Unite the People: Closing the Loop Between 3D and 2D Human Representations

Christoph Lassner[1,2]
classner@tue.mpg.de

Javier Romero[3,*]
javier.romero@bodylabs.com

Martin Kiefel[2]
mkiefel@tue.mpg.de

Federica Bogo[4,*]
febogo@microsoft.com

Michael J. Black[2]
black@tue.mpg.de

Peter V. Gehler[5,*]
pgehler@tue.mpg.de

[1]Bernstein Center for Computational Neuroscience, Tübingen, Germany
[2]MPI for Intelligent Systems, Tübingen, Germany
[3]Body Labs Inc., New York, United States   [4]Microsoft, Cambridge, UK
[5]University of Würzburg, Germany

Abstract

3D models provide a common ground for different representations of human bodies. In turn, robust 2D estimation has proven to be a powerful tool to obtain 3D fits "in-the-wild". However, depending on the level of detail, it can be hard to impossible to acquire labeled data for training 2D estimators on large scale. We propose a hybrid approach to this problem: with an extended version of the recently introduced SMPLify method, we obtain high quality 3D body model fits for multiple human pose datasets. Human annotators solely sort good and bad fits. This procedure leads to an initial dataset, UP-3D, with rich annotations. With a comprehensive set of experiments, we show how this data can be used to train discriminative models that produce results with an unprecedented level of detail: our models predict 31 segments and 91 landmark locations on the body. Using the 91 landmark pose estimator, we present state-of-the art results for 3D human pose and shape estimation using an order of magnitude less training data and without assumptions about gender or pose in the fitting procedure. We show that UP-3D can be enhanced with these improved fits to grow in quantity and quality, which makes the system deployable on large scale. The data, code and models are available for research purposes.

1. Introduction

Teaching computers to recognize and understand humans in images and videos is a fundamental task of computer vision. Different applications require different trade-offs between fidelity of the representation and inference complexity. This led to a wide range of parameterizations for human bodies and corresponding prediction methods ranging from bounding boxes to detailed 3D models.
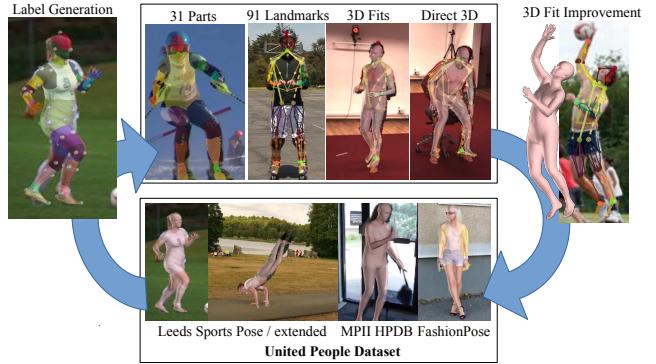

Figure 1: **Lower row:** validated 3D body model fits on various datasets form our initial dataset, *UP-3D*, and provide labels for multiple tasks. **Top row:** we perform experiments on semantic body part segmentation, pose estimation and 3D fitting. Improved 3D fits can extend the initial dataset.

Learning-based algorithms, especially convolutional neural networks (CNNs), are the leading methods to cope with the complexity of human appearance. Their representational power has led to increasingly robust algorithms for bounding box detection [10], keypoint detection [19, 33, 43] and body part segmentation [7, 15, 44]. However, they are usually applied in isolation on separate datasets and independent from the goal of precise 3D body estimation. In this paper we aim to overcome this separation and "unite the people" of different datasets and for multiple tasks. With this strategy, we attack the main problem of learning-based approaches for complex body representations: the lack of data. While it is feasible to annotate a small number of keypoints in images (*e.g.*, 14 in the case of the MPII-HumanPose dataset [1]), scaling to larger numbers quickly becomes impractical and prone to annotation inconsistency. The same is true for semantic segmentation annotations: most datasets provide labels for only a few body parts.

In this paper, we aim to develop a self-improving, scalable method that obtains high-quality 3D body model fits for 2D images (see Fig. 1 for an illustration). To form an initial dataset of 3D body fits, we use an improved version of the recently developed SMPLify method [4] that elevates 2D keypoints to a full body model of pose and shape. A more robust initialization and an additional fitting objective allow us to apply it on the ground truth keypoints of the standard human pose datasets; human annotators solely sort good and bad fits.

This semi-automatic scheme has several advantages. The required annotation time is greatly reduced (Sec. 3.3). By projecting surfaces (Sec. 4.1) or keypoints (Sec. 4.2) from the fits to the original images, we obtain consistent labels while retaining generalization performance. The rich representation and the flexible fitting process make it easy to integrate datasets with different label sets, *e.g.*, a different set of keypoint locations.

Predictions from our 91 keypoint model improve the 3D model fitting method that generated the annotations for training the keypoint model in the first place. We report state-of-the art results on the HumanEva and Human3.6M datasets (Sec. 4.3). Further, using the 3D body fits, we develop a random forest method for 3D pose estimation that runs orders of magnitudes faster than SMPLify (Sec. 4.4).

The improved predictions from the 91 landmark model increase the ratio of high quality 3D fits on the LSP dataset by 9.3% when compared to the fits using 14 keypoint *ground truth* locations (Sec. 5). This ability for self-improvement together with the possibility to easily integrate new data into the pool make the presented system deployable on large scale. Data, code and models are available for research purposes on the project homepage at http://up.is.tuebingen.mpg.de/.

