Contents

1 Disentangling Generative Factors

Disentangled representations of object shape and appearance allow to alter both properties individually to synthesize new images. The ability to flexibly control the generator allows, for instance, to change the pose of a person or their clothing. In contrast to previous work [1, 2, 3, 4, 5, 6], we achieve this ability without requiring supervision *and* using a flexible part-based model instead of a holistic representation. This allows to explicitly control the parts of an object that are to be altered. We quantitatively compare against *supervised* state-of-the-art disentangled synthesis of human figures. Also we qualitatively evaluate our model on unsupervised synthesis of still images, video-to-video translation, and local editing for appearance transfer.

1.1 Disentangling Pose and Appearance

Deep Fashion [7, 8] consists of ca. 53k in-shop clothes images in high-resolution of 256×256. We selected the images which are showing a full body (all keypoints visible, measured with the pose estimator by [9]) and used the provided train-test split. For comparison with Esser *et al*. [1] we used their published code.

On Deep Fashion [7, 8], a benchmark dataset for supervised disentangling methods, the task is to separate person ID (appearance) from body pose (shape) and then synthesize new images for previously unseen persons from the test set in eight different poses. We randomly sample the target pose and appearance conditioning from the test set. Fig. 1.1 shows qualitative results. We quantitatively compare against supervised state-of-the-art disentangling [1] by evaluating *i)* invariance of appearance against variation in shape by the re-identification error and *ii)* invariance of shape against variation in appearance by the distance in pose between generated and pose target image.

1.1.1 Person Re-Identification

To evaluate appearance we fine-tune an ImageNet-pretrained [10] Inception-Net [11] with a re-identification (ReID) algorithm [12] via a triplet loss [13] to the Deep Fashion training

Table 1.1: Mean average precision (mAP) and rank-n accuracy for person re-identification on synthesized images after performing shape/appearance swap. Input images from Deep Fashion test set. Note [1] is supervised w.r.t. shape.

	mAP	rank-1	rank-5	rank-10
VU-Net [1]	88.7%	87.5%	98.7%	99.5%
Ours	90.3%	89.4%	98.2%	99.2%

Figure 1.1: Transferring shape and appearance on Deep Fashion. Without annotation the model estimates shape, 2nd column. Target appearance is extracted from images in top row to synthesize images. Note that we trained without image pairs only using synthetic transformations. All images are from the test set.

Table 1.2: Percentage of Correct Keypoints (PCK) for pose estimation on shape/appearance swapped generations. α is pixel distance divided by image diagonal. Note that [1] serves as upper bound, as it uses the groundtruth shape estimates.

α	2.5%	5%	7.5%	10%
VU-Net [1]	95.2%	98.4%	98.9%	99.1%
Ours	85.6%	94.2%	96.5%	97.4%

set. On the generated images we evaluate the standard metrics for ReID, mean average precision (mAP) and rank-1, -5, and -10 accuracy in Tab. 1.1. Although our approach is unsupervised it is competitive compared to the supervised VU-Net [1].

1.1.2 Pose Estimation

To evaluate shape, we extract keypoints using the pose estimator [9]. Tab. 1.2 reports the difference between generated and pose target in percentage of correct keypoints (PCK), Fig. 1.3 shows the comparison of PCK curves. As would be expected, VU-Net performs better, since it is trained with exactly the keypoints of [9]. Still our approach achieves an impressive PCK without supervision underlining the disentanglement of appearance and shape.

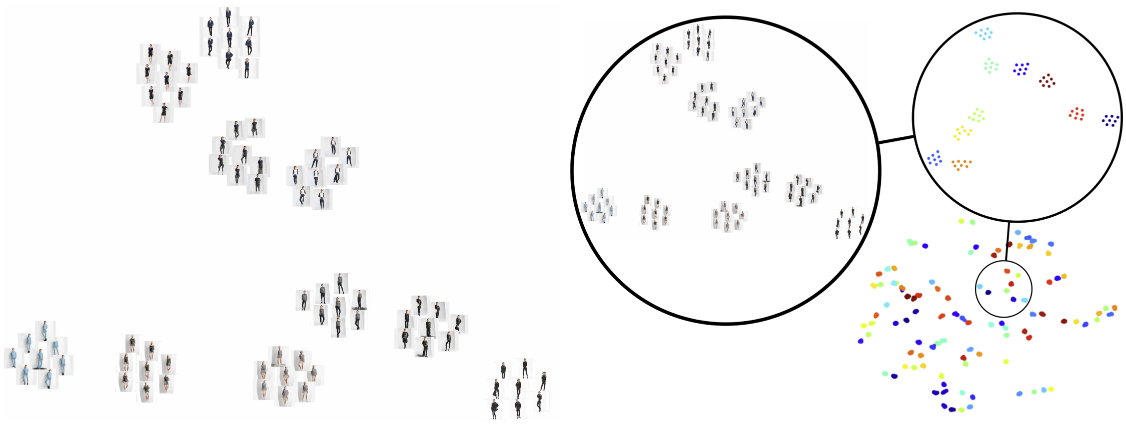


Figure 1.2: Visualization of feature distribution for generated person. (Right) t-SNE (perplexity 16) of 10 generated IDs, (left) color-coded t-SNE (perplexity 12) for 10, 15, 20 and 100 IDs. Each ID has 8 samples. The different IDs are clearly separable, despite variation in pose: Hence, generated appearance is invariant to pose.

