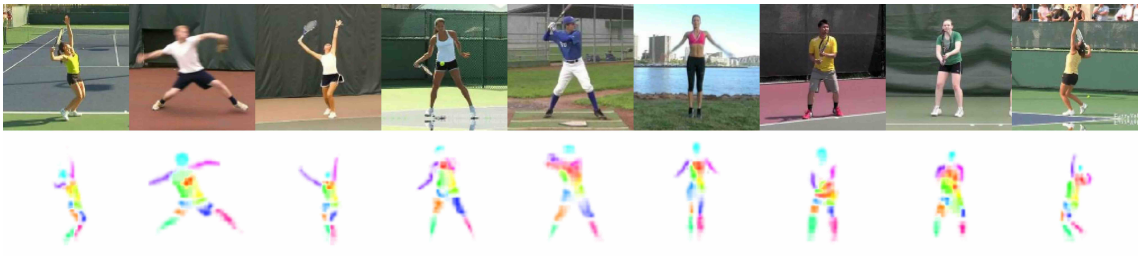
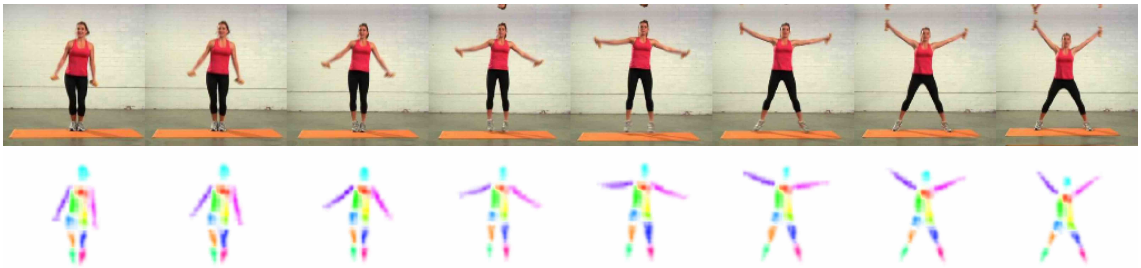
Contents

1 Object 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.

In this section we will establish that the proposed method (Sec. **??**) outperforms the state-of-the-art in unsupervised object shape learning by a large margin. The learned shape representation is visualized in Fig. 1.1. To quantitatively evaluate the shape estimation, we measure how well groundtruth landmarks (only during testing) are predicted from it. We obtain landmarks from our part-region based shape representation by designating the mean of a part shape $\mu[\sigma^i(\mathbf{x})]$ as the landmark position. To quantify the quality of these landmark estimates, we linearly regress them to human-annotated groundtruth landmarks and measure the test error. For this, we follow the protocol of Thewlis *et al.* [1], 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.

In the following, we proceed through our shape learning results: we present the quantitative and qualitative results by object category (Sec. 1.1). On the way we introduce the

datasets for each category. In the next section we highlight and discuss the challenges, which the datasets present (Sec. 1.2) and subsequently argue for the importance of the transformations and modelling assumptions as a means to reach disentanglement and to overcome those challenges (Sec. 1.3). This confirms that disentangled modelling aids the learning of shape (hypothesis I, Sec. **??**).

1.1 Diverse Object Categories

We test our approach on a diverse set of object classes ranging from human and cat faces to articulated bodies and animals. In the following we go through the results sorted by object category. Where possible we compared to state-of-the-art methods quantitatively in terms of unsupervised landmark prediction, additionally we show qualitative results.

1.1.1 Human and Cat Faces

(a)

(b)

Figure 1.2: Unsupervised discovery of landmarks the object classes of (a) human (CelebA dataset) and (b) cat faces (Cat Head dataset).

For human and cat faces we use the popular datasets CelebA and CatHead, since our predecessors for unsupervised shape learning established baselines here. Both object categories are rather rigid and non-articulated - meaning that the relations between object parts are not changing from instance to instance. Due to different breeds, the Cat Head dataset exhibits large variations between instances. Cat faces feature more complicated

Table 1.1: Error of unsupervised methods for landmark prediction on the Cat Head, MAFL (subset of CelebA) testing sets. The error is in % of inter-ocular distance.

