Learning to Regress Bodies from Images using Differentiable Semantic Rendering

Supplementary Material

Sai Kumar Dwivedi[1] Nikos Athanasiou[1] Muhammed Kocabas[1,2] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]ETH Zurich
{sdwivedi, nathanasiou, mkocabas, black}@tue.mpg.de

1. Clothing Semantic Information

It is difficult to obtain ground-truth clothing segmentation masks for in-the-wild datasets. Hence, we use Graphonomy [1], which is an off-the-shelf human clothing segmentation model that provides reasonably reliable pseudo-ground truth.

1.1. Clothing Segmentation Masks

Graphonomy has three different models depending on the granularity of the segmentation mask and we choose the one with 20 labels, also known as the *Universal Model*. This model provides the best clothing segmentation performance compared to other Graphonomy variants. The different labels are: *Background, Hat, Hair, Glove, Sunglasses, UpperClothes, Dress, Coat, Socks, Pants, Jumpsuits, Scarf, Skirt, Face, LeftArm, RightArm, LeftLeg, RightLeg, LeftShoe* and *RightShoe*.

During inference, to get more accurate predictions – as suggested in the original implementation – we use 4 different scaling factors for the input image – $0.5, 0.75, 1.0, 1.5$ – to account for different image resolutions. Then, we merge the outputs for different scaling factors using appropriate upsample and downsample functions (bilinear) to produce an output size the same as the original image. For images more than 1080×1080, we use a single scaling factor of 1.0. We also flip the image horizontally and average the output predictions of the flipped image with the original one.

1.2. Processing Pseudo Ground-Truth Masks

The generated pseudo ground-truth cannot be directly used for supervising existing human body estimator networks because of incompatibility between Graphonomy's output and 3D pose regressor's training procedure [4].

Graphonomy is not an instance segmentation model, which means it is hard to differentiate between people in the image. However, standard human body estimators [3–5] use a single person during training. To circumvent this problem, we use 2D keypoints to get a rough estimate of the region of the person in the image. Furthermore, we add/subtract an offset of 30 pixels in both x and y direction according to the maximum/minimum keypoint location.

Due to occlusion or inaccuracies in the prediction, the spread of pixels for a particular label of Graphonomy may cover an extremely small part of the image. As DSR-MC tries to tightly supervise the rendered SMPL body with the target binary mask, it is important to ensure the target masks are reliable. Hence, we remove labels that cover less than 60 pixels from the predefined set of five labels *(LeftArm, RightArm, LeftShoe, RightShoe, Face)*.

There is a one to one mapping from the DSR-MC labels to Graphonomy labels. The same is not true for DSR-C as there are several clothing labels. Consequently, for DSR-C, we define a coarse mapping as per Table 1.

DSR-C Labels	Graphonomy Labels
Background	Background
LowerClothes	Pants, Skirts
UpperClothes	Upperclothes, Dress, Coat, Jumpsuits
MinimalClothing	Hat, Hair, Glove, Sunglasses, Socks, Scraf, Face, LeftArm, RightArm, LeftLeg, RightLeg, LeftShoe, RightShoe

Table 1: Mapping of DSR-C labels to Graphonomy labels.

2. Semantic Prior for SMPL

To supervise the human body regressor network with semantic information, we need a term that captures the *a priori* probability that describes what parts of the SMPL body correspond to a particular semantic label. To this end, we use 2500 clothed human scans from the AGORA dataset [8] with varied clothing, pose and identity. AGORA contains clothed 3D people with ground truth SMPL-X bodies fit to

Figure 1: **Clothed Human Scans.** Examples of clothed human scans in different clothing, pose and camera views (*Columns* 1,3,5) along with the corresponding SMPL bodies where each vertex is colored based on the output of the clothing segmentation model [1] (*Columns* 2,4,6) applied on the respective scan images. We only show 3 camera views here.

the scans. We convert SMPL-X fits to SMPL. For each scan, we render it from 10 different camera views to cover different angles and generate scan images. We run Graphonomy on each of these images to obtain 10 2D clothing segmentation images for each scan. An illustration of the output from this process is depicted in Fig. 1. We also render the fitted SMPL model with the known camera parameters to obtain the correspondences between the vertices of the SMPL body and the pixels in the image.

Given this training data, we can very simply compute the prior probability of a SMPL vertex having one of the 20 Graphonomy labels. We estimate this by calculating the occurrences of a particular label being present at the vertex divided by the total occurrences of other labels–excluding the *Background* label. Finally, this gives us the prior per-vertex probability that a SMPL vertex has given a Graphonomy label. We also assign a small probability of a vertex being assigned the background label; this increases robustness to occlusion. As an additional step, we use the SMPL body part segmentation to clean the semantic prior. Graphonomy gives incorrect predictions for some clothed body scan images and this will affect downstream tasks. Hence, if the semantic label probability of a "leg" vertex (denoted by SMPL part segmentation) has a higher probability of being hand, we set it to zero. This approach helps to avoid obvious failures when Graphonomy produces incorrect predictions. Note that a more sophisticated prior model could also capture spatial correlations of clothing but we did not find this necessary.