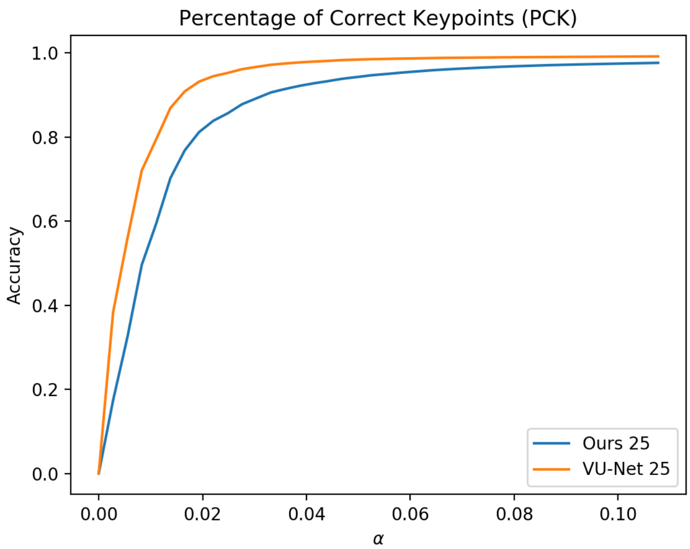


Figure 1.3: PCK Curve for VU-Net [1] and Ours for re-estimating pose with a 25 keypoint human pose detector.

1.2 Factorizing into Parts

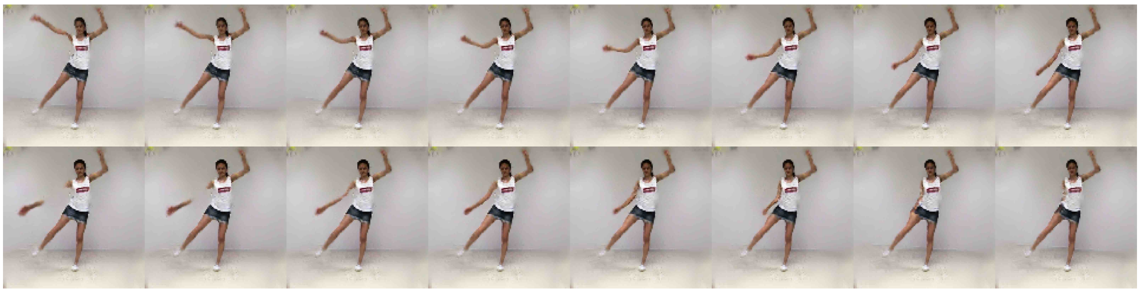
(a) (b) (c) (d)

Figure 1.4: Swapping part appearance on Deep Fashion. Appearances can be exchanged for parts individually and without altering shape. We show part-wise swaps for (a) head (b) torso (c) legs, (d) shoes. All images are from the test set.

- Own Dataset: Move KP
- DeepFashion: exchange parts

1.3 Follow-Up

- make generative:(KP distribution estimation, variational features).
- make video generation possible (RNN on KP vector).
- better transformations -> appearance locally (around parts changed), appearance changed perceptually -> style transfer
- local appearance change (as TPS)



(a)

(b)

Figure 1.5: Moving individual body landmarks for conditional generation (a) arm (b) head.

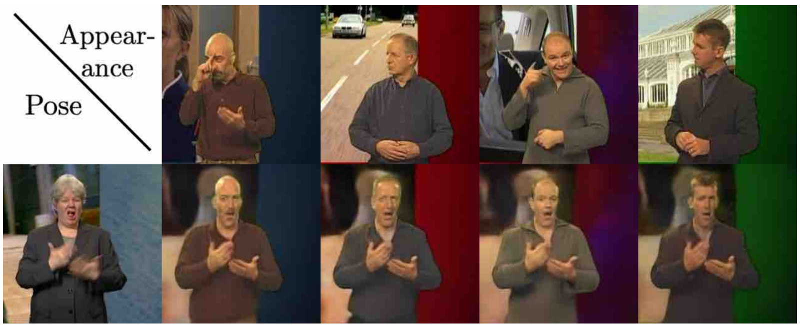


Figure 1.6: Video-to-video translation on BBC Pose. Top-row: target appearances, left: target pose. Note that even fine details in shape are accurately captured. See supplementary for videos.

2 Bibliography

[1] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018. 2, 3, 4

[2] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017. 2

[3] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 2

[4] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2017. 2

[5] Rodrigo de Bem, Arnab Ghosh, Thalaiyasingam Ajanthan, Ondrej Miksik, N Siddharth, and Philip H S Torr. Dgpose: Disentangled semi-supervised deep generative models for human body analysis. *arXiv*, 2018. 2

[6] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018. 2

[7] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 2

[8] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 2

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 3

[10] 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. *ICCV*, 2015. 2

[11] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[12] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*. IEEE, 2017. 2

[13] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 2