces cascaded Pose-

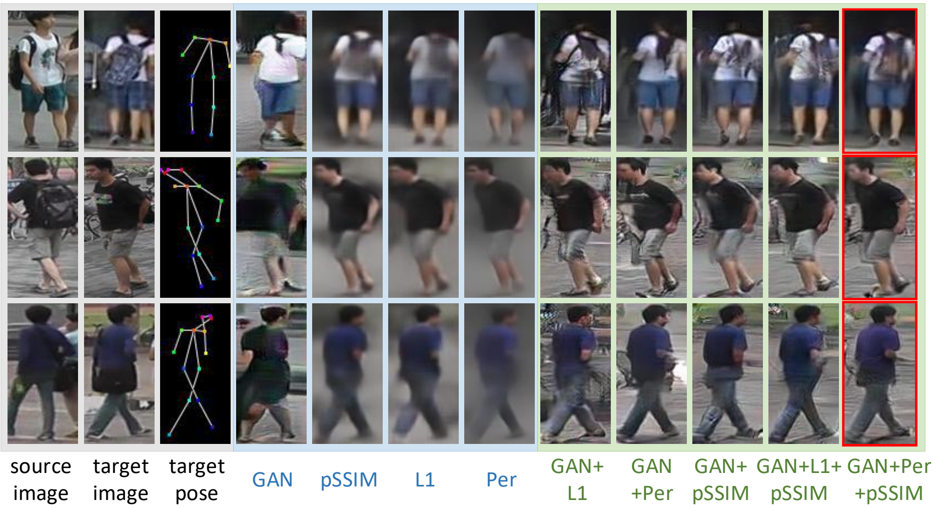

Figure 1: Person images generated by models trained with different loss functions.

Attentional Transfer Blocks into the generator to transform the source data. The most recent work by Ren *et al.* [16] designs a differentiable global-flow local-attention framework to reassemble the input to a target pose. However, these prior approaches focus on new network architectures and simply adopt the loss functions commonly used for generic image synthesis and restoration.

Image quality evaluation. The image quality evaluation is essential for image generation methods to synthesize desired outputs. Recent image synthesis research [1, 4, 6, 13, 15, 24] commonly uses simple loss functions to measure the difference between the generated image and the ground truth, e.g., L1-norm loss, adversarial loss, and perceptual loss. The conditional approaches [6, 13] solving the image restoration task typically use the L1-norm loss to compute the pixel-to-pixel difference in images. The adversarial loss proposed by Goodfellow *et al.* [5] discriminates generated samples from real ones, which is commonly used in image generation tasks [4, 6, 13]. The perceptual loss [7] widely used in the style transfer task [4] computes the element-by-element difference in feature maps. However, the adversarial loss ignores the detailed texture and the element-by-element losses are too sensitive to the spatial misalignment.

As the most popular image quality evaluation metric, the structural similarity (SSIM) index [25] aims to compare the luminance, contrast and structure information in images based on the assumption that the Human Visual System (HVS) is sensitive to changes in local information. However, the computation of SSIM [25] index in each corresponding pixel of two images only looks at a fixed neighborhood patch region, which neglects the specific structure of human body. The SSIM [25] index has also been used in image compression [24], image reconstruction [1], denoising and super-resolution [15]. To the best of our knowledge, we are the first to use the structural similarity (SSIM) index as a loss function for the pose guided person image generation task. Furthermore, we observe that a person can be divided into part regions with homogeneous pixel values or textures. This motivates us to extend the SSIM to a novel part-based SSIM loss to explicitly account for the articulated body structure.

3 Method

3.1 Problem statement

The source image I_s and the target image I_t are two images of the same person in different poses (denoted as P_s and P_t respectively). Given I_s, P_s and P_t, we aim to synthesize a new image $\hat{I}_t$ which is a prediction of I_t. The human pose of the source image is extracted by an off-the-shelf human pose estimator [2]. It uses a sequence of 2D coordinates to describe the locations of body joints in an image. In order to leverage the spatial nature of pose, we use K heat maps making up a 3D volume in $R^{W \times H \times K}$ to represent a 2D pose, e.g., P_s or P_t, where K, W and H are respectively the number of body joints, the width and height of the input image. Each heat map contains a Gaussian mask centered at the corresponding body joint location.

