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

C. Lassner[1,*], J. Romero[3,*], M. Kiefel[1], F. Bogo[4,*], M. J. Black[1], P. V. Gehler[5,*]

[1]Max-Planck Institute for Intelligent Systems, [2]BCCN Tübingen, [3]Bodylabs, [4]Microsoft, [5]University of Würzburg

MAX-PLANCK-GESELLSCHAFT

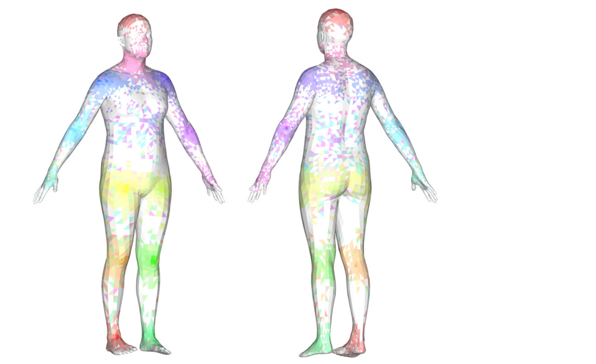
1 Motivation

Goal: highly detailed 2D models and 3D fits of people.

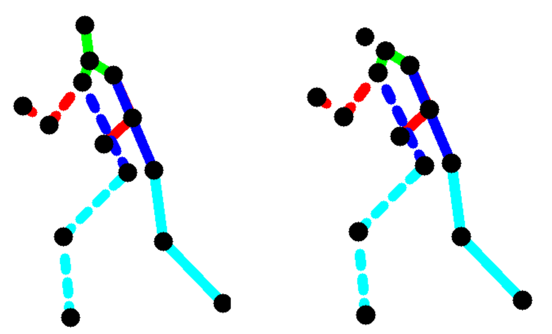
The usually used **14 keypoints / few segments** provide **too little information**.

Challenges:

- **Annotation time** quickly become infeasible: e.g., ~8 min. for 91 keypoints.
- **Inconsistentencies** become more frequent for fine-grained annotations:
- **Different label sets** make it impossible to fuse datasets:



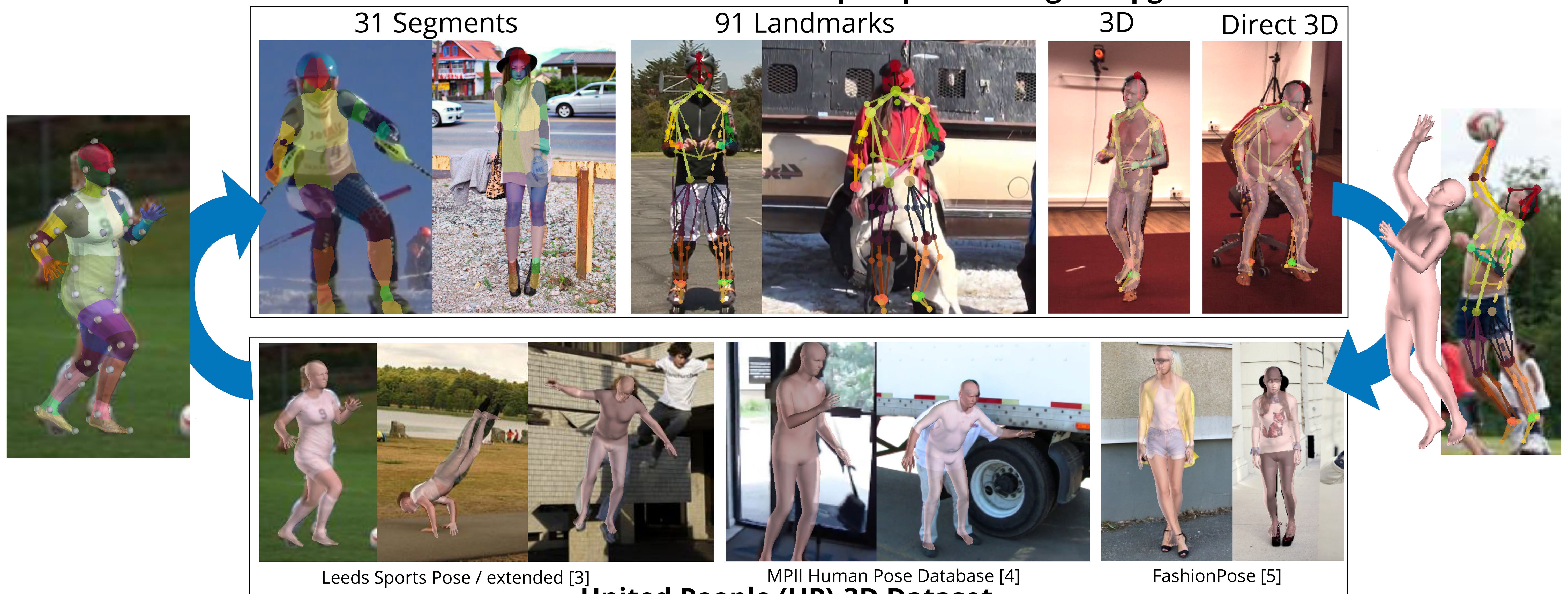
Heatmap for label positins of human annotators proj. to a common 3D body.



