# Unite the People: Closing the loop between 3D and 2D Human Representations Supplementary Material

Christoph Lassner[1,2] classner@tue.mpg.de
Javier Romero[2] jromero@tue.mpg.de
Martin Kiefel[2] mkiefel@tue.mpg.de
Federica Bogo[2,3] fbogo@tue.mpg.de
Michael J. Black[2] black@tue.mpg.de
Peter V. Gehler[1,2] pgehler@tue.mpg.de

Bernstein Center for Comp. Neuroscience[1] Otfried-Müller-Str. 25, Tübingen
Max-Planck Institute for Intelligent Systems[2] Spemannstr. 41, Tübingen
Microsoft[3] 21 Station Rd., Cambridge

## 1. Introduction

We have obtained human segmentation labels to integrate shape information into the SMPLify 3D fitting procedure and for the evaluation of methods introduced in the main paper. The labels consist of foreground segmentation for multiple human pose datasets and six body part segmentation for the LSP dataset. Whereas we discuss their use in the context of the UP dataset in the main paper, we discuss the annotation tool that we used for the collection (see Sec. 2.1) as well as the direct use of the human labels for model training (see Sec. 2.2) in this document.

In Sec. 3.1, we show additional evaluation data of our fine-grained models and conclude with further examples for the applications showcased in the paper in 3.2.

## 2. Human Segmentation Labels

### 2.1. Openpose

To get segmentation labels on large scale, we built an interactive annotation tool on top of the Opensurfaces package [2]: Openpose. It works with Amazon Mechanical Turk and uses the management capabilities of Opensurfaces.

However, it is tedious to collect fine-grained segmentation annotations: it cannot be done with single clicks, and making an annotation border consistent with image edges can be frustrating if done without guidance. To tackle the aforementioned problems, we use the interactive Grabcut algorithm [7] to make the segmentation as easy and fast as possible.

The worker task was to scribble into (part) foreground and background regions until the part of interest was accurately marked. An experienced user can segment images in less than 30 seconds. We received many positive comments for our interface.

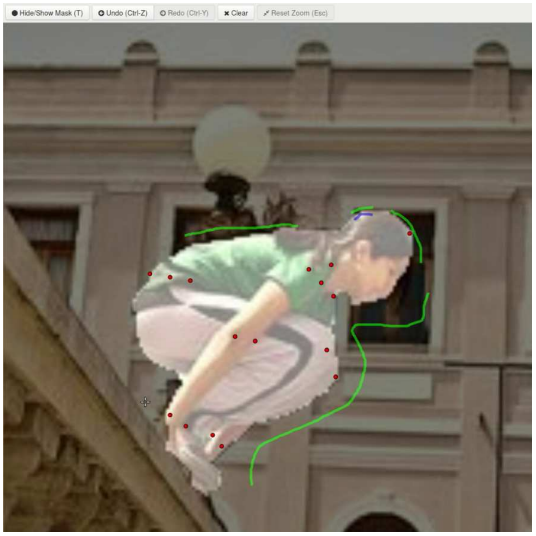



Figure 1: The labeling interface of our Openpose toolbox. Green scribbles mark background, blue scribbles foreground. The red dots indicate annotated pose keypoints. Keypoints are used to initialize the Grabcut [7] mask.

### 2.2. Models and Results

To explore the versatility of the human labeled data, we combine all 25,030 images from our annotated datasets with foreground labels to form a single training corpus. For this series of experiments, we use a Deconvnet-model [6].

We found that a person size of roughly 160 pixels works best for training, therefore we normalize and cut out the people accordingly (for the LSP core dataset this is not necessary since they are roughly in the expected size range). The images are mirrored and rotated up to 30 degrees in both directions to augment the training data as much as possible.

To obtain the final scores, we finetune the model to the datasets they will be tested on. A summary of scores before

| | LSP | MPI-HPDB | Fashionista | Human3.6M |
|---|---|---|---|---|
| Mean acc. | 0.9625 | 0.9584 | 0.9738 | 0.9884 |
| Mean f1 | 0.9217 | 0.9092 | 0.9407 | 0.8166 |
| Mean acc. (ft.) | 0.9684 | 0.9628 | | |
| Mean f1 (ft.) | 0.9336 | 0.9169 | | |

Table 1: Person vs. background segmentation results for our base model. The Fashionista and Human3.6M datasets are used to demonstrate the generalization capability without finetuning.

Figure 2: Example segmentations from the fashionista dataset. Large handbags are the biggest source of uncertainty without adaptation to the dataset domain.

Figure 3: HumanEva example (left) and Human3.6M examples (middle, right) and ground truth.

and after finetuning can be found in Tab. 1.

Additionally to our two annotated test datasets from LSP and MPII-HumanPose single person, we test the models on the three datasets Fashionista [10], HumanEva [8] and Human3.6M [3] without finetuning to analyze their generalization behavior and the difficulty and variety of the training dataset (for scores where applicable, see Tab. 1).

