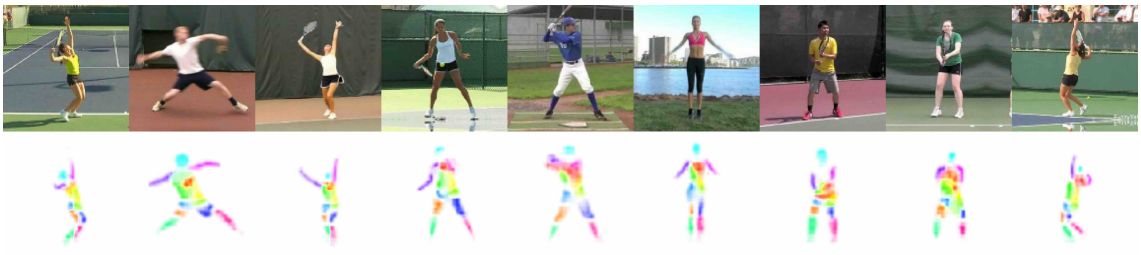
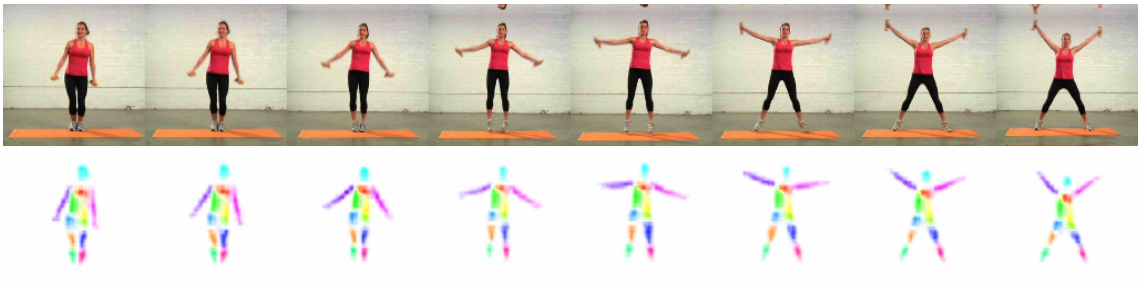
1 Experiments

1.1 Shape Learning



(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.

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 $\mu[\sigma_i(x)]$ 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. 1.2. On all datasets we outperform the state-of-the-art by a significant margin.

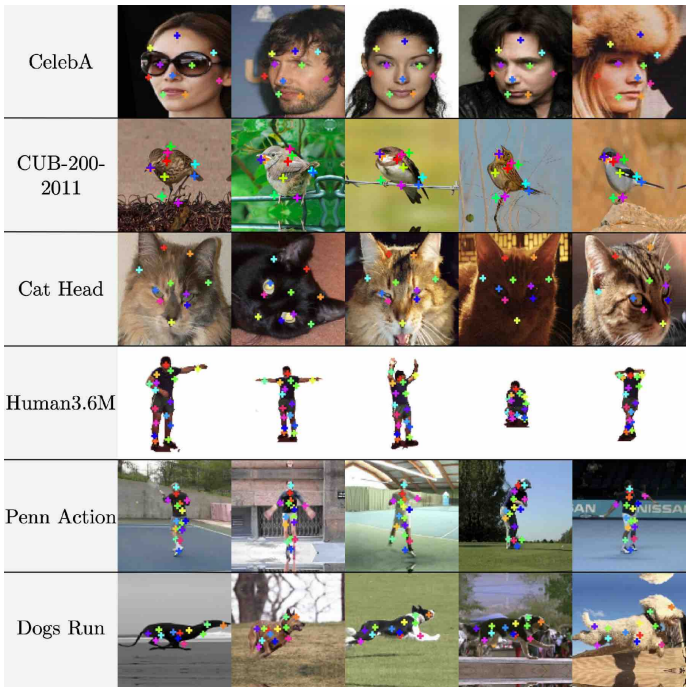


Figure 1.2: Unsupervised discovery of landmarks on diverse object classes such as human or cat faces and birds and for highly articulated human bodies and running dogs.

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 1.1: 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**

1.1.1 Dataset Preprocessing

CelebA [4] contains ca. 200k celebrity faces of 10k identities. We resize all images to 128×128 and exclude the training and test set of the MAFL subset, following [2]. As [2, 1], we train the regression (to 5 ground truth landmarks) on the MAFL training set (19k images) and test on the MAFL test set (1k images).

Cat Head [5] has nearly 9k images of cat heads. We use the train-test split of [1] for training (7,747 images) and testing (1,257 images). We regress 5 of the 7 (same as [1]) annotated landmarks. The images are cropped by bounding boxes constructed around the mean of the ground truth landmark coordinates and resized to 128×128.

