MsCGAN: Multi-scale Conditional Generative Adversarial Networks for Person Image Generation

Wei Tang¹, Gui Li¹, Xinyuan Bao¹, Fudong Nian², Teng Li¹
¹ Anhui University, Hefei, 230601, China
² Hefei University, Hefei, 230601, China
E-mail: li teng@ahu.edu.cn

Abstract: To synthesize high-quality person images with arbitrary poses is challenging. In this paper, we propose a novel Multi-scale Conditional Generative Adversarial Networks (MsCGAN), aiming to convert the input conditional person image to a synthetic image of any given target pose, whose appearance and the texture are consistent with the input image. MsCGAN is a multi-scale adversarial network consisting of two generators and two discriminators. One generator transforms the conditional person image into a coarse image of the target pose globally, and the other is to enhance the detailed quality of the synthetic person image through a local reinforcement network. The outputs of the two generators are then merged into a synthetic, discriminant and high-resolution image. On the other hand, the synthetic image is downsampled to multiple resolutions as the input to multi-scale discriminator networks. The proposed multi-scale generators and discriminators handling different levels of visual features can benefit to synthesizing high-resolution person images with realistic appearance and texture. Experiments are conducted on the Market-1501 and DeepFashion datasets to evaluate the proposed model, and both qualitative and quantitative results demonstrate the superior performance of the proposed MsCGAN.

Key Words: person image generation; multi-scale discriminators; generative adversarial networks; image synthesis

1 INTRODUCTION

To generate high-resolution and photo-realistic person image with arbitrary poses is important in multimedia and computer vision for its wide range of applications such as data augmentation in person re-ID[1], person image editing or inpainting[2] and video forecasting[3]. Various methods have been proposed to handle this problem, such as VAE[4], ARMs[5], and GANs[6]. Owing to impressive performances of GANs, recent works on person image generation[2, 7–10] and high-resolution image synthesis[11–13] mostly focus on it.

From the perspective of the prior knowledge, existing person image generation methods based on GANs can be categorized into two groups. The first is the global pose-guided strategy [7–9, 14] and the second is the segmentation-guided multi-module strategy [10, 15]. The former is to synthesize the target person image through a global model by inputting the conditional person image and the human pose estimation simultaneously. The latter is to parse the conditional person image into foreground, background and pose information, then synthesize the target image using multi-module models. Although the pose-guided person image generation methods can be accurate regarding target human poses, they usually neglect the detailed appearance and the texture information of the conditional person images. Segmentation-guided multi-module person image generation techniques can preserve appearance and texture features of the conditional person image more completely. However, most multi-module models are complex and difficult to train. In many cases, above-mentioned methods are still problematic to produce person image with precise pose and preserve appearance details simultaneously.

In this paper, we propose a novel Multi-scale Conditional Generative Adversarial Networks (MsCGAN) for person image generation. In Figure 1, the inputs are the conditional image and the target pose. Our goal is to synthesize the person image with the target pose, whose appearance and the texture information are also consistent with the conditional images. MsCGAN is based on pose guidance so that it can synthesize the person images with high pose accuracy. Meanwhile, MsCGAN contains two
strategies to ensure the visual quality of the generated images to be consistent with the conditional image. One of them is using the global-to-local generators, which generate a coarse image of the specific pose globally, and refine the coarse image locally. The other is that multi-scale discriminators are adopted to discriminate the generated image and its downsampled images respectively, which aim to handle the visual features on multiple levels.

Compared with existing methods, our model has several advantages: 1) Joint the global generation and the local refinement can improve both the accuracy and quality of the synthetic image. 2) Images with different resolutions contain different levels of visual features, so the proposed multi-scale discriminators can increase the receptive field of the discriminator. 3) The combination of the global-to-local generators and multi-scale discriminators ensure the synthetic person images have the target pose and more detailed appearance features than existing methods.