2. Related Work

Acquiring human pose annotations in 3D is a long-standing problem with several attempts from the computer vision as well as the 3D human pose community.

The classical 2D representation of humans are 2D keypoints [1, 6, 23, 24, 39, 40]. While 2D keypoint prediction has seen considerable progress in the last years and could be considered close to being solved [19, 33, 43], 3D pose estimation from single images remains a challenge [4, 37, 45].

Bourdev and Malik [5] enhanced the H3D dataset from 20 keypoint annotations for 1,240 people in 2D with relative 3D information as well as 11 annotated body part segments.

In contrast, the HumanEva [42] and Human3.6M [21] datasets provide very accurate 3D labels: they are both recorded in motion capture environments. Both datasets have high fidelity but contain only a very limited level of diversity in background and person appearance. We evaluate the 3D human pose estimation performance on both.

Recent approaches target 3D pose ground truth from natural scenes, but either rely on vision systems prone to failure [11] or inertial suits that modify the appearance of the body and are prone to motion drift [45].

Body representations beyond 3D skeletons have a long history in the computer vision community [17, 31, 32, 36]. More recently, these representations have taken new popularity in approaches that fit detailed surfaces of a body model to images [4, 14, 16, 26, 45]. These representations are more tightly connected to the physical reality of the human body and the image formation process.

One of the classic problems related to representations of the extent of the body is body part segmentation. Fine-grained part segmentation has been added to the public parts of the VOC dataset [12] by Chen et al. [8]. Annotations for 24 human body parts and also part segments for all VOC object classes, where applicable, are available. Even though hard to compare, we provide results on the dataset. The Freiburg Sitting People dataset [34] consists of 200 images with 14 part segmentation and is tailored towards sitting poses. The ideas by Shotton et al. [41] for 2.5D data inspired our body part representation. Relatively simple methods have proven to achieve good performance in segmentation tasks with "easy" backgrounds like Human80k, a subset of Human3.6M [20].

Following previous work on cardboard people [25] and contour people [13], an attempt to work towards an intermediate-level person representation is the JHMDB dataset and the related labeling tool [22]. It relies on 'puppets' to ease the annotation task, while providing a higher level of detail than solely joint locations.

The attempt to unify representations for human bodies has been made mainly in the context of human kinematics [2, 30]. In their work, a rich representation for 3D motion capture marker sets is used to transfer captures to different targets. The setup of markers to capture not only human motion but also shape has been explored by Loper et al. [29] for motion capture scenarios. While they optimized the placement of markers for a 12 camera setup, we must ensure that the markers disambiguate pose and shape from a single view. Hence, we use a denser set of markers.

3. Building the Initial Dataset

Our motivation to use a common 3D representation is to (1) map many possible representations from a variety of datasets to it, and (2) generate detailed and consistent labels for supervised model training from it.

We argue that the use of a full human body model with a prior on shape and pose is necessary: without the visualization possibilities and regularization, it may be impossible to create sufficiently accurate annotations for small body parts. However, so far, no dataset is available that provides human body model fits on a large variety of images.

To fill this gap, we build on a set of human pose datasets with annotated keypoints. SMPLify [4] presented promising results for automatically translating these into 3D body model fits. This helps us to keep the human involvement to a minimum. With strongly increasing working times and levels of label noise for increasingly complex tasks, this may be a critical decision to create a large dataset of 3D body models.

3.1. Improving Body Shape Estimation

In [4], the authors fit the pose and shape parameters of the SMPL [27] body model to 2D keypoints by minimizing an objective function composed of a data term and several penalty terms that represent priors over pose and shape. However, the connection length between two keypoints is the only indicator that can be used to estimate body shape. Our aim is to match the shape of the body model as accurately as possible to the images, hence we must incorporate a shape objective in the fitting.

The best evidence for the extent of a 3D body projected on a 2D image is encoded by its silhouette. We define the silhouette to be the set of all pixels belonging to a body's projection. Hence, we add a term to the original SMPLify objective to prefer solutions for which the image silhouette, S, and the model silhouette, $\hat{S}$, match.

Let $M(\vec{\theta}, \vec{\beta}, \vec{\gamma})$ be a 3D mesh generated by a SMPL body model with pose, $\vec{\theta}$, shape, $\vec{\beta}$, and global translation, $\vec{\gamma}$. Let $\Pi(\cdot, K)$ be a function that takes a 3D mesh and projects it into the image plane given camera parameters K, such that $\hat{S}(\vec{\theta}, \vec{\beta}, \vec{\gamma}) = \Pi(M(\vec{\theta}, \vec{\beta}, \vec{\gamma}))$ represents the silhouette pixels of the model in the image. We compute the bi-directional distance between S and $\hat{S}(\cdot)$

$$
\begin{aligned}
E_S(\vec{\theta}, \vec{\beta}, \vec{\gamma}; S, K) = &\sum_{\vec{x} \in \hat{S}(\vec{\theta}, \vec{\beta}, \vec{\gamma})} \text{dist}(\vec{x}, S)^2 \\
&+ \sum_{\vec{x} \in S} \text{dist}(\vec{x}, \hat{S}(\vec{\theta}, \vec{\beta}, \vec{\gamma})), \quad (1)
\end{aligned}
$$

where $\text{dist}(\vec{x}, S)$ denotes the absolute distance from a point $\vec{x}$ to the closest point belonging to the silhouette S.