CUB-200-2011 [6] comprises ca. 12k images of birds in the wild from 200 bird species. We excluded bird species of seabirds, roughly cropped using the provided landmarks as bounding box information and resized to 128×128. We aligned the parity with the information about the visibility of the eye landmark. For comparing with [1] we used their published code.

BBC Pose [7] contains videos of sign-language signers with varied appearance in front of a changing background. Like [3] we loosely crop around the signers. The test set includes 1000 frames and the test set signers did not appear in the train set. For evaluation, as [3], we utilized the provided evaluation script, which measures the PCK around $d = 6$ pixels in the original image resolution.

Human3.6M [8] features human activity videos. We adopt the training and evaluation procedure of [1]. For proper comparison to [1] we also removed the background using the off-the-shelf unsupervised background subtraction method provided in the dataset.

Penn Action [9] contains 2326 video sequences of 15 different sports categories. For this experiment we use 6 categories (tennis serve, tennis forehand, baseball pitch, baseball

Table 1.2: 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	✓	✓	✓	✓

swing, jumping jacks, golf swing). We roughly cropped the images around the person, using the provided bounding boxes, then resized to 128×128.

Dogs Run is made from dog videos from YouTube totaling in 1250 images under similar conditions as in Penn Action. The dogs are running in one direction in front of varying backgrounds. The 17 different dog breeds exhibit widely varying appearances.

Deep Fashion [10, 11] 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 [12]) and used the provided train-test split. For comparison with Esser *et al*. [13] we used their published code.

1.1.2 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. 1.1 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 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

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 [7]	79.9%
	Pfister [14]	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 [15]	2.16
semi-supervised	Zhang [1]	4.14
unsupervised	Thewlis [2]	7.51
	Zhang [1]	4.91
	Ours	**2.79**

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 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.

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 [13] is supervised w.r.t. shape.

	mAP	rank-1	rank-5	rank-10
VU-Net [13]	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 [13] serves as upper bound, as it uses the groundtruth shape estimates.

α	2.5%	5%	7.5%	10%
VU-Net [13]	95.2%	98.4%	98.9%	99.1%
Ours	85.6%	94.2%	96.5%	97.4%

1.1.3 Effect of Transformations

Parity

birds parity salsa parity

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

1.1.4 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, for instance, to change the pose of a person or their clothing. In contrast to previous work [13, 16, 17, 18, 19, 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

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.

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.

1.2.1 Disentangling Pose and Appearance

On Deep Fashion [10, 11], 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 [13] 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.

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.

ReID

- t-SNE of IDs
- Own, Other (stronger statement)

To evaluate appearance we fine-tune an ImageNet-pretrained [20] Inception-Net [21] with a re-identification (ReID) algorithm [22] via a triplet loss [23] to the Deep Fashion training 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 [13].

Pose

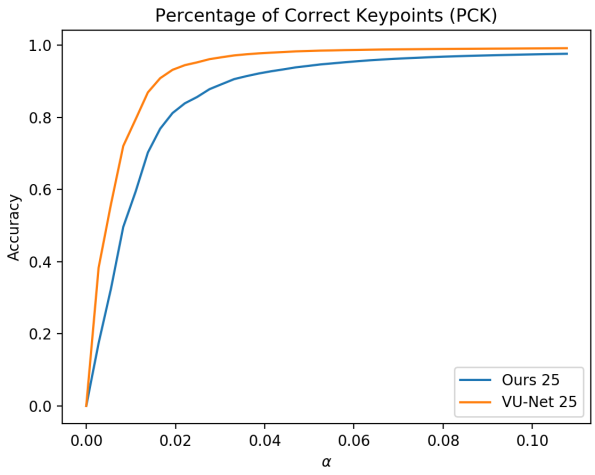


Figure 1.6: PCK Curve for VU-Net [13] and Ours for re-estimating pose with a 25 keypoint human pose detector.

To evaluate shape, we extract keypoints using the pose estimator [12]. Tab. 1.6 reports the difference between generated and pose target in percentage of correct keypoints (PCK), Fig. 1.6 shows the comparison of PCK curves. As would be expected, VU-Net performs better, since it is trained with exactly the keypoints of [12]. Still our approach achieves an impressive PCK without supervision underlining the disentanglement of appearance and shape.

1.2.2 Factorizing into Parts

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

Figure 1.7: 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

2 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. 1, 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, 6

[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 3

[5] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008. 3

[6] C Wah, S Branson, P Welinder, P Perona, and S Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 3

[7] 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. 3, 5

[8] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 3

[9] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 3

[10] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 4, 7

[11] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 4, 7

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

[13] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018. 4, 6, 7, 8

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

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

[16] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017. 6

[17] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 6

[18] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2017. 6

[19] 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. 6

[20] 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. 8

[21] 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. 8

[22] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*. IEEE, 2017. 8

[23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 8