Dataset	Cat Head		MAFL
# Landmarks	10	20	10
Thewlis [1]	26.76	26.94	6.32
Jakab [5]	-	-	4.69
Zhang [4]	15.35	14.84	3.46
Ours	**9.88**	**9.30**	**3.24**

texture and locally variant silhouettes [2], hence, require a better learning of both shape and appearance.

CelebA [3] 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 [1]. As [1, 4], 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 [2] has nearly 9k images of cat heads. We use the train-test split of [4] for training (7,747 images) and testing (1,257 images). We regress 5 of the 7 (same as [4]) annotated landmarks. The images are cropped by bounding boxes constructed around the mean of the ground truth landmark coordinates and resized to 128×128.

Qualitative results

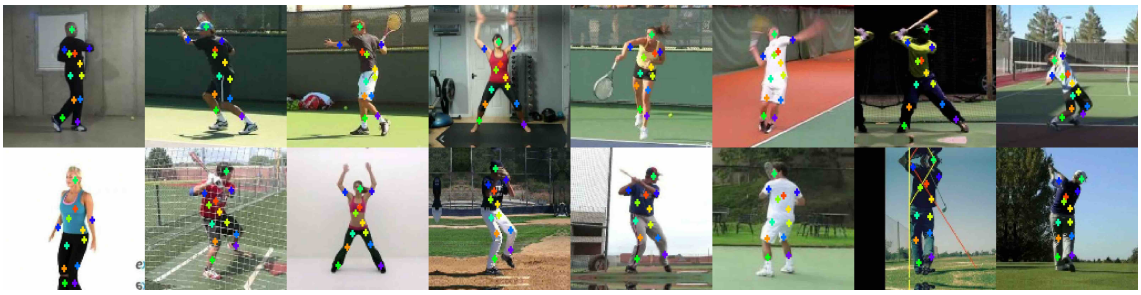
The algorithm successfully defines correspondences between different human individuals, and also generalizes between different cat breeds. On both datasets the performance is visibly near-perfect. Difficulties such as out-of-plane rotation, varying lighting conditions and part occlusions (*e.g.* sunglasses) do not diminish its ability to determine the self-defined keypoints.

Quantitative results

Tab. 1.1 compares against the state-of-the-art. Our approach outperforms competing methods, with a particularly large margin of ca. $4 - 5\%$ on the more challenging Cat Head dataset. The best competitor suffers from an incomplete disentanglement (as we show in Sec. **??**). An interesting side note: the most severe failure modes for Cat Head were human labelling errors, which suggests that the unsupervised performance could be better than the human labelling in this circumstances.

1.1.2 Human Bodies

(a)



(b)

Figure 1.3: Unsupervised discovery of landmarks the object classes of human bodies (a) in constrained (Human3.6M dataset) and (b) unconstrained environments (Penn Action dataset).

Human bodies introduce the challenge of object articulation. We test on three datasets. The BBC Pose dataset is used to compare against Jakab *et al*. [5], the Human3.6M dataset to beat the benchmark of Zhang *et al*. [4]. Additionally, we present the first unsupervised results on Penn Action, which is, as we argue, significantly more difficult.

BBC Pose [6] contains videos of sign-language signers with varied appearance in front of a changing background. Like [5] 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 [5], we utilized the provided evaluation script, which measures the PCK around $d = 6$ pixels in the original image resolution.

Human3.6M [7] features human activity videos. We adopt the training and evaluation procedure of [4]. For proper comparison to [4] we also removed the background using the off-the-shelf unsupervised background subtraction method provided in the dataset.

Penn Action [8] contains 2326 video sequences of 15 different sports categories. For this experiment we use 6 categories (tennis serve, tennis forehand, baseball pitch, baseball swing, jumping jacks, golf swing). We roughly cropped the images around the person, using the provided bounding boxes, then resized to 128×128.