Some examples are shown in Figure 1. The proposed model is evaluated on the dataset Market-1501[16], as well as the fashion dataset DeepFashion[17]. Moreover, we compare our results to the results of the state-of-the-art methods to demonstrate effectiveness. The main contributions of our work are summarized as follows:

- A novel MsCGAN model are proposed for person pose image generation, which is able to handle more detailed appearance and texture information while synthesizing high-quality and realistic person images with arbitrary poses.
- The proposed global-to-local generating strategy, as well as multi-scale discriminators, can be widely used in GANs based image or video synthesis.
- The proposed MsCGAN is a two-stage model, which is more compact than segmentation-guided multimodule methods, and it achieves much superior performance than existing methods on the benchmark Market-1501 and DeepFashion datasets.

The rest of the paper is organized as follows. Section 2 introduces the details of the proposed MsCGAN. In Section 3, we report and analyze extensive experimental results. Finally, we conclude the paper with future work in Section 4.

2 METHODOLOGY

Table 1 lists the key notations for the proposed MsCGAN which will be used in the following parts. The architecture of MsCGAN is shown in Figure 2. In summary, the global-to-local generators consist of two generators that integrate the coarse image and enhance the refined images respectively. To obtain high-level features and the texture information of the conditional person image, we apply the multi-scale discriminators discriminate images with different resolutions. Overall, MsCGAN contains two stages that will be detailed respectively as follows:

2.1 Stage-I: Global Generation

At stage-I, we input the conditional person image ($x$) and the human pose estimation of the target person image ($p$) to synthesize a global structure image ($\hat{y}_1$) containing the pose of the target person image ($y$).

2.1.1 Human Pose Estimation

Here we use the Part Affinity Fields (PAFs)[18] to obtain a highly accurate map of the human pose estimation. In our work, the PAFs estimate the human pose in theconditional person image and generates the coordinates of eighteen 2D keypoints in real-time. Then we encode the eighteen 2D keypoints into eighteen heatmaps to express the human pose more intuitively.

2.1.2 Generator $G_1$: Full Person Image Generator

Generator $G_1$ adopts a variant of the U-Net network [19] and the residual blocks [20]. The input of $G_1$ is a concatenation of the conditional person image ($x$) and the map of human pose estimation generated by the target person image ($p$), while the output of $G_1$ is a coarse image ($\hat{y}_1$) with the pose of the target person image. The poses of the $x$ and the ($\hat{y}_1$) are quite different, but they are closely related to the appearance feature of the person. To change the pose of the person and preserve the detailed appearance of the person simultaneously, we employ skip connection between the symmetric layers of encoder and decoder in the variant of U-Net network. Since the generated image using $L_2$ distance is more blurred than that using $L_1$ distance in generator[21], we use $L_1$ distance as its basic loss function.

2.1.3 Compensation for the Impact of Background

The generators originally have no idea what the background of generates image should be like, which would cause the results unrealistic. Furthermore, when the input image is unsharp, the person and the background are prone to jumble. In order to mitigate the impact of background, we introduce the background loss that uses the pose mask of the target person image ($P_m$) to separate the person when calculating $L_1$ loss. In the $P_m$, 1 indicates the person and 0 indicates the background.

$$L_{Stage1} = L_{L_1} + \lambda_{bg} L_{bg,1}$$

$$= \| y - G_1(x, p) \|_1 + \lambda_{bg,1} \| (y - G_1(x, p)) \odot P_m \|_1.$$ (1)

here $\| y - G_1(x, p) \|_1$ and $\| y - G_1(x, p) \odot P_m \|_1$ indicate $L_1$ loss and background loss, $\lambda_{bg,1}$ is a hyper-parameter.
that indicates the weight of background loss, and ⊙ denotes
the pixels-wise multiplication.

2.2 Stage-II: Local Reinforcement

At stage-II, we correct the detailed feature of \( \hat{y}_1 \) and dis-

criminate against the synthetic image. We design a local
reinforcement generator \( G_2 \) to generate an appearance dif-
ference map (\( \hat{y}_2 \)) between \( x \) and \( \hat{y}_1 \). Then, \( \hat{y}_1 \) and \( \hat{y}_2 \) are

merged into the final synthetic image (\( \hat{y} \)). Multi-scale dis-

criminators discriminate against the synthetic images with

different resolutions. The conditional DCGAN-like net-

work are adopted as our base model.