The first term in Eq. (1) computes the distance from points of the projected model to a given silhouette, while the second term computes the distance from points in the silhouette to the model. We find that the second term is noisier and use the plain L1 distance to measure its contribution to the energy function while we use the squared L2 distance to measure the contribution of the first. We optimize the overall objective including this additional term using OpenDR [28], just as in [4].

Whereas it would be possible to use an automatic segmentation method to provide foreground silhouettes, we decided to involve human annotators for reliability. We also asked for six body part segmentation that we will use

Dataset	Foreground	6 Body Parts	AMT hours logged
LSP [23]	1000 train, 1000 test	1000 train, 1000 test	361h foreground,
LSP-extended [24]	10000 train	0	131h parts
MPII-HPDB [1]	13030 train, 2622 test	0	729h

Table 1: Logged AMT labelling times. The average foreground labeling task was solved in 108s on the LSP and 168s on the MPII datasets respectively. Annotating the segmentation for six body parts took on average more than twice as long as annotating foreground segmentation: 236s.

Figure 2: Examples for six part segmentation ground truth. White areas mark inconsistencies with the foreground segmentation and are ignored.

in Sec. 4 for evaluation. We built an interactive annotation tool on top of the *Opensurfaces* package [3] to work with Amazon Mechanical Turk (AMT). To obtain image-consistent silhouette borders, we use the interactive Grabcut algorithm [38]. Workers spent more than 1,200 hours on creating the labels for the LSP [23, 24] datasets as well as the single-person part of the MPII-HumanPose [1] dataset (see Tab. 1). There is an increase in average annotation time of more than a factor of two comparing annotation for foreground labels and six body part labels. This provides a hint on how long annotation for a 31 body part representation could take. Examples for six part segmentation labels are provided in Fig. 2.

3.2. Handling Noisy Ground Truth Keypoints

The SMPLify method is especially vulnerable to missing annotations of the four torso joints: it uses their locations for an initial depth guess, and convergence deteriorates if this guess is of poor quality.

Finding a good depth initialization is particularly hard due to the foreshortening effect of the perspective projection. However, since we know that only a shortening but no lengthening effect can occur, we can find a more reliable person size estimate $\hat{\theta}$ for a skeleton model with k connections:

$$
\hat{\theta} = \mathbf{x}_i \cdot \arg\max_y f_i(y), \quad i = \arg\max_{j=1,\ldots,k} \mathbf{x}_j, \quad (2)
$$

where f_i is the distribution over ratios of person size to the length of connection $\mathbf{x}_i$. Since this is a skewed distribution, we use a corrected mean to find the solution of the $\arg\max$ function and obtain a person size estimate. This turns out to be a simple, yet robust estimator.

LSP [23]	LSP extended [24]	MPII-HP [1]	FashionPose [9]
45%	12%	25%	23%

Table 2: Percentages of accepted fits per dataset. The addition of the FashionPose dataset is discussed in Sec. 4.2.

3.3. Exploring the Data

With the foreground segmentation data and the adjustments described in the preceding sections, we fit the SMPL model to a total of 27,652 images of the LSP, LSP-extended, and MPII-HumanPose datasets. We use only people marked with the 'single person' flag in MPII-HumanPose to avoid instance segmentation problems. We honor the train/test splits of the datasets and keep images from their test sets in our new, joined test set.

In the next step, human annotators[1] selected the fits where rotation and location of body parts largely match the image evidence. For this task, we provide the original image, as well as four perspectives of renderings of the body. Optionally, annotators can overlay rendering and image. These visualizations help to identify fitting errors quickly and reduce the labeling time to ∼12s per image. The process uncovered many erroneously labeled keypoints, where mistakes in the 3D fit were clear to spot, but not obvious in the 2D representation. We excluded head and foot rotation as criteria for the sorting process. There is usually not sufficient information in the original 14 keypoints to estimate them correctly. The resulting ratios of accepted fits can be found in Tab. 2.

Even with the proposed, more robust initialization term, the ratio of accepted fits on the LSP-extended dataset remains the lowest. It has the highest number of missing keypoints of the four datasets, and at the same time the most extreme viewpoints and poses. On the other hand, the rather high ratio of usable fits on the LSP dataset can be explained with the clean and complete annotations.

The validated fits form our initial dataset with 5,569 training images (of which we use a held-out validation set of 1,112 images in our experiments) and 1,208 test images. We denote this dataset as UPI-3D (**U**nited**P**eople in 3D with an added 'I' for "Initial"). To be able to clearly reference the different label types in the following sections, we add an 'h' to the dataset name when referring to labels from human annotators.

Consistency of Human Labels The set of curated 3D fits allows us to assess the distribution of the human-provided labels by projecting them to the UPI-3D bodies. We did this for both, keypoints and body part segments. Visualizations can be found in Fig. 3.

---

[1]For this task, we did not rely on AMT workers, but only on few experts in close collaboration to maintain consistency.

While keypoint locations in Fig. 3a in completely non-matching areas of the body can be explained by self-occlusion, there is a high variance in keypoint locations around joints. It must be taken into account that the keypoints are projected to the body surface, and depending on person shape and body part orientation some variation can be expected. Nevertheless, even for this reduced set of images with very good 3D fits, high variance areas, *e.g.*, around the hip joints, indicate labeling noise.

