Contents

1 Disentanglement of Shape and Appearance

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

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.

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.

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]) as full visibilty of the object is an assumption to the model and used the provided train-test split. For comparison with Esser *et al*. [1] we used their published code.

1.1.1 Person Re-Identification

Person re-identification (ReID) is a research field on its own (overview in *e.g*. [10, 11]), the goal being to learn a similarity metric for a persons appearance, invariant to a persons posture and the image viewpoint. The key applications are automated person tracking and surveillance [12]. For our purposes, we will treat a ReID algorithm as a metric for measuring the preservation of appearance as well as the invariance to shape on our generated images. For this, we fine-tune an ImageNet-pretrained [13] Inception-Net [14] with a ReID algorithm [15] via a triplet loss [16] to the Deep Fashion training set. On the generated images we report the standard metrics for ReID, mean average precision (mAP) and rank-1, -5, and -10 accuracy. The first question we ask, is, if the appearance encoding is stable to variations in pose, hence invariant to pose (shape). Each ID from the test set is generated in 8 different poses. The task for the ReID algorithm is now to rank the similarity of these pose-changed yet same-appearance generations. Although our approach is unsupervised, it is competitive compared to the supervised VU-Net [1] as shown in Tab. 1.1. The high chance of reidentifying a persons appearance in a different shape (rank-1 accuracy) shows that the appearance is invariant against variation in shape (pose) for both methods. To visualize the closeness of the same-ID generations in the ReID-embedding the show a t-SNE plot in Fig. 1.2.

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%

The second question one could ask is, if appearance is preserved, *i.e*. if the ReID algorithm is able to reidentify the groundtruth appearance image from the generation. Results for this are shown in Tab. 1.2. The result depends strongly on whether the algorithm had been fine-tuned to the DeepFashion image distribution or the DeepFashion and the synthesized image distribution. The stark difference can be explained by the difference in

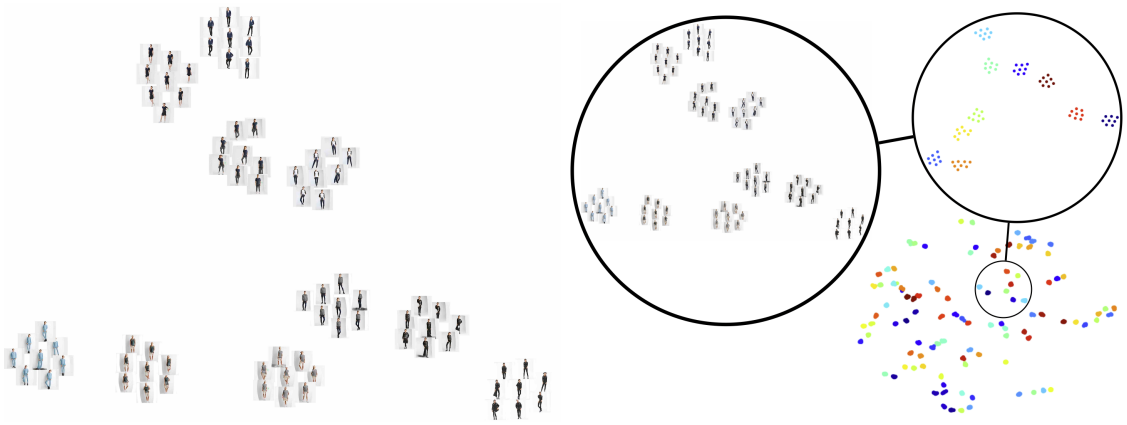


Figure 1.2: Visualization of feature distribution for generated person IDs. (Right) t-SNE (perplexity 16) of 10 generated IDs, (left) color-coded t-SNE (perplexity 12) for 10, 15, 20 and 100 IDs. Each ID has 8 samples. The different IDs are clearly separable, despite variation in pose: Hence, generated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

the feature distribution: high-frequency details (such as patterns and texture of clothing), are not synthesized correctly, as the model is trained by a reconstruction objective which will blurr these high frequencies. On the other hand, the adversarial objective will encourage *some* high frequencies, but not necessarily the ones from the initial appearance conditioning. The ReID algorithm, if not additionally adjusted to this, will pay attention to those details and subsequently fail. unambiguous function between generations and groundtruth can be found.

Table 1.2: Mean average precision (mAP) and rank-n accuracy for person re-identification from synthesized to ground truth appearance images after performing shape/appearance swap. When only fine-tuning the ReID algorithm on DeepFashion, results are much worse that when also adjusting to the synthesized images.

Fine-tune to:	mAP	rank-1	rank-5	rank-10
DeepFashion	17.2%	25.4%	48.8%	60.4%
DeepFashion and Synthesized Images	75.0%	73.8%	89.5%	92.5%

1.1.2 Pose Estimation

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

Table 1.3: 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%

ance and shape. Despite random variation in appearance the shape does not change, this can also be directly observed from the conditioned generations in Fig. 1.1.

1.2 Disentangling in a Temporal Sequence

Figure 1.3: Generation results for conditioning appearances (top row) on pose (bottom, rightmost) on BBCPose. Note that even fine-grained details in shape, such as fingers and facial expression are accurately captured.