2.2.1 Generator \( G_2 \): Local Reinforcement Generator

Local reinforcement generator \( G_2 \) is a network inspired by

conditional DCGAN and U-Net architectures. Specifically,

the encoder of \( G_2 \) is a variant of the conditional DCGAN

network [22] and is symmetric to the decoder of \( G_2 \). The

inputs of \( G_2 \) are the conditional person image (\( x \)) and the

stage-I generation result (\( \hat{y}_1 \)), while the output of \( G_2 \) is an

appearance difference map (\( \hat{y}_2 \)) between \( x \) and \( \hat{y}_1 \). The

fully connected layer destroys the spatial structure of the

image while compressing it. The DCGAN network is a

full-convolution architecture so that the details of the im-

age can be preserved more completely.

2.2.2 Multi-scale Discriminators

In order to generate a high-resolution image, it is neces-
sary to increase the receptive field of the discriminator. We

adopt multi-scale discriminators which consist of two dis-

criminators (\( D_1 \) and \( D_2 \)) that discriminate different resolu-

tion versions of the synthetic image. We downsample the

conditional person image (\( x \)), the final synthetic image (\( \hat{y} \))

and the target person image (\( y \)) by a factor of 2. The dis-

criminators \( D_1 \) and \( D_2 \) are trained to distinguish real and

synthesized images at the two different scales, respectively.

2.3 Overall Loss Function for Stage-II

At stage-II, we use conditional GAN as our basic loss func-
tion and add \( L_1 \) loss, background loss which introduced in

stage-I as extra supervisory signals. The loss function of

the conditional GAN is formulated as:

\[
\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(z)))],
\]

where \( z \) refers to \( G_1(x, p) \). \( G \) tries to minimize this ob-

jective against an adversarial \( D \) which tries to maximize it.

Therefore, the loss function of the conditional GAN can be

optimized as follows:

\[
\min_{G_2} \max_{D_1, D_2} (\lambda_{D_1} \mathcal{L}_{cGAN}(G_2, D_1) + \lambda_{D_2} \mathcal{L}_{cGAN}(G_2, D_2)),
\]

Figure 2: The architecture of the proposed MsCGAN.
where $\lambda_{D_1}$ and $\lambda_{D_2}$ control the importance of the two terms, respectively.

Thus, our full loss combining conditional GAN loss, $L_1$ loss and background loss for Stage-II is formulated as:

$$L_{Stage,2} = \min_{G_2} \max_{D_1,D_2} (\lambda_{D_1}L_{cGAN}(G_2,D_1) + \lambda_{D_2}L_{cGAN}(G_2,D_2)) + \lambda_{L1,2} \parallel y - G_2(x,G_1(x,p)) \parallel_1 + \lambda_{bg,2} \parallel y - G_2(x,G_1(x,p)) \circ p_m \parallel_1),$$

where $\parallel y - G_2(x,G_1(x,p)) \parallel_1$ is $L_1$ loss and $\parallel y - G_2(x,G_1(x,p)) \circ p_m \parallel_1$ is background loss respectively, $\lambda_{L1,2}$ and $\lambda_{bg,2}$ are the weights of them.

3 EXPERIMENTS

This section demonstrates the effectiveness of the proposed MsCGAN. Firstly, the datasets and implementation details are introduced. Then, we show the qualitative and quantitative evaluations of the proposed MsCGAN.

3.1 Datasets and Implementation Details

We conduct evaluations on two public datasets: Market-1501 dataset [16] and DeepFashion dataset [17]. Our MsCGAN is implemented with Tensorflow [23] and we apply the Adam optimizer [24] with adapted hyper-parameters. As suggested in [6], we improve the gradient by training to maximize rather than minimize in the implementation.

3.2 Baselines

To demonstrate the proposed MsCGAN can achieve superior results, we compare our method with several state-of-the-art works: [7, 8, 10, 25–27]. Moreover, to demonstrate the proposed network structure proposed in Section 2 is optimal, we evaluate qualitative and quantitative results of three settings: One discriminator, Two discriminators (MsCGAN) and Three discriminators. All of them have the same global-to-local generators while the number of discriminators are different. Note that we use the same generators, the same datasets and the same hyper-parameters to train and test all baselines mentioned above.

