Contents

1 Experiments

1.1 Shape Learning

Fig. 1.1 visualizes the learned shape representation. To quantitatively evaluate the shape estimation, we measure how well groundtruth landmarks (only during testing) are predicted from it. The part means $\mu[\sigma_i(x)]$ make compatible with method serve as our landmark estimates and we measure the error when linearly regressing the human-annotated groundtruth landmarks from our estimates. For this, we follow the protocol of Thewlis *et al*. [2], fixing the network weights after training the model, extracting unsupervised landmarks and training a single linear layer without bias. The performance is quantified on a test set by the mean error and the percentage of correct landmarks (PCK). We extensively evaluate our model on a diverse set of datasets, each with specific challenges. An overview over the challenges implied by each dataset is given in Tab. 1.1. On all datasets we outperform the state-of-the-art by a significant margin.

1.1.1 Landmark Discovery

On the object classes of human faces, cat faces, and birds (datasets CelebA, Cat Head, and CUB-200-2011) our model predicts landmarks consistently across different instances, cf. Fig. 1.2. Tab. 1.2 compares against the state-of-the-art. Due to different breeds and species the Cat Head, CUB-200-2011 exhibit large variations between instances. Especially on these challenging datasets we outperform competing methods by a large margin. Fig. 1.3 also provides a direct visual comparison to [1] on CUB-200-2011. It becomes evident that our predicted landmarks track the object much more closely. In contrast, [1] have learned a slightly deformable, but still rather rigid grid. This is due to their separation constraint, which forces landmarks to be mutually distant. We do not need such

Table 1.1: Difficulties of datasets: articulation, intra-class variance, background clutter and viewpoint variation

Dataset	Articul.	Var.	Backgr.	Viewp.
CelebA				
Cat Head		✓		
CUB-200-2011		✓	✓	
Human3.6M	✓			✓
BBC Pose	✓		✓	
Dogs Run	✓	✓	✓	
Penn Action	✓	✓	✓	✓

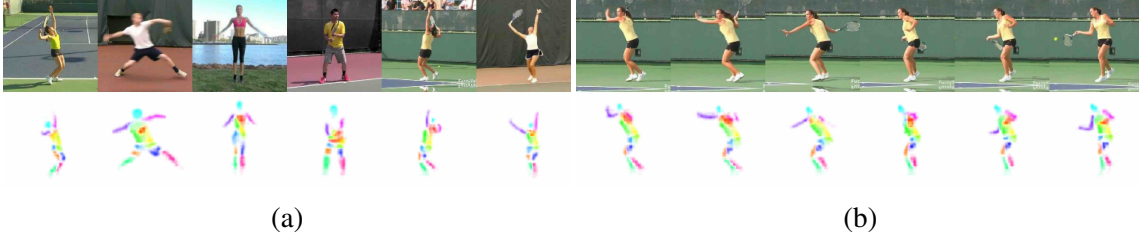(a)   (b)

Figure 1.1: Learned shape representation on Penn Action. For visualization, 13 of 16 part activation maps are plotted in one image. (a) Different instances, showing intra-class consistency and (b) video sequence, showing consistency and smoothness under motion, although each frame is processed individually.

Table 1.2: Error of unsupervised methods for landmark prediction on the Cat Head, MAFL (subset of CelebA), and CUB-200-2011 testing sets. The error is in % of inter-ocular distance for Cat Head and MAFL and in % of edge length of the image for CUB-200-2011.

Dataset	Cat Head		MAFL	CUB
# Landmarks	10	20	10	10
Thewlis [2]	26.76	26.94	6.32	-
Jakab [3]	-	-	4.69	-
Zhang [1]	15.35	14.84	3.46	5.36
Ours	**9.88**	**9.30**	**3.24**	**3.91**

a problematic bias in our approach, since the localized, part-based representation and reconstruction guides the shape learning and captures the object and its articulations more closely.

Composite Objects/Scenes

What is an object? What is a scene? compositional nature of reality Bird on twig object? Bird can also fly, but neural networks learn by correlation in data (-> ref to these "failure modes" Dancing pair as object.

Object/Background Separation

Complexly cluttered background is actually favorable for the method. Correlations of object with background will belong to object.