Example pose with LSP (left) and FashionPose (right) labels.

Proposed solution:
Use 3D SMPL [1] body fits as common, detailed representation and **iterate** between **improving 3D fits and 2D models** (see center figure).

Datasets and code available at http://up.is.tuebingen.mpg.de.



United People (UP)-3D Dataset

2 Fitting 3D Bodies

Use **segmentation data** to estimate **body extent**.
Extend the energy function of [2] with a silhouette term:

$$\begin{aligned} E(\beta,\theta;K,J_{est},S_{est}) = {} & E_J(\beta,\theta;K,J_{est})\ + && \text{(joint matching)}\\ & E_a(\boldsymbol{\theta}) + && \text{(unnatural pose penalty)}\\ & E_\theta(\boldsymbol{\theta}) + && \text{(pose prior)}\\ & E_{sp}(\boldsymbol{\theta};\boldsymbol{\beta}) + && \text{(spheres / inner penetration)}\\ & E_\beta(\boldsymbol{\beta}) + && \text{(shape regularization)}\\ & \boxed{E_S(\beta,\theta,K,S_{est})} && \text{(silhouette matching)} \end{aligned}$$

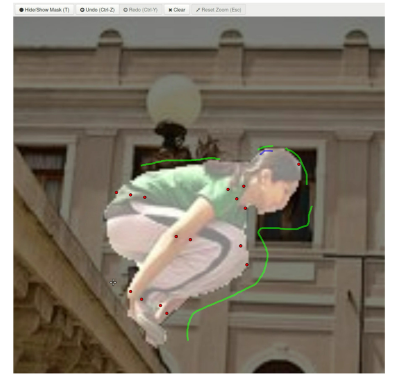
$$\begin{aligned} E_S(\vec{\theta},\vec{\beta},\vec{\gamma};S,K) \quad = \quad & \sum_{\vec{x}\in\hat{S}(\vec{\theta},\vec{\beta},\vec{\gamma})} \operatorname{dist}(\vec{x},S)^2 && \text{(proj. mesh to annotations)}\\ & + \sum_{\vec{x}\in S} \operatorname{dist}(\vec{x},\hat{S}(\vec{\theta},\vec{\beta},\vec{\gamma})),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

Robustify the camera parameter initialization against **missing keypoints**:

$$i = \arg\max_{i=1,\dots,k} \mathbf{x}_i,\ \ \hat{\theta} = \mathbf{x}_i \cdot \arg\max_y f_i(y),$$

3 The Datasets

Foreground-background semantic **segmentation annotations** for the **LSP** [2], **LSP extended** [3] and **MPII Human Pose** [4] (single person) datasets.



Openpose annotation interface with Grabcut labeling support.

Dataset	Foreground	6 Body Parts	AMT hours logged
LSP [3]	1000 train, 1000 test	1000 train, 1000 test	361h foreground, 131h parts
LSP-extended [3]	10000 train	0	
MPII-HPDB [4]	13030 train, 2622 test	0	729h

AMT annotation times for the annotated datasets.

LSP [3]	LSP extended [3]	MPII-HP [4]	FashionPose [5]
45%	12%	25%	23%

Ratio of accepted 3D fits per dataset.

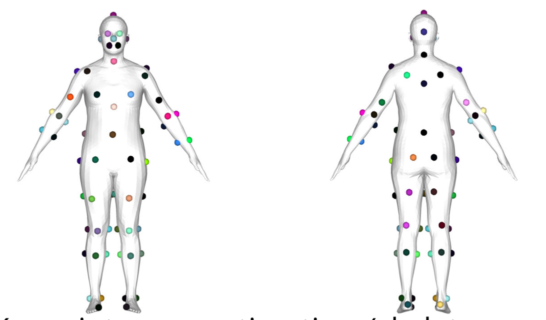
We **fit SMPL** to the **27,652** images and let **human annotators curate** them (+ 7,305 images of FashionPose [5], only to keypoints).

The **curators are trained** and **work in close collaboration** to ensure consistent selection.

6,014 train, 1,112 validation, 1,389 test images with high quality 3D fits (see center figure).