The visualization in Fig. 3b shows the density of part types for six part segmentation with the segments head, torso, left and right arms and left and right legs. While the head and lower parts of the extremities resemble distinct colors, the areas converging to brown represent a mixture of part annotations. The brown tone on the torso is a clear indicator for the frequent occlusion by the arms. The area around the hips is showing a smooth transition from torso to leg color, hinting again at varying annotation styles.

4. Label Generation and Learning

In a comprehensive series of experiments, we analyze the quality of labels generated from UPI-3D. We focus on labels for well-established tasks, but highlight that the generation possibilities are not limited to them: all types of data that can be extracted from the body model can be used as labels for supervised training. In our experiments, we move from surface (segmentation) prediction over 2D- to 3D-pose and shape estimation to a method for predicting 3D body pose and shape directly from 2D landmark positions.

4.1. Semantic Body Part Segmentation

We segment the SMPL mesh into 31 regions, following the segmentation into semantic parts introduced in [41] (for a visualization, see Fig. 3d). We note that the Kinect tracker works on 2.5D data while our detectors only receive 2D data as input. We deliberately did not make any of our methods for data collection or prediction dependent on 2.5D data to retain generality. This way, we can use it on outdoor images and regular 2D photo datasets. The **S**egmentation dataset UPI-S31 is obtained by projecting the segmented 3D mesh posed on the 6,777 images of UPI-3D.

Following [7], we optimize a multiscale ResNet101 on a pixel-wise cross entropy loss. We train the network on size-normalized, cutout images, which could in a production system be provided by a person detector. Following best practices for CNN training, we use a validation set to determine the optimal number of training iterations and the person size, which is around 500 pixels. This high resolution allows the CNN to reliably predict small body parts. In this challenging setup, we achieve an intersection over union (IoU) score of 0.4432 and an accuracy of 0.9331. Qualitative results on five datasets are shown in Fig. 4a.
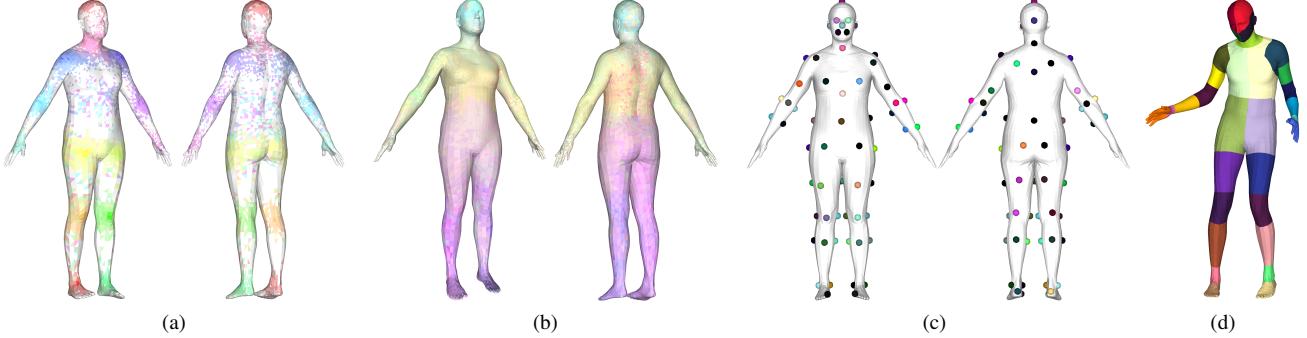
Figure 3: Density of human annotations on high quality body model fits for **(a)** keypoints and **(b)** six part segmentation in front and back views. Areas of the bodies are colored with (1) hue according to part label, and (2) saturation according to frequency of the label. Keypoints on completely 'wrong' bodyparts are due to self-occlusion. The high concentration of 'head' labels in the nose region originates from the FashionPose dataset, where the 'head' keypoint is placed on the nose. The segmentation data originates solely from the six part segmentation labels on the LSP dataset. (Must be viewed in color.) **(c)** Placement of the 91 landmarks (left: front, right: back). **(d)** Segmentation for generating the 31 part labels.

The overall performance is compelling: even the small segments around the joints are recovered reliably. Left and right sides of the subjects are identified correctly, and the four parts of the head provide an estimate of head orientation. The average IoU score is dominated by the small segments, such as the wrists.

The VOC part dataset is a hard match for our predictor: instead of providing instances of people, it consists of entire scenes, and many people are visible at small scale. To provide a comparison, we use the instance annotations from the VOC-Part dataset, cut out samples and reduce the granularity of our segmentation to match the widely used six part representation. Because of the low resolution of many displayed people and extreme perspectives with, *e.g.*, only a face visible, the predictor often only predicts the background class on images not matching our training scheme. Still, we achieve an IoU score of 0.3185 and 0.7208 accuracy over the entire dataset without finetuning.

Additional examples from the LSP, MPII-HumanPose, FashionPose, Fashionista, VOC, HumanEva and Human3.6M datasets are shown in the supplementary material available on the project homepage[2]. The model has not been trained on any of the latter four, but the results indicate good generalization behavior. We include a video to visualize stability across consecutive frames.

4.2. Human Pose Estimation

With the 3D body fits, we can not only generate consistent keypoints on the human skeleton but also on the body surface. For the experiments in the rest of this paper, we designed a 91-landmark[3] set to analyze a dense keypoint set.

---

[2]http://up.is.tuebingen.mpg.de/

[3]We use the term 'landmark' to refer to keypoints on the mesh surface to emphasize the difference to the so-far used term 'joints' for keypoints located inside of the body.