Object Articulation

Object articulation makes consistent landmark discovery challenging. Fig. 1.2 shows that our model exhibits strong landmark consistency under articulation and covers the full human body meaningfully. Even fine-grained parts such as the arms are tracked
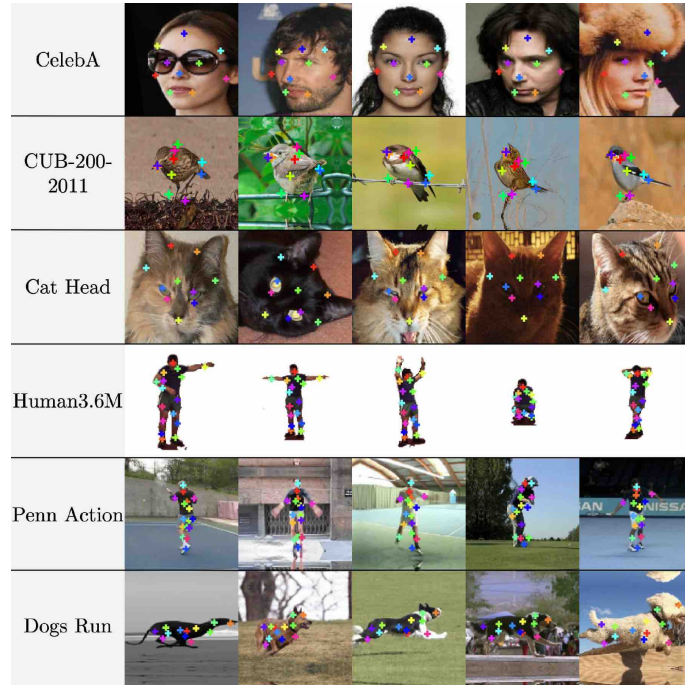
Figure 1.2: Unsupervised discovery of landmarks on diverse object classes such as human or cat faces and birds and for highly articulated human bodies and running dogs.
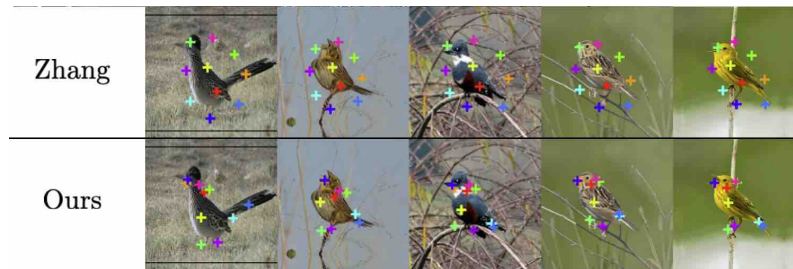

Figure 1.3: Comparing discovered keypoints against [1] on CUB-200-2011. We improve on object coverage and landmark consistency. Note our flexible part placement compared to a rather rigid placement of [1] due to their part separation bias.

Table 1.3: Performance of landmark prediction on BBC Pose test set. As upper bound, we also report the performance of supervised methods. The metric is % of points within 6 pixels of groundtruth location.

BBC Pose		Accuracy
supervised	Charles [4]	79.9%
	Pfister [5]	88.0%
unsupervised	Jakab [3]	68.4%
	Ours	**74.5**%

Table 1.4: Comparing against supervised, semi-supervised and unsupervised methods for landmark prediction on the Human3.6M test set. The error is in % of the edge length of the image. All methods predict 16 landmarks.

Human3.6M		Error w.r.t. image size
supervised	Newell [6]	2.16
semi-supervised	Zhang [1]	4.14
unsupervised	Thewlis [2]	7.51
	Zhang [1]	4.91
	Ours	**2.79**

across heavy body articulations, which are frequent in the Human3.6M and Penn Action datasets. Despite further complications such as viewpoint variations or blurred limbs our model can detect landmarks on Penn Action of similar quality as in the more constrained Human3.6M dataset. Additionally, complex background clutter as in BBC Pose and Penn Action, does not hinder finding the object. Experiments on the Dogs Run dataset underlines that even completely dissimilar dog breeds can be related via semantic parts. Tab. 1.3 and Tab. 1.4 summarize the quantitative evaluations: we outperform other unsupervised and semi-supervised methods by a large margin on both datasets. On Human3.6M, our approach achieves a large performance gain even compared to methods that utilize optical flow supervision. On BBC Pose, we outperform [3] by 6.1%, reducing the performance gap to supervised methods significantly.

