Disentangled Person Image Generation

Wenjie Niu

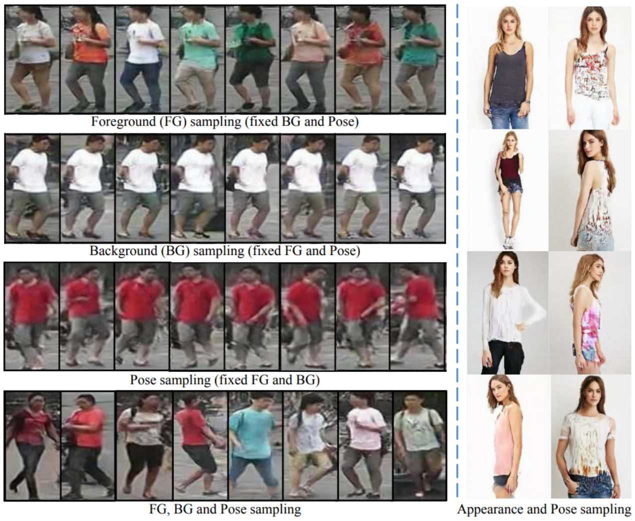
June 24,2018

Abstract

In fact, generating novel, realistic images of person is a complex challenging due to different image factors, such as the foreground, background and pose information. The paper aims at generating such iamges based on a novel, two-stage reconstruction pipeline that learns a disentangled representation of the aforementioned image factors and generates novel person images at the same time. First, a multi-branched reconstruction network is proposed to disentangle and encode the three factors into embedding features, which are then combined to recompose the input image itself. Second, three corresponding mapping functions are learned in an adversarial manner in order to map Gaussian noise to the learned embedding feature space, for each factor, respectively. Using the proposed framework, we can manipulate the foreground, background and pose of the input image, and also sample new embedding features to generate such targeted manipulations, that provide more control over the generation process. Experiments on the Market-1501 and Deepfashion datasets show that our model does not only generate realistic person images with new foregrounds, backgrounds and poses, but also manipulates the generated factors and interpolates the in-between states. Another set of experiments on Market-1501 shows that our model can also be beneficial for the person re-identification task. [5]

1. Introduction

The process of generating realistic-looking images of persons has several applications, like image editing, person re-identification (re-ID), inpainting or on-demand generated art for movie production. The recent advent of image generation models, such as variational autoencoders (VAE) [2], generative adversarial networks (GANs) [3] and autoregressive models (ARMs) (*e.g.* PixelRNN [7]), has provided powerful tools towards this goal. Several papers [8],[1],[6] have then exploited the ability of these networks to generate sharp images in order to synthesize realistic photos of faces and natural scenes. Recently, Ma *et al.* [4] proposed an architecture to synthesize novel person images in arbitrary poses given as input an image of that person and a new pose. From an application perspective however, the user often wants to have more control over the generated images (*e.g.* change the background, a persons appearance and clothing, or the viewpoint), which is something that existing methods are essentially uncapable of. We go beyond these constraints and investigate how to generate novel person images with a specific user intention in mind (*i.e.* foreground (FG), background (BG), pose manipulation). The key idea is to explicitly guide the generation process by an appropriate representation of that intention. Fig. 1 gives examples of the intended generated images.



Foreground (FG) sampling (fixed BG and Pose)

Background (BG) sampling (fixed FG and Pose)

Pose sampling (fixed FG and BG)

FG, BG and Pose sampling

Appearance and Pose sampling

Figure 1. Left: image sampling results on Market-1501. Three factors, *i.e.* foreground, background and pose, can be sampled independently (1st-3rd rows) and jointly (4th row). Right: similar joint sampling results on DeepFashion.This dataset contains almost no background, so we only disentangle the image into appearance and pose factors. [5]

References

[1] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learn-

ing by information maximizing generative adversarial nets. In *International Conference on Neural Information Processing Systems*, 2016. 1

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Auto-encoding variational bayes. *arXiv preprint arXiv:arXiv:1312.6114*, 2013. 1

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, 2014. 1

[4] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *International Conference on Neural Information Processing Systems*, 2017. 1

[5] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz. Disentangled person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[6] L. B. Martin Arjovsky, Soumith Chintala. Wasserstein GAN. In *International Conference on Learning Representations*, 2017. 1

[7] A. V. D. Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, 2016. 1

[8] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2015. 1