Prior work in pose guided person image generation simply adopts the loss functions commonly used for generic image synthesis and restoration, e.g., L1-norm loss, adversarial loss, and perceptual loss. The L1-norm loss calculates the L1 norm of the difference between the synthetic image $\hat{I}_t$ and the ground truth real image I_t:

$$\mathcal{L}_{L1}(I_t, \hat{I}_t) = \|\hat{I}_t - I_t\|_1 \tag{1}$$

The perceptual loss commonly calculates the L1-norm distance between two feature maps respectively extracted from $\hat{I}_t$ and I_t by a pre-trained network. It can be written as:

$$\mathcal{L}_{Per}(I_t, \hat{I}_t) = \sum_i \|\phi_i(\hat{I}_t) - \phi_i(I_t)\|_1 \tag{2}$$

where ϕ_i is the output of the i^{th} layer of a pre-trained network, e.g., *conv1_2* of a VGG-19 [21] pre-trained on ImageNet [18]. The adversarial loss [5] uses a *discriminator* to force the distribution of generated images to mimic that of real images. Zhu *et al.* [28] utlize two discriminators D_A and D_S to respectively ensure the appearance consistency and shape consistency between I_t and $\hat{I}_t \equiv G(I_s, P_t)$:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{I_s,P_t,I_t}\left[\log\left(D_A(I_s,I_t)\cdot D_S(P_t,I_t)\right)\right] + \\ &\mathbb{E}_{I_s,P_t}\left[\log\left((1-D_A(I_s,G(I_s,P_t)))\cdot(1-D_S(P_t,G(I_s,P_t)))\right)\right]\end{aligned} \tag{3}$$

where G denotes a *generator* [5].

3.2 Structural SIMilarity (SSIM) loss

The Structural SIMilarity (SSIM) index [25] is a perceptually motivated metric which decomposes the similarity measurement task into three comparison functions: luminance (l), contrast (c) and structure (s). Given two signals x and y, the three comparison functions are defined as: $l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$, $c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$, $s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$. Here μ_x, σ_x^2 and σ_{xy} are the mean of x, the variance of x, and the covariance of x and y, respectively. C_1, C_2 and C_3 are constants and stabilize the divisions. Then the general form of the SSIM index between x and y is defined as:

$$SSIM(x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma \tag{4}$$

where α, β and γ are parameters to control the relative importance of the three comparison functions. By setting $C_3 = C_2/2$ and $\alpha = \beta = \gamma = 1$ [25], the formula of the SSIM index can be reduced to the form shown below.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

For image quality evaluation, the SSIM index is typically calculated using a sliding Gaussian window. The window can be displaced pixel-by-pixel on the image to create an SSIM quality map of the image, whose mean defines the SSIM index of two images X and Y:

$$MSSIM(X,Y) = \frac{1}{M}\sum_{i=0}^{M} SSIM(X_i, Y_i) \tag{6}$$

where X_i and Y_i are the image contents of the i^{th} local window; and M is the number of local windows in the image.

In this paper, we consider a SSIM loss of the synthetic person image $\hat{I}_t$ and the ground truth real person image I_t for person image generation, which can be defined as:

$$\mathcal{L}_{SSIM}(I_t, \hat{I}_t) = 1 - MSSIM(I_t, \hat{I}_t) \tag{7}$$

It has several advantages over the losses discussed earlier. Comparing the patch statistics between two images not only makes the SSIM loss robust to local spatial misalignment but also enables it to characterize the texture patterns. Furthermore, since the SSIM index has been widely used to evaluate the generated person images, minimizing the SSIM loss should directly improve the performance.

3.3 Part-based SSIM loss

The SSIM index of two images X and Y defined in Eq. (6) is computed by looking at the neighborhood of each pixel as large as the support of a standard window. We observe that a person in an image can be divided into part regions with homogeneous pixel values or textures. This motivates us to extend the SSIM loss into a novel part-based SSIM loss. By calculating the statistics in those part regions instead of the sliding windows, our new loss explicitly accounts for the articulated body structure.

The person image X is composed of a foreground human body and some background. We decompose the human body into $L = 10$ parts (i.e., head, upper arms, lower arms, upper legs, lower legs and torso) using 18 detected 2D body joints. The head and the torso respectively contain 6 joints and 4 joints. Each of the other parts contains 2 joints. Let $\{M^l : l = 0, \cdots, L\}$ be a set of $L+1$ masks. M^0 and $\{M^l : l = 1, \cdots, L\}$ respectively encode the background region and the L body part regions. For each body part mask M^l, there is a 2D Gaussian filter masking out the l^{th} part according to joints contained in it. Then the masked image can be written as $X^l = X \otimes M^l$, where $\otimes$ is the elementwise multiplication.