Qualitative results

We demonstrate (Fig. 1.3), that our model not only exhibits strong landmark consistency under articulation, but also covers the full human body meaningfully. Even fine-grained parts such as the arms are tracked across heavy body articulations, as are present in the Human3.6M or Penn Action datasets. Also with further complications introduced with the Penn Action dataset such as viewpoint variations, blurred limbs and partial self-occlusions we are able to detect landmarks of similar quality and coverage 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.

Quantitative results

Table 1.2: 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 [6]	79.9%
	Pfister [9]	88.0%
unsupervised	Jakab [5]	68.4%
	Ours	**74.5**%

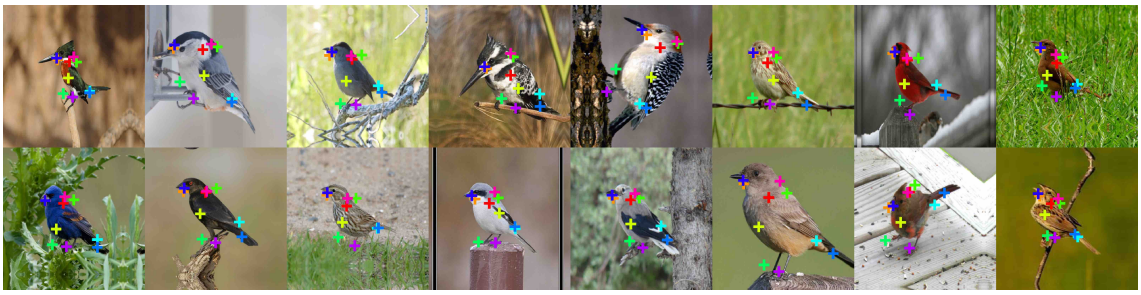
Table 1.3: Comparing against supervised, semi-supervised and unsupervised methods for landmark prediction on the Human3.6M test set. The error is in % of the edge length of the image. All methods predict 16 landmarks.

Human3.6M		Error w.r.t. image size
supervised	Newell [10]	2.16
semi-supervised	Zhang [4]	4.14
unsupervised	Thewlis [1]	7.51
	Zhang [4]	4.91
	Ours	**2.79**

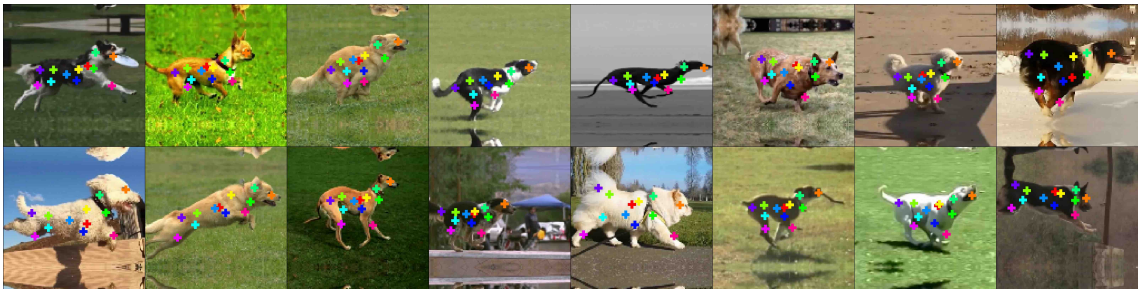
The quantitative comparisons are shown in Tab. 1.2 and Tab. 1.3: other unsupervised and semi-supervised methods are outperformed by a large margin on both datasets.

On Human3.6M, judging by the performance gap, it is questionable whether the other unsupervised method from Thewlis *et al.* [1] learned to deal with articulation at all or whether they just find a mean solution.We beat the best unsupervised result by Zhang *et al.* [4] by an improvement of 2.12%. This is not only cutting the absolute error nearly in half, but also reduces the gap between unsupervised and supervised algorithms by about 77%. Zhang *et al.* [4] additionally used optical flow to stabilize their training by forcing the landmarks to cover the object, which we (and they themselves) classified as semi-supervised. Despite this advantage, our approach is able to achieve a performance gain of 1.35% even over results obtained with optical flow supervision. On BBC Pose, we outperform Jakab *et al.* [5] by 6.1%, which translates into a reduction of the unsupervised performance gap to supervised methods by more than half. An analysis of conceptual differences to both [4] and [5] can be found in Sec. 1.4.