3. Failure Case Analysis

We qualitatively analyse the failure cases using our method and broadly categorise them into two types: occlusion failures as shown in Fig. 2 and multi-person failures as shown in Fig. 3. Note that these are also cases where standard 3D pose estimation methods commonly fail.

First, we observe failures in case of either self-occlusion or scene occlusion producing unreasonable pose. Hence, we tried to analyse the training samples with occlusion. As we can see in Fig. 2, Graphonomy outputs a black patch (Background class) when an object or the scene is occluding the person. As DSR-C tries to minimise the negative

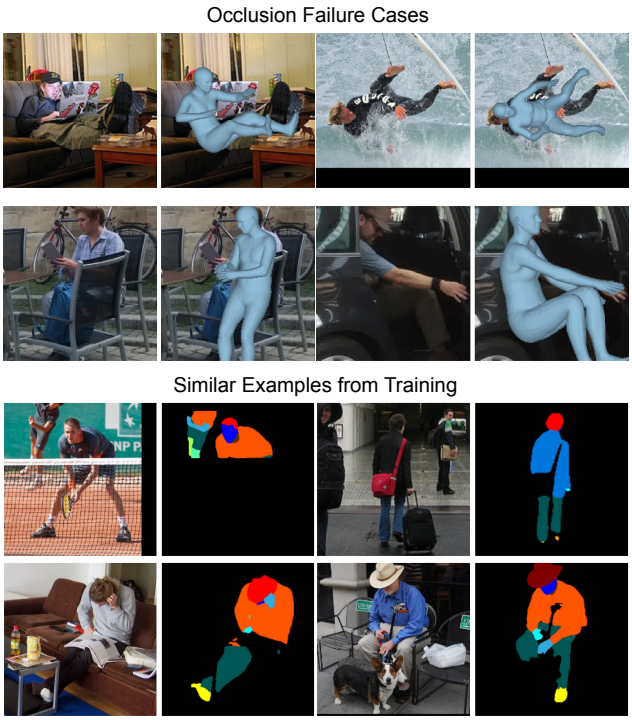

Figure 2: **Occlusion Failure Analysis** Qualitative failure results in case of occlusion. We show outputs from COCO and 3DPW in *Rows 1-2* respectively. *Rows 3-4:* Similar occlusion cases present in the training samples.

Figure 3: **Multi-Person Failure Analysis** Qualitative failure results in case multiple people are present. We show outputs from COCO and 3DPW in *Rows 1-2* respectively. *Rows 3-4:* Similar multi-person failure cases present in the training samples.

log probability of a rendered vertex being a particular label, and the background label has a low probability, occlusions can cause the pose to be incorrect. More complete labeling of things like backpacks or training with synthetic occlusion could improve this. Moreover, it can also hinder detailed fitting of the body where the labels associated with DSR-MC are occluded. Additional occlusion handling techniques could help our approach in such cases.

Furthermore, another failure case occurs when multiple people are present in a scene. As Graphonomy is not an instance segmentation network, the pseudo ground-truth data may still contain other people even after using the heuristics to clean them, as described in Section 1.2. This confuses training, resulting in misaligned bodies at inference time. Figure 3 shows common cases where all the upper body clothing of multiple people are merged into one segment and clothing masks of partially visible people in the background, which affect the quality of the obtained masks. Our entire method could be improved by better instance-level clothing segmentation.

Higher quality of Graphonomy masks leads to increased performance gains in the case of DSR. We demonstrate it by doing an ablation study using Human3.6M [2] dataset where the Graphonomy predictions are more reliable because of the simpler background and single subject. The quantitative results of this experiment are reported in the main paper.

Overall, our performance is affected by the off-the-self model we use to supervise the clothing semantics of the person. However, improvements over the state-of-the-art show that even weak supervision of clothing semantics is crucial for detailed 3D body fits. The success of our approach suggests that more accurate human parsing and clothing segmentation are a good investment for the community.

4. Additional Qualitative Results

We show additional qualitative results comparing our method with other state-of-the-art methods [3, 5] for 3DPW [6] and COCO [7] which are challenging in-the-wild benchmarks for 3D human pose and shape estimation. The results are depicted in Figures 4 and 5. Next to each example, we show the corresponding side view. We observe that our approach produces more accurate pose and shape that are better aligned with the human in the image than current SOTA approaches.