**Results on the Fashionista Dataset** We use our model to predict the images of the *Fashionista* test dataset without using the training part. Example results can be found in Fig. 2. We achieve a competitive accuracy of 0.9738. This compares to 0.9608 (over all 56 classes of the dataset) reported by the current state-of-the art method [5]. Liang et al. achieve an improved score of 0.9706 when including the *Chictopia10k* dataset for training. By using the ratio of $\approx 77\%$ background in the dataset (c.t. [9]), it is possible to give an interval for the foreground vs. background accuracy of their segmentation method: $[0.9608; 0.9960]$, with which we can compete even though we have not used any fashion specific data. We would expect a large improvement from finetuning, because the main failure case of our foreground segmentation are large handbags (see, e.g., Fig. 2, image three). They provide enough visual cues to reliably adapt.

**Results on Human3.6M and HumanEva** We test our model on the two 3D evaluation sets used in the main paper without finetuning. Examples are shown in Fig. 3.

For Human3.6M, there is segmentation information available that was obtained from background subtraction. In the rightmost image in Fig. 3 we show our segmentation result together with the ground truth, where chair and parts of the background are labeled erroneously as foreground. To calculate accuracy and f1 scores on the Human3.6M dataset, we sample 5 images from all of the commonly used test sequences of subjects 9 and 11 randomly, which is a total of 1,190 images, and average their scores.

**Body Part Segmentation** We leverage the base human segmentation model by dropping its last layer and retraining it on our body part segmentation data. This is required due to the very little training data we have for this task (only 1,000 examples). On average, the retrained network achieves a score of 0.9095 accuracy and 0.6046 macro f1. Example results can be found in Fig. 5.

Furthermore, this allows us to make comparisons between a model trained on the human labeled data, a model trained on generated data from SMPL on the exact same set (including potentially erroneous fits) and our model trained on UP-S31 (which contains only the subset of 'good' fits on the LSP training set but additional good fits to the other datasets) reduced to a six part representation. We provide example segmentations for comparison in Fig. 6.

The resulting macro f1 scores are 0.6046 for the human model, 0.5628 for the model trained on LSP SMPL projections, and **0.6101** for the 31 part segmentation model reduced to six parts. The model trained on the generated annotation outperforms the model trained on human labels, highlighting again the versatility and quality of the dataset presented in the main paper.

## 3. Additional Evaluation of the Fine-Grained Models

### 3.1. Part-by-part Evaluation

In Fig. 7, we provide visualizations of the scores for fine-grained segmentation and keypoint localization. All the values are from models trained and tested on the full UP dataset.

Unsurprisingly, the segmentation scores for wrists, hands, ankles and feet are the lowest. This is not only due to the model being unstable in this regions (c.t. Fig. 6, Fig. 8), but also due to our generated ground truth being noisy in these regions, since SMPL does not receive information about foot or hand orientation during the fits other than the foreground segmentation. The mean IOU score over all parts is 0.4560, and the accuracy 0.9132.

The keypoint score visualization shows the same pattern, with the lowest scores at the big toes. The overall stability is very high with an average PCK@0.2 of 0.9450 (for the 14 core keypoints: 0.9388).

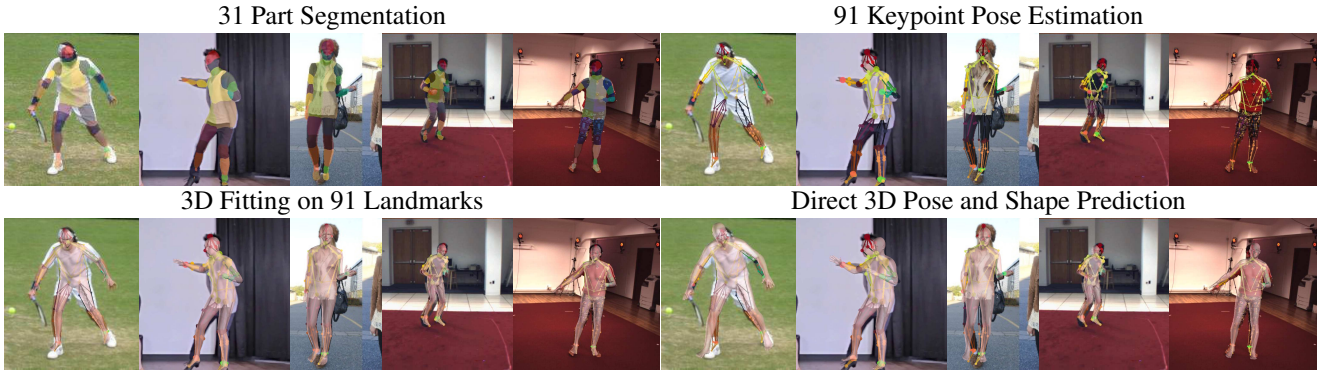



Figure 4: Results from various methods trained on labels generated from the UP dataset (improved Fig. 4, main paper).

Figure 5: Part segmentations on LSP. Due to little training data, the model remains rather unstable.