Different from the conventional SSIM, which calculates a nonlinear function of pixel mean, variance and covariance, i.e., Eq. (5), in each of the overlapping sliding windows and takes their average, i.e., Eq. (6). The proposed part-based SSIM calculates the nonlinear function of these pixel statistics in each of the nonoverlapping part regions and takes their average. Thus, SSIM compares textures or pixel statistics in each patch while Part-SSIM compares them in each body part region. Since the number of body part regions is small, the computation of the part-based SSIM is efficient. Formally, the SSIM of the l^{th} body part is defined as:

$$SSIM^l = \frac{(2\mu_{X^l}\mu_{Y^l} + C_1)(2\sigma_{X^lY^l} + C_2)}{(\mu_{X^l}^2 + \mu_{Y^l}^2 + C_1)(\sigma_{X^l}^2 + \sigma_{Y^l}^2 + C_2)} \tag{8}$$

where μ_{X^l}, σ_{X^l} and $\sigma_{X^lY^l}$ are statistics calculated in the l^{th} body part region.

Since the background region is unconstrained, we calculate its SSIM index using Eq. (6), denoted by $MSSIM(X^0, Y^0)$, in sliding windows and take their average. For the foreground human body, we calculate the SSIM values in different body part regions using Eq. (8) and take their average. Then the proposed part-based SSIM, denoted by $pSSIM(X,Y)$, can be written as:

$$pSSIM(X,Y) = \frac{1}{2}\left(MSSIM(X^0,Y^0) + \frac{1}{L}\sum_{l=1}^{L} SSIM^l\right) \tag{9}$$

We define the part based SSIM loss of the synthetic person image $\hat{I}_t$ and the ground truth real person image I_t as:

$$\mathcal{L}_{pSSIM}(I_t, \hat{I}_t) = 1 - pSSIM(I_t, \hat{I}_t) \tag{10}$$

We also consider a variant of the part-based SSIM loss. It treats the background region as a part and calculates the structure similarity using the part-based SSIM in Eq. (8). Then the loss function becomes:

$$\mathcal{L}_{pSSIM^*}(I_t, \hat{I}_t) = 1 - \frac{1}{L+1}\sum_{l=0}^{L} SSIM^l \tag{11}$$

4 Experiments

4.1 Implementation details

Baseline model. We use the Pose-Attentional Transfer Network (PATN) proposed by Zhu *et al.* [28] as a baseline model due to its state-of-the-art performance. The PATN can effectively transfer the source person image I_s to the target pose P_t using several cascaded Pose-Attentional Transfer Blocks (PATBs) in the generator G. The PATN consists of an encoder, which takes as input I_s, P_s and P_t, a cascade of 9 PATBs and a decoder to generate the target image $\hat{I}_t$. The loss function is a combination of the L1-norm loss, the perceptual loss and the adversarial loss respectively defined in Eq. (1), Eq. (2) and Eq. (3).

Training Details. We use a Gaussian window with $\sigma = 0.8$ and $2\mu = 7$ to compute the SSIM index of background in the part-based SSIM loss. When different loss functions are combined together to get the final loss, we follow PATN and set the coefficients of L1-norm loss, adversarial loss, and perceptual loss as (10, 5, 10) for both two datasets. We set the coefficient of Part-SSIM loss as 10. This combination of coefficients achieves the overall best performance.

4.2 Datasets and evaluation metrics

Datasets. We conduct experiments on the Market-1501 dataset [27] and the DeepFashion dataset (In-shop Clothes Retrieval Benchmark) [10]. The Market-1501 dataset contains 32,668 images of 1,501 persons captured from six different surveillance cameras. The images have a low resolution (128×64) and vary in human poses, viewpoints, background and

Table 1: Quantitative results of models trained by different loss functions.

Loss model	SSIM	IS	mask-SSIM	mask-IS	DS	pSSIM
L1	0.327	3.431	0.822	3.143	0.365	0.676
Per	0.311	3.339	0.820	3.25	0.480	0.668
GAN	0.192	**3.623**	0.731	**3.723**	**0.628**	0.508
SSIM	**0.352**	2.992	**0.825**	3.365	0.495	**0.678**

illumination, which makes the person image generation more challenging. The DeepFashion dataset contains 52,712 in-shop clothes images with a high resolution (256×256) and clean background. We collect the training and testing splits following [28].