1.1.3 Animal Bodies



(a)



(b)

Figure 1.4: Unsupervised discovery of landmarks the object classes of animal bodies (a) birds (CUB-200-2011 dataset) and (b) dogs (Dogs Run dataset).

Table 1.4: Error of unsupervised methods for landmark prediction on the CUB-200-2011 testing set. Both methods predict 10 landmarks.

CUB-200-2011 dataset	Error w.r.t. image edge
Zhang [4]	5.36
Ours	**3.91**

Dataset preprocessing

CUB-200-2011 [11] 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 [4] we used their published code.

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.

Qualitative results

The Dogs Run dataset displays that even completely different dog breeds can be related via semantic parts. Here, the universality of the approach (capturing non-human poses) is underlined once more. This is to be expected, as no prior assumptions about the object-class are introduced in the model. Furthermore, the limited amount of data in Dogs Run is no problem for finding meaningful correspondences, due to the unsupervised nature of the model and the transformations acting as a form of data augmentation.

Quantitative results

For a direct comparison to Zhang *et al.* [4] we apply their published code on the CUB-200-2011 dataset. The results are shown in Tab. 1.4.

1.2 Challenges

An overview over the challenges implied by each of the presented datasets is given in Tab. 1.5. We address the main difficulties in the following: background clutter 1.2.1, intra-class variance 1.2.3, articulation and viewpoint variation 1.2.2. We discuss how the method overcomes these challenges.

Table 1.5: 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	✓	✓	✓	✓

1.2.1 Background Clutter

The question how to separate background from object goes deeper than one might think. Fundamentally the question is equivalent to: *What is an object?* If an unsupervised algorithm is posed the task of finding the object in an image dataset - under the assumption that the object is present in all images - the object is a structure common to these images. By this definition background is everything that is not strongly correlated with the object itself. This dataset-specific object category can be unintuitive: for example for a bird sitting on a twig, the twig can considered as part of the object, if the dataset shows only birds on twigs (*e.g.* two landmarks are on feet/twig on CUB-200-2011 dataset, cf. Fig. 1.4). This dataset-biased pre-categorical thinking of unsupervised algorithms can be seen as a failure, or as a feature: on a dataset of Salsa-dancing humans our method identified the pair of dancers as an object (cf. Fig. 1.7). Technically, we allow the algorithm to focus on the reconstruction of only the object, and not background by a local weighting around the part activation (refer to Sec. **??**). The fundamental issue of strong object-background correlations cannot be solved technically, but requires different data. Interestingly, on with the constrained but repetetive background of the Human3.6M dataset, where traditional background subtraction methods are easily applied, our method struggles: several parts are assigned to background objects. On the other hand, complexly cluttered backgrounds - as long as no correlations to the object exist - such as the background TV screen in the BBCPose dataset (cf. Fig. **??**) are actually favorable for the method. This is due to the crossed reconstruction objective: if reconstructing the background of the target image is possible with the information of the source appearance image, the algorithm will try to do so by assigning parts to the background, if not, not.

1.2.2 Object Articulation and Viewpoint Variation

Object articulation and viewpoint variation makes consistent landmark discovery challenging. In contrast to rigid bodies that can vary in orientation and scale, articulated objects have many degrees of freedom more. Part assignment consistency means equivariance w.r.t. changes in shape due to articulation. Equivariance is enforced twice in the method (cf. Sec. **??**). We showed previously that the method can deal with strongly

articulated human and animal bodies (cf. Sec. 1.1.2 and Sec. 1.1.3).

1.2.3 Intra-Class Variation

Intra-class variation can be both in shape and in appearance. Due to different breeds and species the animal datasets - Dogs Run, CUB-200-2011, Cat Head - present the highest degree of variability within the object class. The method in part coincidentally generalizes and in part inherently enforces in due to the data augmentation with shape and appearance transformations (see also Sec. 1.3).