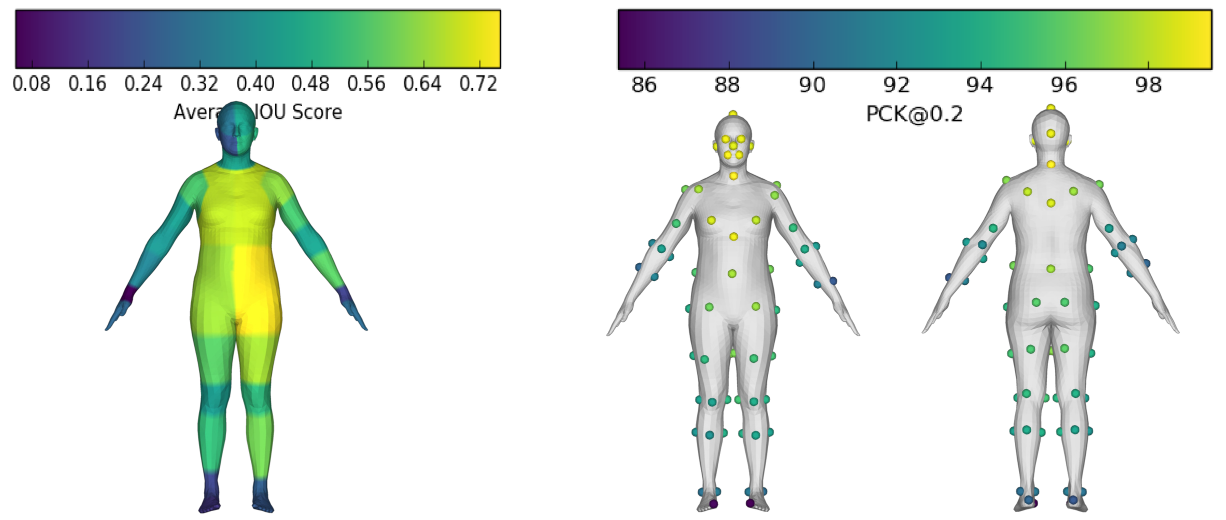
4 2D Appearance Models

Semantic 31 part segmentation (similar to [8]). As predictor, we use a Deeplab-Resnet 101 [6].



91 Keypoint pose estimation (skeleton points not visible on the surface). We use a Deepercut CNN [7] as predictor.

Performance:



5 3D Human Pose Estimation

Energy-based fitting **without silhouette information on 91 keypoints**. Alternatively, use a **regression forest** to regress from **91 keypoints to 3D**.

	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. [2]	**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**

Evaluation results on the six part semantic segmentation data annotated by humans.

PCK@0.2	UPI-P14h	UPI-P14	UPI-P91
DeeperCut CNN [7]	**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**

2D Pose estimation performance.

	HumanEva	Human3.6M
Zhou et al. [9]	110.0	106.7
DP from 91 landmarks	93.5	93.9
SMPLify on DeepCut CNN lms. [2]	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**

3D evaluation results.

6 Closing the Loop

Rerun energy-based **fitting on 91 keypoints (no silhouette term needed)**.

Comparison of results compared to fits to ground truth keypoints + silhouette:

When re-curating, 20% more than the initial accepted fits were rated 'usable'.

7 References

[1] SMPL: A skinned multi-person linear model. M. Loper, N. Mahmood, J. Romero et al. SIGGRAPH 2015.
[2] Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. F. Bogo, A. Kanazawa, C. Lassner et al. ECCV 2016.
[3] Clustered pose and nonlinear appearance models for human pose estimation. S. Johnson and M. Everingham. BMVC 2010.
[4] 2D human pose estimation: New benchmark and state of the art analysis. M. Andriluka, L. Pishchulin, P. V. Gehler, B. Schiele. CVPR 2014.
[5] Body Parts Dependent Joint Regressors for Human Pose Estimation in Still Images. M. Dantone, J. Gall, C. Leistner, L. v. Gool. TPAMI 2014.
[6] Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. L. Chen, G. Papandreou, I. Kokkinos et al. ICLR 2015.
[7] DeeperCut: A Deeper, Stronger, and Faster Multi-Person Pose Estimation Model. E. Insafutdinov, L. Pishchulin, B. Andres et al. CVPR 2016.
[8] Efficient human pose estimation from single depth images. J. Shotton, R. Girshick, A. Fitzgibbon et al. TPAMI 2013.
[9] Sparse representation for 3D shape estimation: A convex relaxation approach. X. Zhou, M. Zhu, S. Leonardos et al. CVPR 2015.

* When this work was performed, Christoph was with BCCN and MPI-IS, Javier with MPI-IS, Federica with MPI-IS, Peter with BCCN and MPI-IS.

Perceiving Systems – http://ps.is.tuebingen.mpg.de