We distributed the landmarks according to two criteria: disambiguation of body part configuration and estimation of body shape. The former requires placement of markers around joints to get a good estimation of their configuration. To satisfy the latter, we place landmarks in regular intervals around the body to get an estimate of spatial extent independent of the viewpoint. We visualize our selection in Fig. 3c and example predictions in Fig. 4b.

In the visualization of predictions, we show a subset of the 91 landmarks and only partially connect the displayed ones for better interpretability. The core 14 keypoints describing the human skeleton are part of our selection to describe the fundamental pose and maintain comparability with existing methods.

We use a state-of-the-art DeeperCut CNN [19] for our pose-related experiments, but believe that using other models such as Convolutional Pose Machines [43] or Stacked Hourglass Networks [33] would lead to similar findings.

To assess the influence of the quality of our data and the difference of the loss function for 91 and 14 keypoints, we train multiple CNNs: (1) using all human labels but on our (smaller) dataset for 14 keypoints (UPI-P14h) and (2) on the dense 91 landmarks from projections of the SMPL mesh (UPI-P91). Again, models are trained on size-normalized crops with cross-validated parameters. We include the performance of the original DeeperCut CNN, which has been trained on the full LSP, LSP-extended and MPII-HumanPose datasets (in total more than 52,000 people) in the comparison with the models being trained on our data (in total 5,569 people). The results are summarized in Tab. 3. Even though the size of the dataset is reduced by nearly an order of magnitude, we maintain high performance compared to the original DeeperCut CNN. Comparing the two models trained on the same amount of data, we find that the model trained on the 91 landmarks from
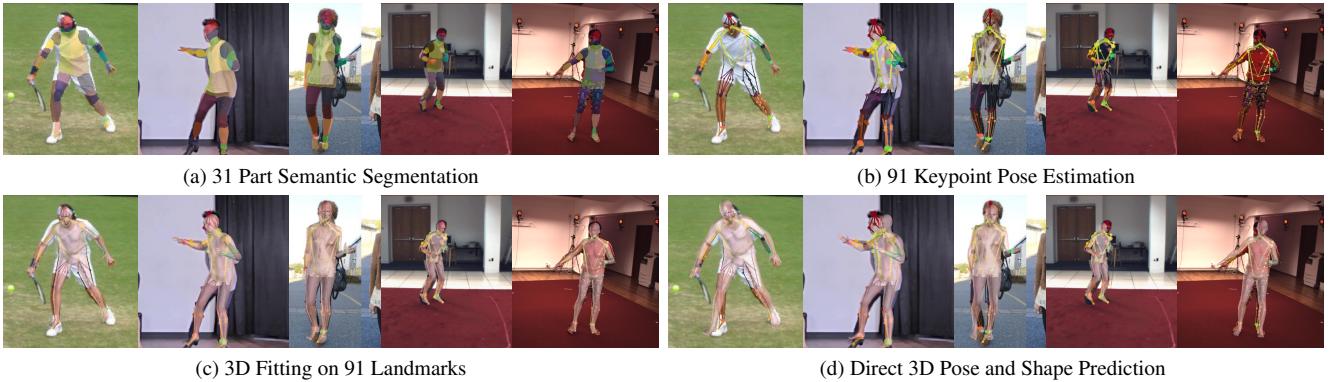
5

(a) 31 Part Semantic Segmentation

(b) 91 Keypoint Pose Estimation

(c) 3D Fitting on 91 Landmarks

(d) Direct 3D Pose and Shape Prediction

Figure 4: Results from various methods trained on labels generated from the UP-3D dataset.

PCK@0.2	UPI-P14h	UPI-P14	UPI-P91
DeeperCut CNN [19]	**93.45**	92.16	NA
Ours (trained on UPI-P14h)	89.11	87.36	NA
Ours (trained on UPI-P91)	**91.15**	**93.24**	**93.54**

Table 3: Pose estimation results. Even though the DeeperCut CNN has been trained on almost by factor ten more examples, our model remains competitive. The third row shows the results of our 91 landmark model evaluated on the 14 core keypoints on human, 14 and 91 SMPL generated landmark labels. It outperforms the model trained on the data labeled by humans (row 2 vs. 3, column 1) by more than two score points. Fair comparisons can only be made within offset boxes.

the SMPL data has a notable advantage of nearly six score points on the SMPL labeled data (row 2 vs. 3, column 2). Even when evaluating on the human labeled data, it maintains an advantage of two score points (row 2 vs. 3, column 1). This shows that the synthetic keypoints generalize to the human labels, which we take as an encouraging result.

We provide the third column for giving an impression of the performance of the additional 77 landmarks. When including the additional landmarks in the evaluation, the score rises compared to evaluating on the 14 core keypoints, indicating their overall performance is above average. A direct comparison to the 14 keypoint values is not valid, because the score is averaged over results of differing 'difficulty'.

Integrating a Dataset with a Different Label Set The current state-of-the art pose estimators benefit from training on all human pose estimation datasets with a similar label set. The FashionPose dataset [9] would complement these well, but is annotated with a different set of keypoints: the neck joint is missing and the top head keypoint is replaced by the nose. Due to this difference, it is usually not included in pose estimator training.

