SUPPLEMENTARY MATERIAL: Learned Vertex Descent: A New Direction for 3D Human Model Fitting

Enric Corona[1], Gerard Pons-Moll[2,3], Guillem Alenyà[1], and Francesc Moreno-Noguer[1]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]University of Tübingen, Germany, [3]Max Planck Institute for Informatics, Germany

In this supplementary material, we provide a detailed description of the implementation details and the data augmentation we used. We also include more qualitative examples and a supplementary video which summarizes the method and the contributions of the paper.

1 Implementation details

We next describe the main implementation details. The code will be made publicly available.

The clipping factor for the learnt gradient is to 18% of the vertical size of the scan, which we normalize between -0.75 and 0.75. In our experiments, $H = W = 256$, f is a stacked hourglass network [10] trained from scratch with 4 stacks and batch normalization replaced with group normalization [17]. The feature embeddings have size 128×128 with 256 channels each. Therefore, query points have a feature size of $F = 256 \times 4 = 1024$. The MLP f is formed by 3 fully connected layers with Weight Normalization [15], and deeper architectures or positional encoding did not help to improve performance. We attribute this to the fact that the MLP is already obtaining very rich representations from feature maps. For images, we assume a weak-perspective projection although our approach is compatible with perspective cameras.

The networks are trained end-to-end with batch size 4, learning rate 0.001 during 500 epochs, and then with linear learning rate decay during 500 epochs more. We use Adam Optimizer [7] with $\beta_1 = 0.9$ $\beta_2 = 0.999$. When considering point-clouds as input we train an IF-Net backbone[5] from scratch with the same training conditions and number of iterations.

Implementation-wise f has an output dimension of $N = 6890$. When estimating an SMPL shape, we input a surface of 6890×3 and obtain a prediction tensor with shape $6890 \times 6890 \times 3$, from which we sample the diagonal to obtain per-vertex displacements (6890×3) and move each vertex in the correct direction. For the task of registration of the MANO model [13], we instead predict 778 vertices.

To compare LVD against other baselines, we used their available code. For SMPL-X, we fitted the SMPL model for better comparison with ours and previous works, using their most recent code (SMPLify-X) with the variational prior.

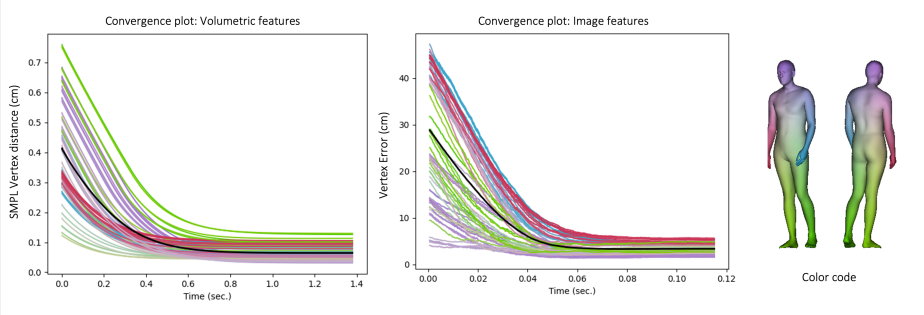

Fig. 1. Convergence plot of the proposed optimization, for voxel-based experiments in comparison to image-based reconstruction. In comparison with the reported results on image-based reconstruction (which also are shown in the main paper), volumetric reconstruction takes almost a second to converge with our settings. Experiments were run on a single GeForce® GTX 1080 Ti GPU. The black line represents the average of all vertex errors while the remaining colors show how the error is distributed among different body parts, *e.g.* . arms and feet accumulate the biggest error while torso or head generally are the most accurately reconstructed parts.