Conditional image generation can also be extended to the task of video-to-video translation. The two conditioning images can be frames from different videos. One frame is acting as the appearance conditioning and the other as shape conditioning. By generating each frame conditioned on the shape and appearance from two videos, one effectively transfers the appearance of one video to the shape of the other on a frame-to-frame level. We evaluate this frame-to-frame video translation on the BBCPose dataset. The datasets videos of sign language present a delicate and complex articulation of arms and hands. We condition on appearance from videos in the training set and on shape from videos in the test set. A sample for generated frames is shown in Fig. 1.3, for the complete

videos please refer to the supplementary. We want to point out two features of the model here: Firstly, despite no use of smoothing or interpolation between frames the generated sequence is smooth in the temporal domain. This is enabled by a temporally consistent part assignment which is stable across articulation. Secondly, the training on the natural spatial transforms in video data enables the model to encapsulate realistic transitions such as out-of-plane rotation and complex 3D articulation of *e.g*. hands and even fingers (note the correct translation of the thumbs position in Fig. 1.3).

1.3 Part-wise Disentanglement

The second type of disentanglement we approach is to factorize the object into local parts. Obviously, object parts are in general not disentangled factors in a sense that they have an independent probability distribution. Since the parts are geometrically connected, in their spatial layout they are conditioned on each other. For example, you cannot move your head arbitrarily far away from your shoulders. Still, there is a local freedom and modularity - especially in appearance features - that renders a local factorization efficient. As an illustration, the color of the shoes you wear need only be mildly correlated with your hair color. We show that the model disentangles these local modes of variation for a persons appearance (Sec. 1.3.1) and shape (Sec. 1.3.2).

1.3.1 Part-wise Appearance Transfer

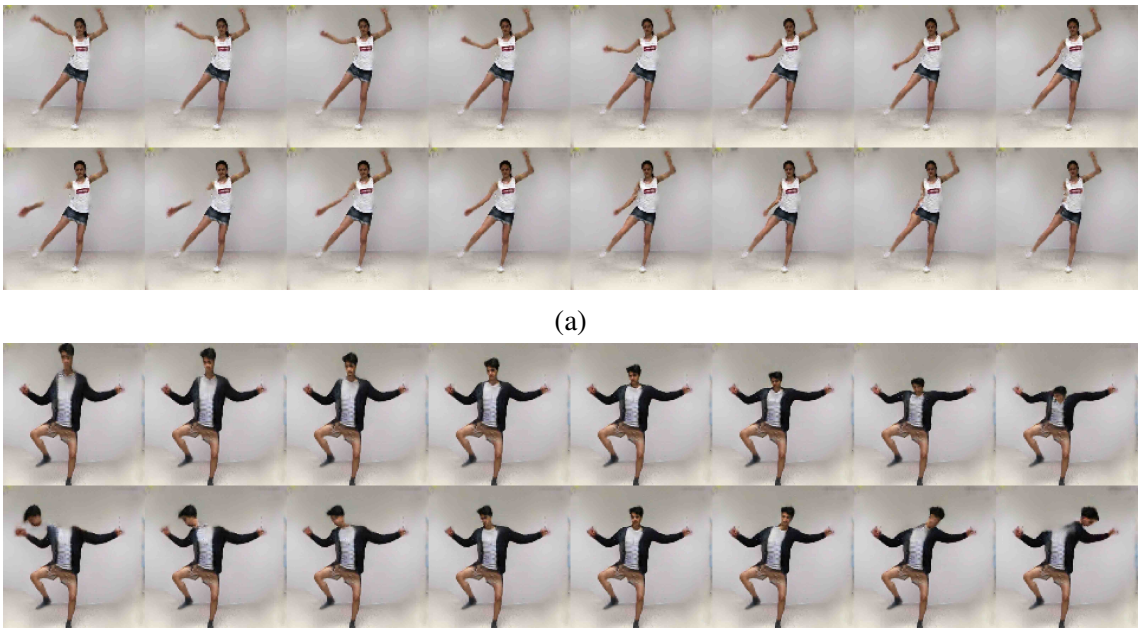
The local modelling of parts allows for a part-wise transfer of appearance. In Fig. 1.4 we show the image generation conditioned on a target shape and appearance from a single image, but for several parts the appearance is transferred from another image. This shows a possible application as a virtual try-on generation, as in [17].

1.3.2 Part-wise Shape Changes

One can also change the position of individual parts in the shape conditioning, which leads to generations as shown in Fig. 1.5. One can observe that the other non-moved parts shapes also lead to stationary parts in the generation, indicating that these parts are spatially disentangled. In the unnatural (never seen in data) regime *e.g*. if the head is to far from the shoulders, the model still hallucinates a head next to the body - similar to supervised results [5].

(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.



(a) (b)

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

2 Bibliography

[1] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018. 2, 3, 5

[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

[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, 3

[8] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 2, 3

[9] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3, 4

[10] Jon Almazán, Bojana Gajic, Naila Murray, and Diane Larlus. Re-ID done right: towards good practices for person re-identification. *arXiv*, 2018. 3

[11] Apurva Bedagkar-Gala and Shishir K. Shah. A survey of approaches and trends in person re-identification. *Image Vision Comput.*, 2014. 3

[12] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person Re-identification: Past, Present and Future. *arXiv*, 2016. 3

[13] 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. 3

[14] 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. 3

[15] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*. IEEE, 2017. 3

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 3

[17] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: An Image-based virtual try-on network. *arXiv*, 2017. 6