Disentangled Person Image Generation

Xuewen Yang

July 22 2018

Abstract

Generating novel, yet realistic, images of persons is a challenging task due to the complex interplay between the different image factors, such as the foreground, background and pose information. In this work,the author aim at generat- ing such images based on a novel, two-stage reconstruc- tion pipeline that learns a disentangled representation of the aforementioned image factors and generates novel per- son images at the same time.First, a multi-branched recon- struction network is proposed to disentangle and encode the three factors into embedding features, which are then com- bined to re-compose the input image itself. Second, three corresponding mapping functions are learned in an adver- sarial manner in order to map Gaussian noise to the learned embedding feature space, for each factor, respectively.

1. Introduction

The author disentangle the input image into intermediate embedding features, i.e. person images can be reduced to a composition of features of foreground, background, and pose. Compared to existing approaches,they rely on a different technique to generate new samples. In particular,they aim at sampling from a standard distribution, e.g. a Gaussian distribution, to first generate new embedding features and from them generate new images.

2. Method

The goal is to disentangle the appearance and structure factors in person images, so that they can manipulate the foreground, background and pose separately. To achieve this, the author propose a two-stage pipeline shown in Figure 1. In stage-I, they disentangle the foreground, background and pose factors using a reconstruction network in a divide-and-conquer manner. In particular,they reconstruct person images by first disentangling into intermediate embedding features of the three factors, then recover the input image by decoding these features. In stage-II,they treat these features as real to learn mapping functions for mapping a Gaussian distribution to the embedding feature distribution adversarially.

2.1. StageI:Disentangled image reconstruction

At stage-I, they propose a multi-branched reconstruction architecture to disentangle the foreground, background and pose factors.

To separate the foreground and background information, they apply the coarse pose mask to the feature maps instead of the input image directly. By doing so, they can alleviate the inaccuracies of the coarse pose mask. Then, in order to further disentangle the foreground from the pose information, they encode pose invariant features with 7 Body Regions-Of-Interest instead of the whole image similar to [3].

For the background branch,the author apply the inverse pose mask to get the background feature maps and pass them into the background encoder to obtain a 128-dim embedding feature. Then, the foreground and background features are concatenated and tiled into $128 \times 64 \times 352$ appearance feature maps.

For the pose branch,they concatenate the 18- channel heatmaps with the appearance feature maps and pass them into the a U-Net-based architecture [2], i.e., convolutional autoencoder with skip connections, to generate the final person image following $PG^2(G1 + D)$ [1].

2.2. StageII:Embedding feature mapping

They propose a two-step mapping technique as illustrated in Figure 1. Instead of directly learning to decode Gaussian noise to the image space, they first learn a mapping function Φ that maps a Gaussian space Z into a continuous feature embedding space E, and then use the pretrained decoder to map the feature embedding space E into the real image space X. The encoder learned in stage-I encodes the FG, BG and Pose factors x into low-dimensional real embedding features e. Then, they treat the features mapped from Gaussian noise z as fake embedding features and learn the mapping function Φ adversarially. In this way, we can sample fake embedding features from noise and

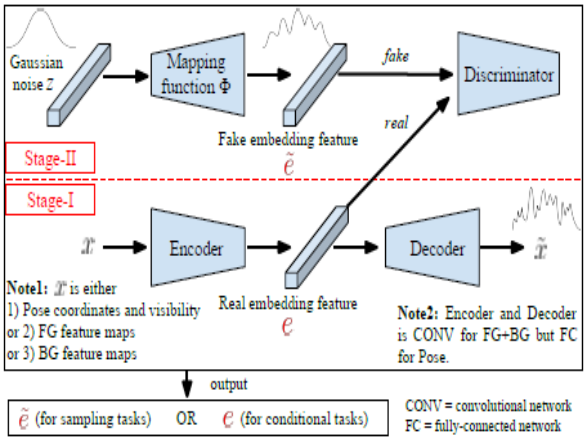

Figure 1. The two-stage framework.

then map them back to images using the decoder learned in stage-I. The proposed two-step mapping technique is easy to train in a piecewise style and most importantly can be useful for other image generation applications.

References

[1] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In *NIPS*, 2017. 1

[2] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[3] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *CVPR*, 2017. 1