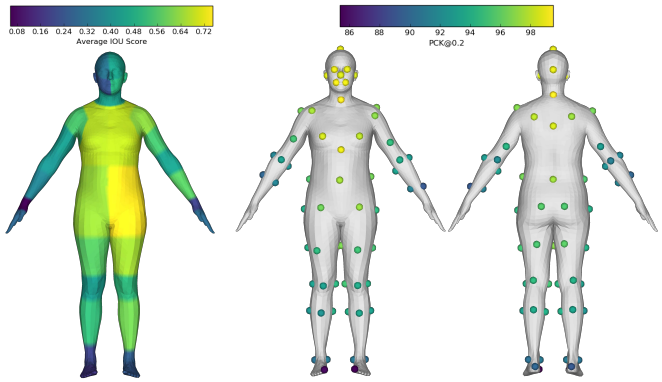
Figure 6: Comparison of part segmentation on two examples from the LSP test set. For each example from left to right: CNN trained on the 1,000 human labeled training examples, CNN trained on projections from SMPL fits on the same 1,000 samples, CNN trained on UP-S31, which contains only the high quality fits from the 1,000 samples as well as high quality fits from the other datasets.

### 3.2. Examples

Due to an image size-independent visualization parameter, the 91 landmark pose visualizations are rather small in the main paper. An improved version of Fig. 4 (main) is provided in Fig. 4 and has been integrated into the main paper.

Additionally to that, we provide more examples from many datasets for all our discussed learning and fitting methods in Fig. 8. We added examples from the Fashionista dataset [10], from which we did not use data for finetuning the models. Further samples from the LSP dataset [4] and the MPII Human Pose Database [1] are provided, which have harder background than the motion capture datasets.

Additional examples of improved fits from using 91 keypoints from our predictor over fits to the 14 ground truth




(a) Segmentation. (b) Landmark detection.

Figure 7: Visualization of the discriminative model scores.

keypoints can be found in Fig. 9. We only show one fit that improves due to wrong ground truth labels (first row, rightmost triple), even though this is a frequent source of improvement. Instead, we want to highlight that the fits take more details into account and the additional keypoints provide hints to disambiguate perspective and rotation. With the cues about limb rotation, the fits look more realistic. We expect this to improve the pose estimator even further once the additional samples are integrated into the dataset.

## References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3, 4

[2] 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. 1

[3] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, July 2014. 2

(a) 31 Part Segmentation.

(b) 91 Keypoint Pose Estimation.

(c) 3D Fitting on 91 Landmarks.

(d) Direct 3D Pose and Shape Prediction.

Figure 8: Additional results on the Fashionista dataset [10] (first three images), LSP dataset [4] (images three to six) and MPII Human Pose Database [1] (last two images).

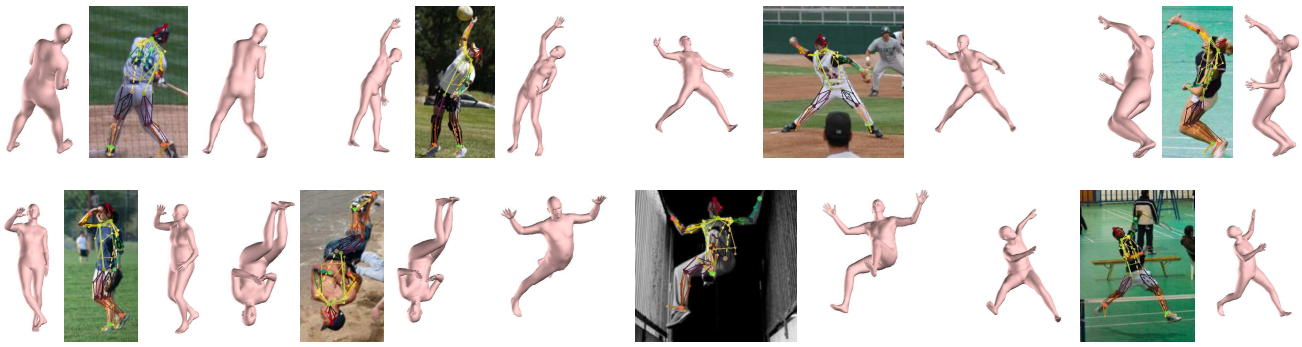


Figure 9: For each image triple: improvement over fits to 14 ground truth keypoints (left) by using 91 keypoints from our predictor (center, right) on the LSP dataset.

[4] S. Johnson and M. Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *British Machine Vision Conference (BMVC)*, 2010. doi:10.5244/C.24.12. 3, 4

[5] X. Liang, C. Xu, X. Shen, J. Yang, J. Tang, L. Lin, and S. Yan. Human parsing with contextualized convolutional neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99), 2016. 2

[6] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[7] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, August 2004. 1

[8] 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

[9] K. Yamaguchi, H. Kiapour, L. E. Ortiz, and T. L. Berg. Retrieving similar styles to parse clothing. *IEEE Transactions on Parrern Analysis and Machine Intelligence (TPAMI)*, 37(5):1028–1040, May 2015. 2

[10] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3570–3577, June 2012. 2, 3, 4