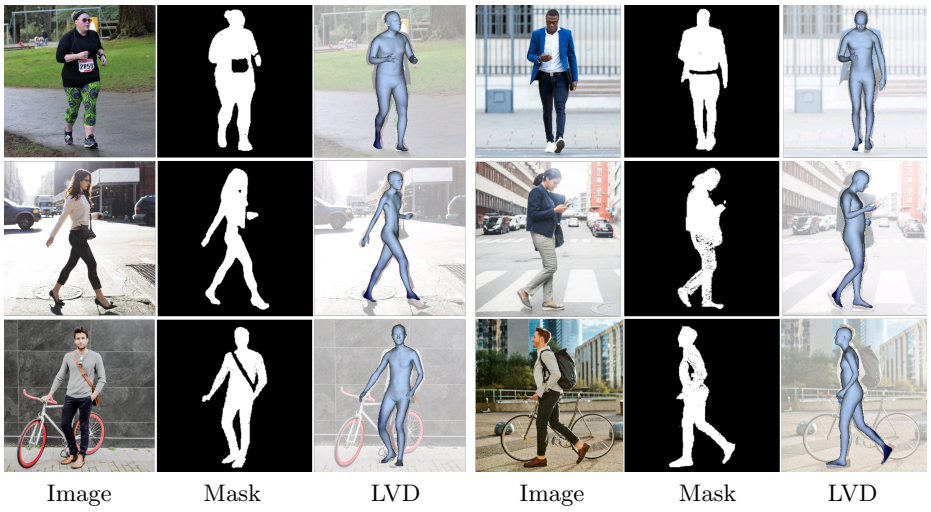

Fig. 2. SMPL reconstruction on images in-the-wild, and the predicted foreground masks[18]. Even with noisy segmentations, the predicted SMPL accurately represents the body shapes and poses of the target people.

Fig. 3. More examples of body shapes estimated on images in-the-wild.

2 Data Augmentation for image data

As mentioned in the main document, we use the RenderPeople, AXYZ and Twindom datasets [12, 1, 16], which consist of 767 3D scans. We first obtain SMPL registrations and manually annotate the correct fits, leaving 750 scans. Due to the reduced number of 3D scans, we augment each of them by changing its pose and body shape. On one side, we label pose vectors for humans walking and running, and automatically select a random pose + noise for each new augmentation. To pose the 3D scan, we simply assign the skinning weights of each 3D surface vertex to those of the closest SMPL vertex. This can lead to several artifacts, for body parts that are in contact, such as hands, which will generate very large triangles. We manually prune the generated 3D scans to remove these cases.

Next, we tune the body shape of each 3D scan by changing the first shape parameter in the PCA space. We discretize a number of augmentations with respect to the initial shape and calculate the linear displacement for each body vertex. For the 3D scan we apply the displacement of the closest vertex. This augmentation is proven to be really useful and does not significantly create artifacts since it retains self-contact information. We perform 6 augmentations for each scan.

For the task of human reconstruction from images, we then render each augmentation by rotating around the yaw axis to gather views with different illuminations. As mentioned in the main document, in total we obtain $\sim 680k$ rendered images that are used for training and validation.

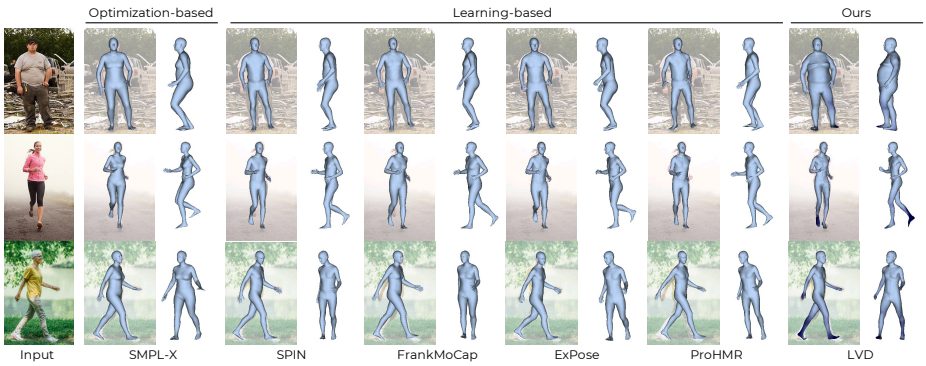

Fig. 4. Qualitative comparisons with more methods. For each method, we show front and side views of the reconstruction.

Note that the original data consisted only of a few hundred 3D scans, all with very average body shapes. The augmentation led the model represent more diverse shapes and avoid overfitting, but the proposed Learned Vertex Descent paradigm was necessary for it to represent them well. The baseline that predicts SMPL parameters directly did not manage to generalize well beyond the training set.

3 Experiments

As mentioned in the main document, we train our model without backgrounds when taking images as input. Therefore at test time we use RP-R-CNN [18] to automatically segment the foreground person before running the forward pass. However, this can still generate masks with artifacts or missing parts. We show in Fig. 2 that the proposed approach is robust to these noisy masks or parts that were incorrectly segmented.