Evaluation metrics. We use the same evaluation metrics as Def-GAN [20], including Structural Similarity (SSIM) [25], Inception Score (IS) [19], masked version of Structural Similarity (mask-SSIM) [11], masked version of Inception Score (mask-IS) [11] and Detection Score (DS) [20]. IS computes the classification score of generated images using the Inception Net [23] trained on ImageNet [18]. Mask-SSIM and Mask-IS are respectively the masked versions of SSIM and IS. Note that the masks have been built following the procedure proposed in [26]. DS computes the person-class detection scores using SSD [9] on each generated image, which measures the sharpness of an image to some extent. We additionally use the proposed Part-SSIM (pSSIM) as a new evaluation metric to measure the generated image quality.

4.3 Ablation study

Loss analysis. In order to analysis the impact of different losses on the person image generation task, we train different image generation models on Market-1501 by using different losses separately (L1 loss, perceptual loss,adversarial loss, and SSIM loss). Tab. 1 shows that the model trained by the adversarial loss (GAN) has the highest IS scores and DS score but performs worst under the SSIM metric. The model trained by the SSIM loss (SSIM) has the highest SSIM scores but the lowest IS score. Models trained by the L1 loss (L1) and the perceptual loss (Per) have moderate performance in all metrics. We also present some qualitative results in Fig. 2. The adversarial loss often produces distortions in the generated images but keeps their sharpness. The rest three losses often result in blurred images but the images generated via the SSIM loss look better than those generated by the L1 loss model and the perceptual loss model. Note that the IS metric and DS metric respectively use a classification model and a detection model trained by a deep neural network based on an entropy loss, while the discriminator used in the adversarial loss is also trained via an entropy loss. Thus, the model trained by the adversarial loss is more likely to capture the semantic information which are easily recognized by a classification model or a detection model. This explains why the model trained by the adversarial loss has the highest IS score and DS score. For the same reason, the model trained by the SSIM loss can achieve the highest SSIM score.

Then we combine different loss functions for further study. Considering the strengths and weaknesses of different loss functions analyzed above, we combine the adversarial loss with each of the other three loss functions to train person image generation models, denoted by GAN+L1, GAN+Per and GAN+SSIM, respectively. Quantitative results in Tab. 2 indicate that DS scores of all models improve greatly, which means the generated images are sharper. The model trained by GAN+L1 has higher IS scores. The model trained by GAN+Per has a higher DS score. The model trained by GAN+SSIM has higher SSIM scores.

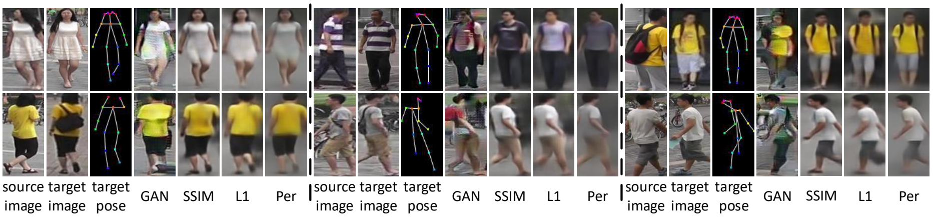

Figure 2: Qualitative results of models trained by different loss functions.

Table 2: Quantitative comparison of different models trained by three loss combinations.

Loss model	SSIM	IS	mask-SSIM	mask-IS	DS	pSSIM
GAN+L1	0.275	**3.616**	0.794	**3.898**	0.731	0.611
GAN+Per	0.299	3.313	0.804	3.791	**0.761**	0.619
GAN+SSIM	**0.308**	3.410	**0.807**	3.723	0.713	**0.626**

Part-based SSIM loss analysis. We make a series of loss function combinations to verify the effectiveness of the proposed part-based SSIM loss and all models are trained on Market-1501. Tab. 3 shows the quantitative comparison. We first compare the models trained by GAN+L1 and GAN+L1+pSSIM. Results show that the model with the part-based SSIM loss has a higher SSIM score but a lower IS score. We also compare GAN+Per and GAN+Per+pSSIM. Results show that both the SSIM and IS metrics improve when we train the model with the part-based SSIM loss. However, when the model is trained via the combination of the adversarial loss, perceptual loss, part-based SSIM loss and L1-norm loss, the performance degrades under most metrics. The quantitative comparison also shows that the model trained with the part-based SSIM loss always performs better than that without the part-based SSIM loss. The model trained with the perceptual loss always performs better than the model trained with the L1-norm loss. Finally, the comprehensive comparison shows that the model trained by the adversarial loss, perceptual loss and part-based SSIM loss is the best model because it has comparable quantitative results under all evaluation metrics.