1.1.2 Effect of Transformations

in this section: effect of transformations on learning, disentangling effectively connecting samples from the dataset, spreading the

Parity

birds parity salsa parity

Figure 1.4: 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 test set.

Rotation, Scaling, Translation

on Cats -> black cats different set of KP than rest -> connect these samples via transformation to reach intra-class consistency

Mimicking Appearance

Color, Contrast, Hue

1.1.3 Natural Changes

Video data: Penn, Own

1.2 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,

Table 1.5: 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 [7] is supervised w.r.t. shape.

	mAP	rank-1	rank-5	rank-10
VU-Net [7]	88.7%	87.5%	98.7%	99.5%
Ours	90.3%	89.4%	98.2%	99.2%

Table 1.6: Percentage of Correct Keypoints (PCK) for pose estimation on shape/appearance swapped generations. α is pixel distance divided by image diagonal. Note that [7] serves as upper bound, as it uses the groundtruth shape estimates.

α	2.5%	5%	7.5%	10%
VU-Net [7]	95.2%	98.4%	98.9%	99.1%
Ours	85.6%	94.2%	96.5%	97.4%

for instance, to change the pose of a person or their clothing. In contrast to previous work [7, 8, 9, 10, 11, 3], 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.

t-SNE of Shape Representation t-SNE of Appearance Representation

1.2.1 Disentangling Pose and Appearance

On Deep Fashion [12, 13], 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. 1.4 shows qualitative results. We quantitatively compare against supervised state-of-the-art disentangling [7] 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.

ReID

- t-SNE of IDs

- Own, Other (stronger statement)

To evaluate appearance we fine-tune an ImageNet-pretrained [14] Inception-Net [15] with a re-identification (ReID) algorithm [16] via a triplet loss [17] to the Deep Fashion training

Figure 1.5: 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.

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. 1.5. Although our approach is unsupervised it is competitive compared to the supervised VU-Net [7].

Pose

To evaluate shape, we extract keypoints using the pose estimator [18]. Tab. 1.6 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 key-points of [18]. Still our approach achieves an impressive PCK without supervision under-lining the disentanglement of appearance and shape.

- PCK Curve

1.2.2 Factorizing into Parts

- 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
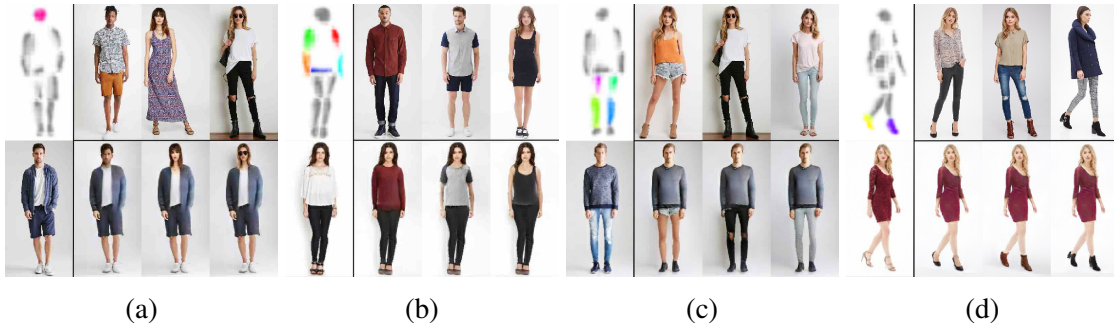
Figure 1.6: Swapping part appearance on Deep Fashion. Appearances can be exchanged for parts individually and without altering shape. We show part-wise swaps for (a) head (b) torso (c) legs, (d) shoes. All images are from the test set.

Part I

Appendix

A Bibliography

[1] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018. 2, 3, 4, 5

[2] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 2, 3, 5

[3] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018. 3, 5, 7

[4] James Charles, Tomas Pfister, Derek R Magee, David C Hogg, and Andrew Zisserman. Domain adaptation for upper body pose tracking in signed tv broadcasts. In *BMVC*, 2013. 5

[5] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 5

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016. 5

[7] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018. 7, 8

[8] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017. 7

[9] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 7

[10] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2017. 7

[11] 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. 7

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

[13] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 7

[14] 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. 7

[15] 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. 7

[16] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*. IEEE, 2017. 7

[17] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 7

[18] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 8