We also show more qualitative examples of 3D reconstruction from a single view in-the-wild in Fig. 3, and Fig. 4 shows comparisons with the rest of the methods that are not shown in the main document. In particular, we noted several differences between optimization-based and learning-based body pose/shape estimation methods. On one hand, optimization-based methods [4, 11] are often accurate, but have severe failure cases and are slow. On the other hand, learning based methods [8, 14, 6, 9] regress global parameters from the full image. Hence, the shape estimates have a strong bias towards the mean. Moreover, learning-based methods are not able to verify their initial estimates against the image.

Our goal in this paper is to combine the advantages of both methods. LVD produces varied shape estimates thanks to the learned per vertex descent directions which are conditioned on local image evidence, and can work in real time.

Fig. 5. SMPL registration of 3D scans showing SMPL and SMPL-D for LoopReg, IP-Net and LVD.

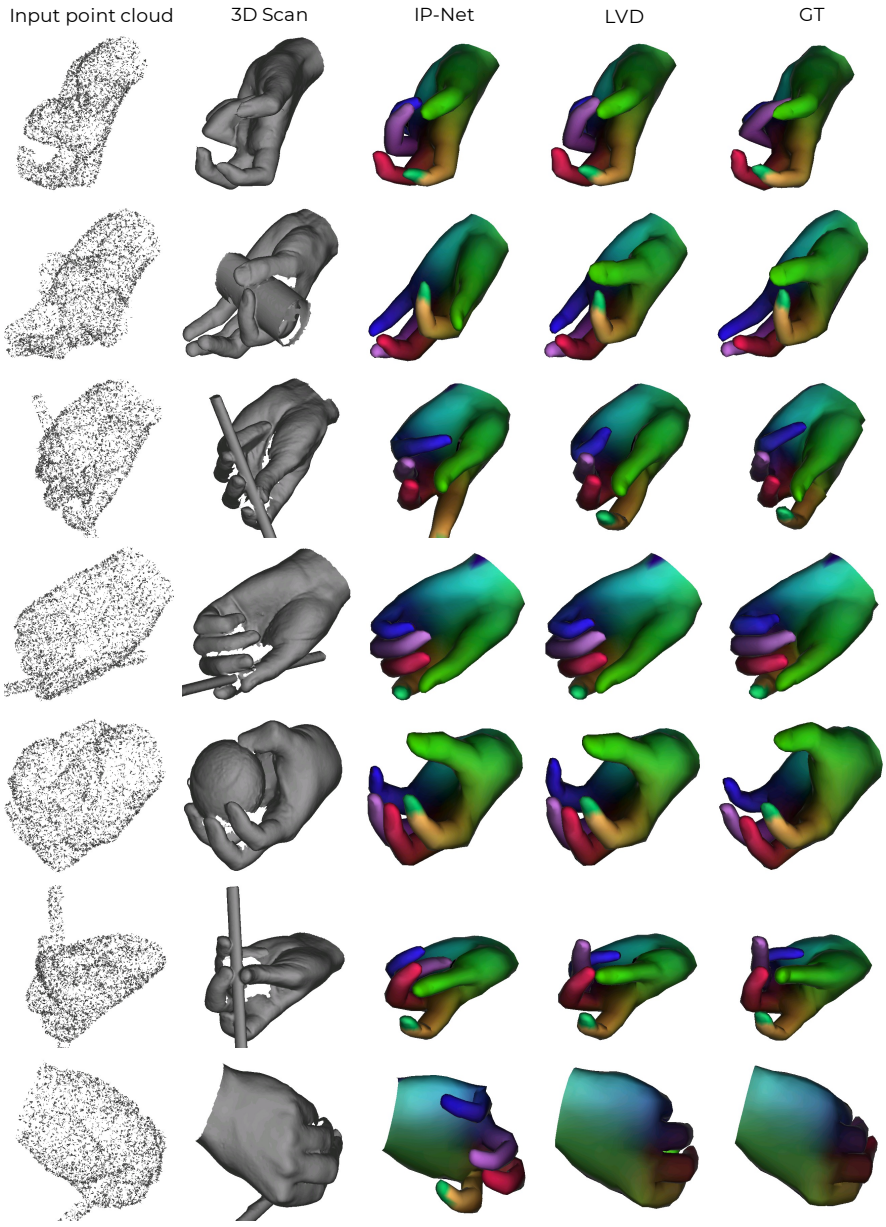

Fig. 6. Registration of 3D Hands using MANO [13]. The input to IP-Net [2] or LVD is the input point cloud in the left column, while the groundtruth 3D scan is shown in the second column. IP-Net performs similarly well in most cases, but is most confused in the presence of other objects or very noisy pointclouds.