We also compare the performance of models trained with $\mathcal{L}_{pSSIM}$, $\mathcal{L}_{SSIM}$ and $\mathcal{L}_{pSSIM^*}$. Quantitative results in Tab. 4 show that the model trained with the proposed part-based SSIM loss has the best performance under most evaluation metrics because it helps the model capture the articulated body structure. We also make a qualitative comparison in Fig. 3. Results show that the model trained with the part-based SSIM loss produces more realistic person images than the model trained with the other two losses.

Table 3: Quantitative comparison of models trained by different loss combinations with or without part-based SSIM loss.

Model	SSIM	IS	mask-SSIM	mask-IS	DS	pSSIM
GAN+L1	0.275	3.616	0.794	3.898	0.731	0.611
GAN+l1+pSSIM	0.291	3.454	0.799	3.801	0.718	0.630
GAN+Per	0.299	3.313	0.804	3.791	0.761	0.619
GAN+Per+pSSIM	**0.312**	**3.326**	**0.810**	**3.807**	**0.742**	**0.642**
GAN+Per+l1+pSSIM	0.306	3.338	0.807	3.779	0.733	0.633

Table 4: Quantitative results of the models trained with three types of SSIM-based losses.

Model	SSIM	IS	mask-SSIM	mask-IS	DS	pSSIM
GAN+Per+SSIM	0.307	3.306	0.808	3.756	0.776	0.627
GAN+Per+pSSIM*	0.303	3.376	0.807	3.786	0.760	0.636
GAN+Per+pSSIM	**0.312**	**3.326**	**0.810**	**3.807**	**0.742**	**0.642**

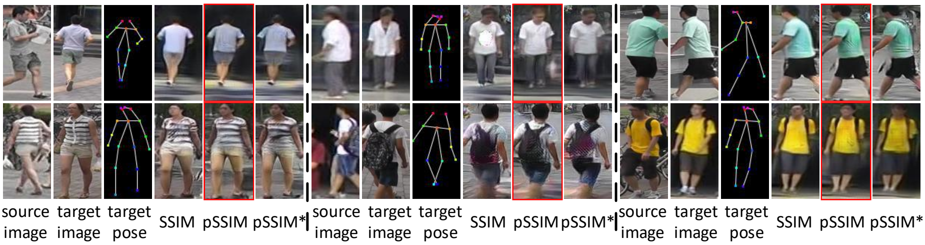

Figure 3: Qualitative results of models trained with different types of SSIM-based losses.

4.4 Comparison with state-of-the-art methods

We give qualitative and quantitative comparisons in this part to verify the effectiveness of our proposed part-based SSIM loss.

Quantitative results. Quantitative results on Market-1501 [27] and DeepFashion [10] are shown in Tab. 5. Several state-of-the-art methods including PG^2 [11], Def-GAN [20] and PATN [28] [1] are compared with our best model trained by the combination of the adversarial loss, part-based SSIM loss and perceptual loss. Considering both datasets, our method is comparable with PG2 and Def-GAN under IS but outperforms them under all other metrics. Tab. 5 shows the IS of PATN is much lower than PG2 or Def-GAN, and after adding the Part-SSIM loss, the IS improves on both datasets, which demonstrates the effectiveness of the proposed method. While our DS is lower than PATN on Market1501 but higher on DeepFashion, both their scores are higher than real data, which means the images generated by both methods are as good as real images under this metric. In addition, our method outperforms PATN under all other metrics.

Table 5: Quantitative comparison with state-of-the-art methods.