3.3 Results and Discussions

3.3.1 Qualitative Results

We apply global-to-local generators and multi-scale discriminators consisting of two discriminators. To illustrate the difference between multi-scale discriminators, we perform different number of discriminators for qualitative comparisons. Moreover, we compare our results with six different methods [7, 8, 10, 25–27]. The results for different models are shown in Figure 3. It is clear that MsCGAN can generate more photo-realistic results than other methods. The major reason is that our MsCGAN plays more attention on the appearance and the texture information of the conditional person image.

3.3.2 Quantitative Results

We adopt the evaluation protocol from previous person image generation works [7–10, 28, 29]. We use Structural Similarity Index Measure (SSIM) [30] and Inception Score (IS) [31] and their corresponding masked versions mask-SSIM and mask-IS [7]. It is objective to evaluate the IS on a large enough number of samples [31].

Figure 3: Example of results on Market-1501 dataset and DeepFashion dataset.
Table 2: Quantitative results on two public datasets, higher scores are better.

| Model                   | DeepFashion |       |       | Market-1501 |       |       |
|-------------------------|-------------|-------|-------|-------------|-------|-------|
|                         | SSIM        | IS    |       | SSIM        | IS    | mask-SSIM |
| One discriminator       | 0.737       | 3.212 |       | 0.248       | 3.167 | 0.791   |
| MsCGAN                  | 0.725       | 3.335 |       | 0.219       | 3.588 | 0.764   |
| Three discriminators    | 0.691       | 3.284 |       | 0.216       | 3.459 | 0.778   |
| Ma et al.[7]            | 0.762       | 3.090 |       | 0.253       | 3.460 | 0.792   |
| Ma et al.[10]           | 0.614       | 3.228 |       | 0.099       | 3.483 | 0.614   |
| Pumarola et al.[25]     | 0.747       | 2.97  |       | —           | —     | —      |
| Siarohin et al.[8]      | 0.756       |       | 3.439 | 0.290       | 3.185 | 0.805   |
| Esser et al.[26]        | 0.786       |       | 3.087 | 0.353       | 3.214 | —       |
| Dong et al.[27]         | 0.793       | 3.314 |       | 0.356       | 3.409 | —       |

Figure 4: A few typical failure cases.

results are shown in Table 2. From Table 2, it is clear that MsCGAN achieves the highest or high score in IS and mask-IS criteria. There is still a gap between MsCGAN and the state-of-the-art approach in SSIM and mask-SSIM criteria.

3.3.3 Failure Cases

Although MsCGAN can synthesize more photo-realistic results than other methods, it still has failure cases. We select a few typical failure cases that are shown in Figure 4. In row 1, we can see that objects in the conditional person image have a great influence on the human pose estimation so the proposed MsCGAN synthesizes poor-quality result. From the result in row 2, the occlusion of the person in the conditional person image causes the appearance of the synthetic image to be disorganized.

In row 3, we can find that MsCGAN causes an error of gender which transforms a female into a male. In row 4, the body in the target person image does not appear in the conditional person image, resulting in an incomplete body of the generated image. There are two main reasons why image generation fails. The first one is that the conditional person image is too blurred and there are else objects in it, which affects the detection of keypoints on the human body. The second one is that the conditional person image is occluded, which affects the appearance features and texture information of the synthetic image.

4 CONCLUSION

In this paper, we propose a novel multi-scale conditional generative adversarial networks (MsCGAN) for generating high-quality and photo-realistic person image with arbitrary poses. The proposed MsCGAN can simultaneously model target pose as well as the appearance of the conditional person image in a unified and principled way. In the experiments, the qualitative and quantitative results clearly demonstrate the effectiveness of the proposed model. In the future, we will evaluate our model on more practical datasets and investigate more applications.

REFERENCES

[1] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in IEEE International Conference on Computer Vision, 2017.

[2] C. Si, W. Wang, L. Wang, and T. Tan, “Multistage adversarial losses for pose-based human image synthesis,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[3] J. Walker, K. Marino, A. Gupta, and M. Hebert, “The pose knows: Video forecasting by generating pose futures,” in IEEE International Conference on Computer Vision, 2017.

[4] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” arXiv preprint arXiv:1312.6114, 2013.