1.3 Transformational Effects

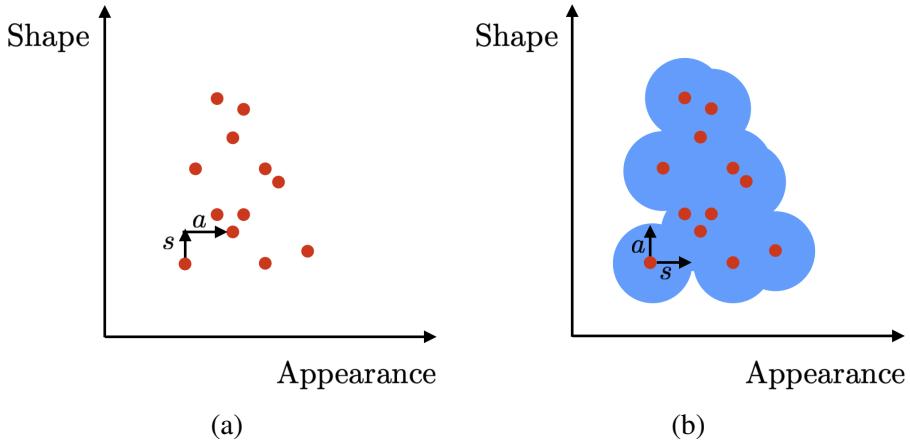


Figure 1.5: Effect of transformations on data distribution: (a) Data points (red) can be connected via a shape s and an appearance a transformation. (b) Applying transformations effectively blurs the data distribution.

In this section we discuss the effect of the transformations on learning a consistent and comprehensive representation. Since strong image transformations can make the learning curve for the algorithm too steep, we exponetially schedule the increase in magnitude, finally resulting in image changes as shown in Fig. 1.6. In effect, the transformations teach the algorithm what changes in shape and appearance are. Assuming that samples from the data distribution are - showing the same object class - related via a change in shape and appearance, the transformations blur the distribution. This data augmentation is sketched in Fig. 1.5.

1.3.1 Spatial Transformations

We perform thin-plate spline (TPS) warps to mimick spatial transformations. These changes incorporate rotation, scaling and translation as a special case. While irreplacable

for calculating the direct equivariance loss, they can result in artificial shape changes. After all, most objects - such as human beings or animals, do not warp, but articulate their parts/limbs. Natural shape changes are needed to learn a model of the objects articulation. These changes are presented in video data. Hence, for videos we enforce the reconstruction to function across different frames. This results in a much stabler performance and greater part consistency especially for highly articulated parts such as arms.

1.3.2 Appearance Transformations

Figure 1.6: Examples for shape and appearance transformation on CUB-200-2011. Images from the upper row relate to images directly below.

We mimick appearance changes with image transformations in color, contrast, hue and brightness. Exemplars for the combined effect of spatial and appearance transformations are shown in Fig. 1.6. Especially for datasets with high intra-class appearance variance, connecting the data points via appearance changes is crucial. On Cat Head for example, without them, the method assigned different landmarks to black cats than to other-color cats. The model will incur no loss, as long as it always has to reconstruct black cats from images of black cats. If it has to relate black and white cats (*e.g.* via color inversion) this intra-class inconsistency has to vanish.
Ideally one would want more "natural" appearance changes as well. This could be a line of future work (cf. Sec. **??**).

1.3.3 Parity

The model has problems with consistent part assignment under parity changes, if these changes do not change the object enough. For example human body appearance in frontal and back view are not dissimilar enough from each other. So the model can assign the same landmark to *e.g.* the right arm in the frontal view and the left arm in the back view. Similarly, there is no distinction, for dog or bird side views facing to the left or right. However, if features on the left and on the right are very different *e.g.* for the dancing humans the model automatically learns the distinction (see Fig. 1.7).