Model	**Market-1501**						**DeepFashion**			
	SSIM	IS	mask-SSIM	mask-IS	DS	pSSIM	SSIM	IS	DS	pSSIM
PG2	0.261	**3.495**	0.782	3.367	0.390	-	0.773	3.163	0.951	-
Def-GAN	0.291	3.230	0.807	3.502	0.720	-	0.760	**3.362**	0.976	-
PATN (baseline)	0.281	3.162	0.799	3.737	**0.796**	0.6186	0.771	3.201	0.976	0.799
Ours	**0.312**	3.326	**0.810**	**3.807**	0.742	**0.6415**	**0.776**	3.262	**0.982**	**0.813**
Real Data	1.000	3.890	1.000	3.706	0.740	1	1.000	4.053	0.968	1

Qualitative results. We further visualize some typical qualitative examples of our method and PATN [28] in Fig. 4. Qualitative comparison indicates that our part-based SSIM loss not only keeps the person structure of synthetic images but also preserves the detailed texture to some extent. The right eight examples in the blue rectangle are performed on the Market-1501. We can clearly see that our model produces synthetic images with sharper borders of

[1] We reproduce the results of PATN using the code provided by the authors.

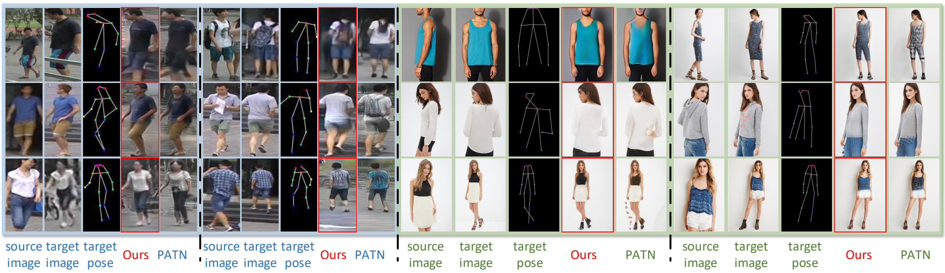

Figure 4: Qualitative comparison with state-of-the-art methods.

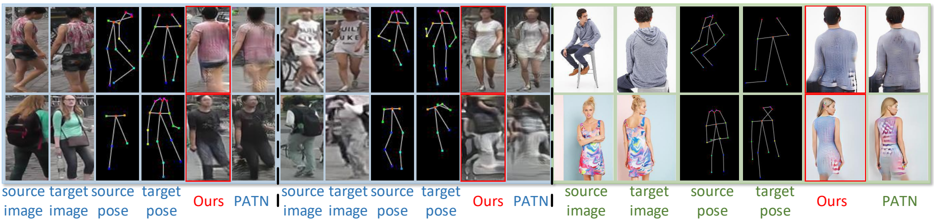

Figure 5: Failure cases on Market-1501 and DeepFashion.

persons and preserves more detailed texture of the human body. For the eight typical examples obtained on the DeepFashion shown in the green rectangle, our model produces more realistic images with the accurate person postures, clean background and detailed texture. We also provide some failure cases in Fig. 5. As we can see, our method generates low quality images with coarse texture and blurred postures when the source image has complicated texture or large pose transfer. Our code and trained models are publicly available [2].

5 Conclusion

In this paper, we first make a comprehensive study of the loss functions (e.g., L1-norm loss, adversarial loss, and perceptual loss) for pose-guided person image generation. We also propose a novel part-based SSIM loss function to account for the unique appearance and structure patterns of person. Experimental results demonstrate the effectiveness of the proposed approach.

Acknowledgment

This work was supported partly by National Key Research and Development Program of China Grant 2018AAA0101400, NSFC Grants 61629301, 61773312, and 61976171, China Postdoctoral Science Foundation Grant 2019M653642, Young Elite Scientists Sponsorship Program by CAST Grant 2018QNRC001, and Natural Science Foundation of Shaanxi Grant 2020JQ-069.

[2] https://github.com/shyern/Pose-Transfer-pSSIM.git

References

[1] Dominique Brunet, Edward R Vrscay, and Zhou Wang. Structural similarity-based approximation of signals and images using orthogonal bases. In *International Conference Image Analysis and Recognition*, pages 11–22. Springer, 2010.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[3] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018.

[4] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[7] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[8] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019.

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[10] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[11] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017.

[12] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[13] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[14] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8620–8628, 2018.

[15] Abdul Rehman, Mohammad Rostami, Zhou Wang, Dominique Brunet, and Edward R Vrscay. Ssim-inspired image restoration using sparse representation. *EURASIP Journal on Advances in Signal Processing*, 2012(1):16, 2012.

[16] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. *arXiv preprint arXiv:2003.00696*, 2020.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115 (3):211–252, 2015.

[19] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[20] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2357–2366, 2019.

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[24] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.

[25] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[26] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[27] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[28] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.