Using our framework, we can overcome this difficulty: we adjust the fitting objective by adding the nose to and removing the top-head keypoint from the objective function. We fit the SMPL model to the FashionPose dataset and curate the fits. The additional data enlarges our training set by 1,557 images and test set by 181 images. This forms the full UP-3D dataset, which we use for all remaining experiments.

We train an estimator on the landmarks projected from the full UP-3D dataset. This estimator outperforms the plain DeeperCut CNN with a small margin from 0.897 PCK@0.2 (DeeperCut) to 0.9028 PCK@0.2 (ours) on the full, human labeled FashionPose test set.

4.3. 3D Human Pose Estimation

In this section, we analyze the impact of using the 91 predicted keypoints instead of 14 for the SMPLify 3D fitting method. For the fitting process, we rely solely on the 91 predicted 2D landmarks and no additional segmentation or gender information (in contrast to the SMPLify method as described in [4], where gender information is used for fitting on the 3D datasets). Segmentation information is not required anymore to estimate body extent due to the landmarks on the body surface.

LSP dataset On the LSP dataset, there is no ground truth for 3D body model fitting available. To be independent of biases towards a specific keypoint set, we rely on the acquired six body part segmentation to obtain meaningful performance scores (see Tab. 4).

The six-part manual segmentation annotations consist of head, torso, left and right leg, and left and right arm (see Fig. 2). While this representation is coarse, it provides a good estimate of the overall quality of a fit. It takes into account the body shape and not only keypoints, hence it is a fair judge for pose estimators aiming for slightly different keypoint locations.

	FB Seg. acc., f1	P Seg acc., f1
SMPLify on GT lms.	0.9176, 0.8811	0.8798, 0.6584
SMPLify on GT lms. & GT seg.	**0.9217, 0.8823**	**0.8882, 0.6703**
SMPLify on DeepCut CNN lms. [4]	**0.9189, 0.8807**	**0.8771, 0.6398**
SMPLify on our CNN lms., tr. UPI-P14h	0.8944, 0.8401	0.8537, 0.5762
SMPLify on our CNN lms., tr. UP-P14	0.8952, 0.8475	0.8588, 0.5798
SMPLify on our CNN lms., tr. UP-P91	**0.9099, 0.8619**	**0.8732, 0.6164**
DP from 14 landmarks	0.8649, 0.7915	0.8223, 0.4957
DP from 91 landmarks	0.8666, 0.7993	0.8232, 0.5102
DP from 14 lms., rotation opt.	0.8742, 0.8102	0.8329, 0.5222
DP from 91 lms., rotation opt.	**0.8772, 0.8156**	**0.8351, 0.5304**

Table 4: Scores of projected body parts of the fitted SMPL model on the full LSP test set six part human labels (*landmarks* is abbreviated to *lms.*). Fair comparisons can only be made within offset boxes. 'DP' refers to 'Direct Prediction' (see Sec. 4.4). The landmarks for these experiments are always predictions from our CNN trained on UP-P91.

Unsurprisingly, the segmentation scores of the SMPLify method improve when the segmentation term (*c.f.* Sec. 3.1) is added. Due to the longer-trained pose estimator, SMPLify as presented in [4] still has an overall advantage on the LSP dataset (compare rows three and four).

Training on our generated data for 14 joints and then using SMPLify improves the scores (compare lines four and five) thanks to cleaner data and better correspondence of keypoints and SMPL skeleton. Using our 91 landmark model gives a large performance boost of 3.6 f1 score points. We do not reach the performance of the fits performed on the DeepCut CNN [35] predictions, largely because of few extreme poses that our pose estimator misses with a large influence on the final average score.

HumanEva and Human3.6M Datasets We evaluate 3D fitting on the HumanEva and Human3.6M datasets where 3D ground truth keypoints are available from a motion capture system. We follow the evaluation protocol of SMPLify [4] to maintain comparability, except for subsampling to every 5th frame. This still leaves us with a framerate of 10Hz which does not influence the scores. We do this solely due to practical considerations, since the SMPLify fitting to 91 landmarks can take up to twice as long as fitting to 14 keypoints. We provide a summary of results in Tab. 5.

We do not use actor or gender specific body models, but one hybrid human model, and rely on the additional landmarks for shape inference. This makes the approach fully automatic and deployable to any sequence without prior knowledge. Even with these simplifications and a magnitude fewer training examples for our pose estimator, we achieve an improvement of 5.4mm on average on the HumanEva dataset and an improvement of 1.6mm on average on the Human3.6M dataset (4th versus 5th row).

	HumanEva	Human3.6M
Zhou et al. [46]	110.0	106.7
DP from 91 landmarks	93.5	93.9
SMPLify on DeepCut CNN lms. [4]	79.9	82.3
SMPLify on our CNN lms., tr. UPI-P14h	81.1	96.4
SMPLify on our CNN lms., tr. UP-P14	79.4	90.9
SMPLify on our CNN lms., tr. UP-P91	**74.5**	**80.7**

Table 5: Average error over all joints in 3D distance (mm).

The use of a pose estimator trained on the full 91 keypoint dataset UP-P91 improves SMPLify even more. Compared to the baseline model trained on UPI-P14h, performance improves by 6.6mm on the simpler HumanEva dataset, and by 15.7mm on Human3.6M. Even when training a 14 keypoint pose estimator, the higher consistency of our generated labels helps to solve this task, which becomes apparent comparing lines four and five.