A tentative solution to the problem can be to incorporate parity flips in the equivariance loss. This is impossible for parity-symmetrical objects (such as frontal view humans),

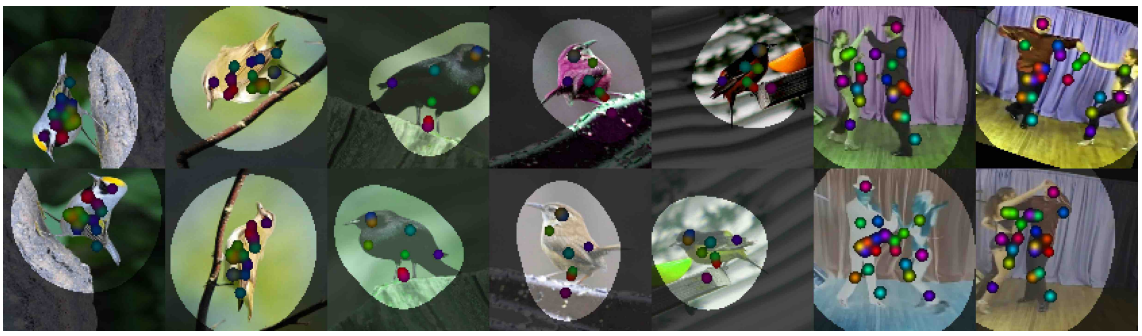


Figure 1.7: Parity changes: the images of the upper and lower row relate via the usual transformations and an additional parity flip. For the bird (1-5th column) images induced artificially, for the dancing humans (6-7th column) via sampling different frames from a video.

but works *e.g.* for side view dogs or birds. One has to be careful in scheduling these random parity flips, as landmarks tend to align along the mirror axis as a trivial solution. A successfully learned parity model for CUB-200-2011 is shown in Fig. 1.7.

1.4 Comparative Advantages

1.4.1 Non-Disentangling Approach

Dataset	Human3.6M					
Actions	Directions	Discussion	Waiting	Greeting	Posing	Walking
Newell [10] (supervis.)	1.88	1.92	2.15	1.62	1.88	2.21
Zhang [4] (semi-superv.)	5.01	4.61	4.76	4.45	4.91	4.61
Thewlis [1] (unsuperv.)	7.54	8.56	7.26	6.47	7.93	5.40
Ours (unsuperv.)	**2.58**	**2.26**	**2.87**	**3.08**	**2.67**	**3.35**

Table 1.6: Comparison with unsupervised, semi-supervised and supervised methods for annotated landmark prediction on the Human 3.6M testing sets for selected actions. The error is in % regarding the edge length of the image. All methods predict 16 landmarks, from which the 32 ground truth landmarks are regressed.

The method of Zhang *et al.* [4] is similar to our method, but does not disentangle shape from appearance. In Tab. 1.6 we list the detailed results of a comparison to Zhang *et al.* [4]. Fig. 1.8 provides a direct visual comparison to [4] on CUB-200-2011. It becomes evident that our predicted landmarks track the object much more closely. In contrast, [4] 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

Figure 1.8: Comparing discovered keypoints against [4] on CUB-200-2011. We improve on object coverage and landmark consistency. Note our flexible part placement compared to a rather rigid placement of [4] due to their part separation bias.

and reconstruction guides the shape learning and captures the object and its articulations more closely.

1.4.2 Holistic Approach

The method of Jakab *et al*. [5] aims disentangling shape and appearance with video information. Shape is then - similar to ours - defined by the change between consecutive frames. However, they do not model the link between local parts of shape and appearance, but use a holistic appearance embedding. Constrained by reality.

Dataset	BBCPose				
Landmarks	Head	Wrists	Elbows	Shoulders	Avg.
Charles et al. [6] (supervised)	95.40	72.95	68.70	90.30	79.90
Pfister et al. [9] (supervised)	98.00	88.45	77.10	93.50	88.01
Jakab [5] (unsupervised)	76.10	56.50	70.70	74.30	68.44
Ours (unsupervised)	**96.34**	**71.39**	**62.12**	**80.28**	**74.85**

Table 1.7: Comparison with supervised and unsupervised methods for annotated landmark prediction on the BBCPose testing sets. %-age of points within 6 pixels of ground-truth is reported.