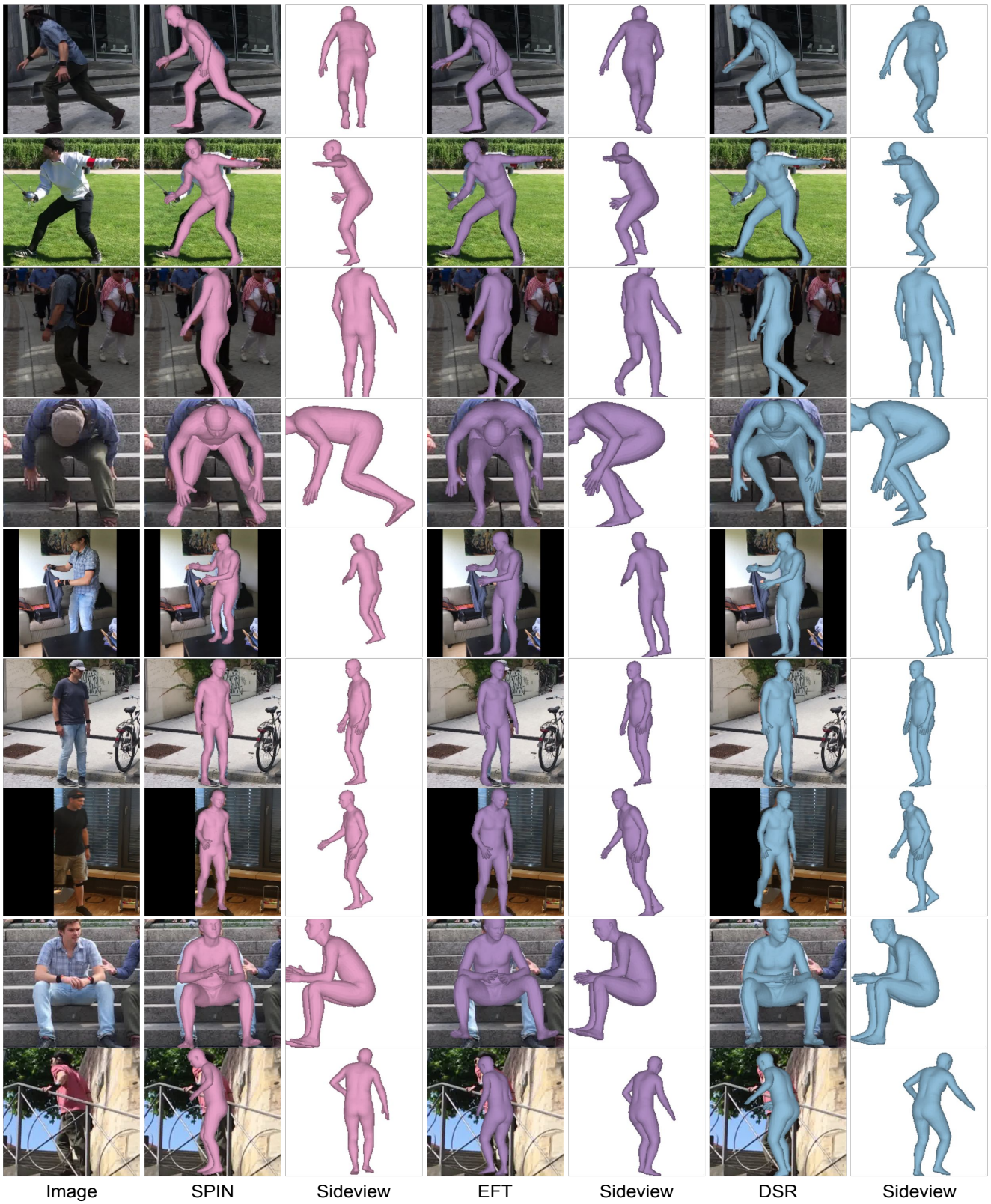

Figure 4: **Additional Qualitative Results of 3DPW.** From left to right - Input image, SPIN [5], SPIN Sideview, EFT [3], EFT Sideview, DSR and DSR Sideview results

Figure 5: **Additional Qualitative Results of COCO.** From left to right - Input image, SPIN [5], SPIN Sideview, EFT [3], EFT Sideview, DSR and DSR Sideview results

References

[1] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[2] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI)*, 2014. 3

[3] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *arXiv preprint arXiv:2004.03686*, 2020. 1, 3, 4, 5

[4] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[5] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 4, 5

[6] Sijin Li and Antoni B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*, 2014. 3

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 3

[8] Priyanka Patel, Chun-Hao Paul Huang, Joachim Tesch, David Hoffmann, Shashank Tripathi, and Michael J Black. AGORA: Avatars in geography optimized for regression analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1