4.4. Direct 3D Pose and Shape Prediction

The 91 landmark predictions in 2D enable a human observer to easily infer the 3D shape and pose of a person: the keypoints on the body surface provide a good hint to estimate person shape and in combination with the skeleton orientation and pose can usually be identified (*c.f.* Fig. 4b). This observation inspired us to explore the limits of a predictor for the 3D body model parameters directly from the 2D keypoint input. For this purpose, we use the 3D poses and shapes from the UP-3D dataset to sample projected landmarks with the full 3D parameterization of SMPL as labels. We move a virtual 'camera' for every pose on 5 elevations to 36 positions around the 3D model to enhance the training set. On this data, we experimented with multi-layer perceptrons as well as Decision Forests. We preferred the latter regressor, since Decision Forests are less susceptible to noise. We train a separate forest to predict the axis-angle rotation vector for each of the 24 SMPL joints, as well as one to predict the depth. The input landmark positions are normalized w.r.t. position and scale to improve generalization. We experimented with distance-based features and dot-product features from the main skeleton connections, but these were not as robust as plain 2D image coordinates. It turned out to be critical to use full rotation matrices as regression targets: the axis-angle representation has discontinuities, adding noise to the loss function.

One Decision Forest predicts pose or shape in 0.13s[4]. The predictions of all forests are independent, which means that the full pose and shape prediction can be obtained in between one and two orders of magnitudes faster than with SMPLify. This could allow the use of 3D pose and shape estimation for video applications, *e.g.*, action recognition.

---

[4]For all timings, a test system with a 3.2Ghz six core processor and an NVIDIA GeForce GTX970 has been used.

Whereas the 3D model configuration does not always match the image evidence (see Fig. 4d), it recovers the rough pose. We provide scores in Tab. 4 and Tab. 5 with the name 'DP' (**D**irect **P**rediction). We additionally add the scores for a hybrid version, for which we predict pose and shape using Decision Forests and take few optimization steps to make the global rotation of the body model match the image evidence (with varying runtime depending on the initialization, but less than one second on our data).

The difference between the full optimization on the LSP dataset in f1 score is 0.1062 for the 91 landmark based method and reduces with rotation optimization to 0.086. On the 3D datasets, the direct prediction method outperforms all optimization based methods except for SMPLify that runs in the order of tens of seconds.

Together with our ResNet101-based CNN model, it is possible to predict a full 3D body model configuration from an image in 0.378s. The pose-predicting CNN is the computational bottleneck. Because our findings are not specific to a CNN model, we believe that by using a speed-optimized CNN, such as SqueezeNet [18], and further optimizations of the direct predictor, the proposed method could reach real-time speed.

5. Closing the Loop

With the improved results for 3D fitting, which helped to create the dataset of 3D body model fits in the first place, a natural question is, whether the improved fitting method helps to enlarge the dataset.

We ran SMPLify on the 91 landmark predictions from our pose estimator and again asked human annotators to rate the 3D fits on all LSP images that were not accepted for our initial dataset. Of the formerly unused 54.75% of the data (1095 images), we found an improvement in six body part segmentation f1 score for 308 images (*c.f.* Fig. 5a). We show three example images with high improvement in f1 score in Fig. 5, (b) to (d): improvement due to left-right label noise, depth ambiguity and perspective resolution compared to fits on the 14 ground truth keypoints. Human annotators accepted additional 185 images, which is an improvement of 20% over the number of accepted initial fits and an absolute improvement of 9.3% in accepted fits of the LSP dataset.

The most common reasons for improvement are (1) noisy annotations, (2) better perspective resolution and (3) the better match of keypoints to the SMPL skeleton. An even higher ratio of improvement can be expected for the datasets with more annotation noise, such as the LSP-extended and MPII-HumanPose datasets. This enlarged set of data could be used to again train estimators and continue iteratively.
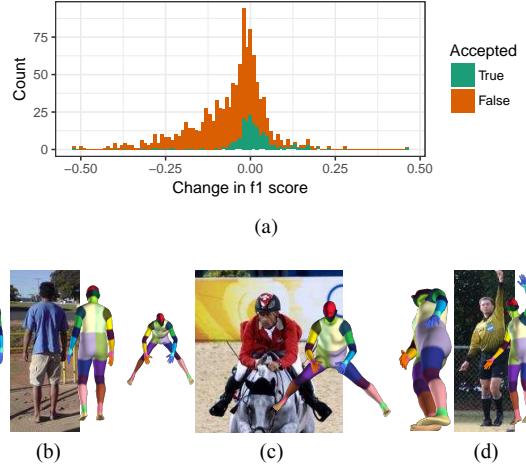

Figure 5: Improvement of body model fits on the unused part of the LSP dataset. Compared are fits to 91 predicted keypoints vs. fits to 14 ground truth keypoints. **(a)**: histogram of change in f1 score of projected six body part segmentation agreement with human annotator ground truth. Green color indicates that the formerly unaccepted fit to the 14 ground truth keypoints is accepted as valid when performed with the 91 predicted keypoints. For each image triple in **(b)**, **(c)**, **(d)**: left: SMPLify fit to the 14 ground truth keypoints, right: fit to the predicted 91 landmarks from our predictor.

6. Discussion

With the presented method and dataset, we argue for a holistic view on human related prediction tasks. By improving the representation of humans we could integrate datasets with different annotations and approach established tasks at a new level of detail. The presented results include high fidelity semantic body part segmentation into 31 parts and 91 landmark human pose estimation. This sets a new mark in terms of levels of detail that previous work did not reach. At the same time, it helps to improve the state-of-the art for 3D human pose estimation on the two standard benchmark datasets HumanEva and Human3.6M.

