Learning Realistic Human Reposing using Cyclic Self-Supervision with 3D Shape, Pose, and Appearance Consistency

Soubhik Sanyal[2] Alex Vorobiov[1] Timo Bolkart[2] Matt Looper[1]
Betty Mohler[1] Larry Davis[1] Javier Romero[1] Michael Black[1]
[1]Amazon [2]Max Planck Institute for Intelligent Systems, Tübingen
{ssanyal,tbolkart}@tue.mpg.de {vorobioo,mattl,bettym,larryd,javier,mjblack}@amazon.com

1. FID and LPIPS evaluation

We pad all generated images to size 256x256 with white border. Reference images are obtained by resizing the original images from DeepFashion dataset to height of 256 and then padding them to size 256x256 with white border. We use PyTorch implementations of FID [8] and LPIPS [10] with AlexNet as feature extractor. The FID is calculated using training images as reference distribution, where generated images are generated using the test split from Ren et al. [7] and Zhu et al. [11].

2. FID for different JPEG quality levels

Common practice to calculate metrics on generated images is to save the images on disk in JPEG format as an intermediate step. We noticed that this affects the FID calculation significantly, as shown in Table 1. The FID increases when it is calculated on image distributions with different levels of JPEG quality and decreases if it is calculated on higher levels of JPEG compression applied to both distributions.

GEN \ REF	80	90	95	RAW
80	6.9	7.3	8.1	12.1
90	7.5	7.1	7.4	10.6
95	8.7	7.8	7.4	9.6
RAW	12.4	10.4	9.1	**7.8**

Table 1: FID as a function of the JPEG quality level for generated (GEN) and reference images (REF).

3. Additional metrics

To assess the similarity between the source and generated image and target and generated image we calculate CX scores [6]. This score measures the cosine similarity between deep features extracted using VGG19 model between two not aligned images. We used the original implementation of [5].

Another important metric is the distance between the target pose and pose on the generated image, which can be evaluated using object keypoint similarity (OKS) [1]. We used OpenPose [2] to extract the keypoints from target and from generated images.

Additional metrics are shown in Table 2. SPICE outperforms other unsupervised methods on both CX scores and OKS.

DeepFashion	CX-GS(↑)	CX-GT(↑)	OKS(↑)
VU-Net [3]	0.182	0.245	0.93
DPIG [4]	0.164	0.197	0.86
PGSPT [9]	0.169	0.222	0.90
SPICE (Ours)	**0.236**	**0.311**	**0.94**

Table 2: Additional quantitative comparison of our method with other unpaired state-of-the-art methods.

References

[1] Object keypoints similarity. https://cocodataset.org/#keypoints-eval. 1
[2] 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 (PAMI)*, 2019. 1
[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 (CVPR)*, pages 8857–8866, 2018. 1
[4] 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 (CVPR)*, pages 99–108, 2018. 1
[5] Yifang Men. CX score source code. https://github.com/menyifang/ADGAN. 1
[6] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma,

and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5084–5093, 2020. 1

[7] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[8] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.1.1. 1

[9] 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 (CVPR)*, 2019. 1

[10] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. lpips: LPIPS metric for PyTorch. https://github.com/richzhang/PerceptualSimilarity. 1

[11] 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 (CVPR)*, pages 2347–2356, 2019. 1