1.4.3 Ablating Contributions

We ablate the main components of our proposed framework: reconstruction loss $\mathcal{L}_{\text{rec}}$, the equivariance loss $\mathcal{L}_{\text{equiv}}$, the appearance augmentation and the crossing task for disentangling shape and appearance. For the ablation study we use the Cat Head dataset, following the already introduced train-test setup on the task of landmark ground truth regression. Tab. 1.8 illustrates the ablation results.
Leaving out the reconstruction task naturally leads to the largest drop in performance since only training on equivariance leads to collapsed landmark solutions as discussed

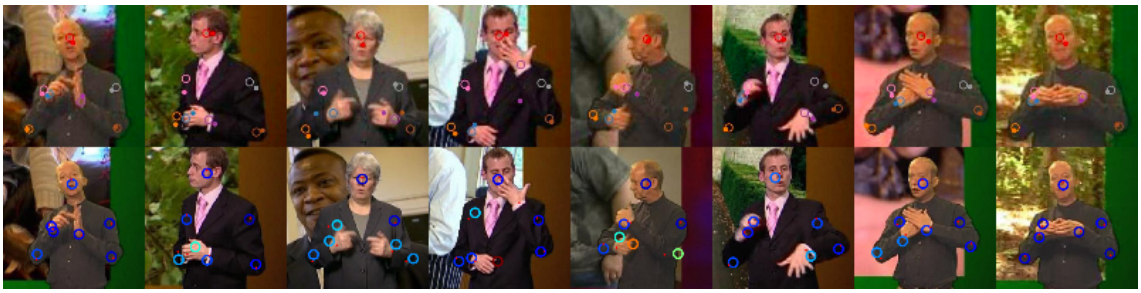


Figure 1.9: Comparison of regression results of our method (bottom rows) to [5] (top rows) on BBC POSE. For visualization by Jakab *et al*. (from their paper) ground truth is in circles and the corresponding regression in the same color. For our visualization the red dots mark the ground truth, the colored circles the regressed locations. The color coding is in terms of the error w.r.t. the image edge length.

Dataset # Landmarks	Cat Head 20
full model	9.30
w/o $\mathcal{L}_{\text{equiv}}$	11.32
w/o $\mathcal{L}_{\text{rec}}$	35.0
w/o appearance transform	12.46
w/o shape transform	14.72

Table 1.8: Ablation studies on Cat Head dataset. We ablate the reconstruction loss $\mathcal{L}_{\text{rec}}$, equivariance loss $\mathcal{L}_{\text{equiv}}$, the color augmentation and the transformations

in [4]. Training our model without color augmentations or appearance crossing between image pairs (i.e. the crossing task) weakens, respectively neglects the disentanglement of appearance and shape and hence the performance of our model significantly. Note that without the crossing task our models performs on par with Zhang *et al*. [4], indicating, that this novel task could be explaining overall performance gain w.r.t. [4]. Leaving out the explicit equivariance leads to the smallest drop in performance. This is not surprising, as equivariance is implicitly also enforced in the crossing framework.

1.4.4 Crucial role of Transformations

The proposed method enables to abstract away object appearance from shape. Despite the multifarious challenges in the diverse range of datasets, the method is able to learn a dedicated part representation for shape. We compare to other approaches and reach state-of-the-art performance on the task of regressing human-annotated landmarks from the part representation. The key difference to the most competitive approach [4] is the emphasis on disentanglement via a crossed reconstruction with shape and appearance transformations. Enforcing disentanglement via targeted transformations enhances the shape representation in two ways: *(i)* it asserts that no appearance information is encoded in the shape representation and vice versa and *(ii)* it requires visual features to be equivariant under a spatial transformation. With regards to the considerations earlier, the crucial role of the transformations are to be expected, as they enable to reach a *disentanglement*.

2 Bibliography

[1] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 2, 4, 6, 7, 12

[2] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008. 4

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

[4] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018. 4, 5, 6, 7, 8, 12, 13, 15

[5] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018. 4, 5, 6, 7, 13, 14

[6] 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, 6, 13

[7] 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. 5

[8] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 6

[9] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 6, 13

[10] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016. 6, 12

[11] 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. 8