We present a regression tree model that predicts the 3D body configuration from 2D keypoints directly. This method runs orders of magnitude faster than optimization based methods. This direct prediction captures the overall pose from simple 2D input reasonably well and we are optimistic that it can be scaled to reach near real-time performance. We show that the improved 3D fitting method allows more good fits that enlarge the training set. Here, we only took one iteration but are confident that a system that iterates over the two generative and discriminative stages can be deployed on large scale to continuously learn and improve with very limited human feedback.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2, 3, 4

[2] P. Azad, T. Asfour, and R. Dillmann. Toward an unified representation for imitation of human motion on humanoids. In *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 2558–2563, April 2007. 2

[3] S. Bell, P. Upchurch, N. Snavely, and K. Bala. OpenSurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics (SIGGRAPH)*, 32(4), 2013. 3

[4] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 561–578. Springer, 2016. 2, 3, 6, 7

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *Proc. of the International Conference on Computer Vision (ICCV)*, sep 2009. 2

[6] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman. Personalizing human video pose estimation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[7] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proc. of the 2016 IEEE Conference on Computer Vision (CVPR)*, 2016. 1, 4

[8] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[9] M. Dantone, J. Gall, C. Leistner, and L. van Gool. Body parts dependent joint regressors for human pose estimation in still images. *Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 4, 6

[10] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 34(4):743–761, 2012. 1

[11] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3810–3818. IEEE, 2015. 2

[12] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal on Computer Vision (IJCV)*, 88(2):303–338, June 2010. 2

[13] O. Freifeld, A. Weiss, S. Zuffi, and M. J. Black. Contour people: A parameterized model of 2D articulated human shape. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 639–646. IEEE, June 2010. 2

[14] P. Guan, A. Weiss, A. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *Proc. of the International Conference on Computer Vision (ICCV)*, pages 1381–1388, 2009. 2

[15] B. Hariharan, P. Arbelez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 447–456, June 2015. 1

[16] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormhlen, and H. P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1823–1830, June 2010. 2

[17] D. Hogg. Model-based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5 – 20, 1983. 2

[18] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *arXiv:1602.07360*, 2016. 8

[19] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 5, 6

[20] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3D human pose estimation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1661–1668, 2014. 2

[21] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, July 2014. 2

[22] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pages 3192–3199, Sydney, Australia, Dec. 2013. IEEE. 2

[23] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. of the British Machine Vision Conference (BMVC)*, 2010. doi:10.5244/C.24.12. 2, 3, 4

[24] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1465–1472, 2011. 2, 3, 4

[25] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *Proc. of the 2nd Int. Conf. on Automatic Face- and Gesture-Recognition*, pages 38–44, Killington, Vermont, Oct. 1996. 2

[26] S. Li and A. B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Proc. of the Asian Conference on Computer Vision (ACCV)*, pages 332–347. Springer, 2014. 2

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM*

Trans. on Graphics (Proc. SIGGRAPH Asia), 34(6):248:1–248:16, 2015. 3

[28] M. M. Loper and M. J. Black. OpenDR: An approximate differentiable renderer. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 154–169, Cham, 2014. Springer International Publishing. 3

[29] M. M. Loper, N. Mahmood, and M. J. Black. MoSh: Motion and shape capture from sparse markers. *ACM Trans. on Graphics, (Proc. SIGGRAPH Asia)*, 33(6):220:1–220:13, Nov. 2014. 2

[30] C. Mandery, O. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour. The KIT whole-body human motion database. In *Proc. of the International Conference on Advanced Robotics (ICAR)*, pages 329–336, July 2015. 2

[31] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. of the Royal Society of London B: Biological Sciences*, 200(1140):269–294, 1978. 2

[32] K. Nevatia and T. O. Binford. Structured descriptions of complex objects. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2

[33] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 483–499. Springer, 2016. 1, 2, 5

[34] G. L. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox. Deep learning for human part discovery in images. In D. Kragic, A. Bicchi, and A. D. L. 0001, editors, *Proc. of the IEEE International Conference on Robotics and Automation (ICRA) 2016*, pages 1634–1641. IEEE, 2016. 2

[35] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[36] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann, and A. Fitzgibbon. Metric regression forests for correspondence estimation. *International Journal of Computer Vision (IJCV)*, pages 1–13, 2015. 2

[37] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. *Proc. of the European Conference on Computer Vision (ECCV)*, pages 573–586, 2012. 2

[38] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. *ACM Trans. on Graphics (SIGGRAPH)*, August 2004. 3

[39] B. Sapp and B. Taskar. MODEC: Multimodal decomposable models for human pose estimation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[40] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1281–1288. IEEE, 2011. 2

[41] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman,

et al. Efficient human pose estimation from single depth images. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(12):2821–2840, 2013. 2, 4

[42] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1):4–27, Mar. 2010. 2

[43] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5

[44] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human part segmentation with auto zoom net. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2016. 1

[45] S. Zhou, H. Fu, L. Liu, D. Cohen-Or, and X. Han. Parametric reshaping of human bodies in images. *ACM Trans. on Computer Graphics (TOG): Special Issue of ACM SIGGRAPH*, 29(4):126, 2010. 2

[46] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparse representation for 3D shape estimation: A convex relaxation approach. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 7