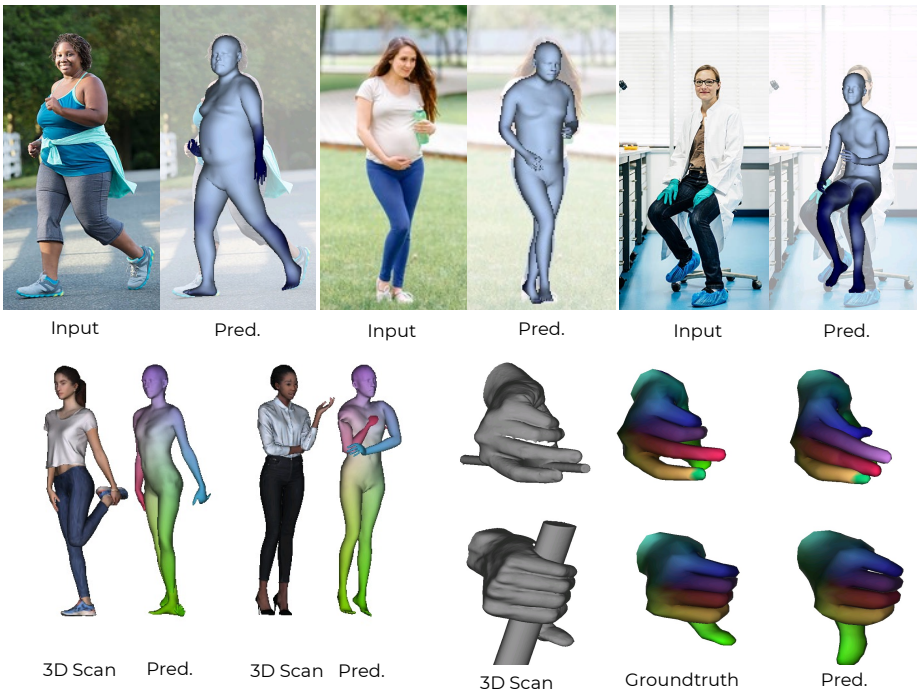

Fig. 7. Failure cases from LVD in body shape estimation from single view images (first row), 3D registration of humans from point clouds (second row - left) and 3D registration from hands (second row - right). See Section 4 for more details.

In addition, we focus on designing a general method that is straightforward to apply to other input modalities such as 3D point clouds. In this direction, Fig. 5 includes more results on the task of 3D registration of 3D scans and Fig. 6 shows 3D registration results of MANO of LVD in comparison to those of IP-Net [2]. IP-Net obtains quantitative results close to LVD, and works generally well for clean 3D scans. However, it might converge to wrong local minima when tackling 3D point clouds with objects or holes.

4 Failure cases.

We finally include failure cases of LVD in all tasks where we evaluate our approach, in Fig. 7. For the task of body shape estimation from single view (First row), the body shapes we can generate are limited by the SMPL model and the training data, and cannot accurately reproduce body shapes of *e.g.* pregnant women (second example). Furthermore, our training data is rather limited in the diversity of body poses, so challenging body poses is another reason for failure cases. For instance, examples in Fig. 7 top-left and top-right show scenarios

that are rare in the train data, and the predicted body does not correctly adjust to the input image. However, note that the wrong body parts are predicted to have a big uncertainty (in dark blue).

In Fig. 7 (Second row) we show more examples of failure cases in 3D registration of human scans and hands.

References

1. Axyz dataset. https://secure.axyz-design.com/
2. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: ECCV. pp. 311–329. Springer (2020)
3. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. NeurIPS **33** (2020)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: ECCV. Springer (2016)
5. Chibane, J., Pons-Moll, G.: Implicit feature networks for texture completion from partial 3d data. In: ECCV. pp. 717–725. Springer (2020)
6. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: ECCV. pp. 20–40. Springer (2020)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
8. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
9. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV. pp. 11605–11614 (2021)
10. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. pp. 483–499. Springer (2016)
11. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: CVPR (2019)
12. Renderpeople dataset. https://renderpeople.com/
13. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ToG (2017)
14. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. arXiv preprint arXiv:2008.08324 (2020)
15. Salimans, T., Kingma, D.P.: Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In: Advances in neural information processing systems. pp. 901–909 (2016)
16. Twindom dataset. https://web.twindom.com/
17. Wu, Y., He, K.: Group normalization. In: ECCV. pp. 3–19 (2018)
18. Yang, L., Song, Q., Wang, Z., Hu, M., Liu, C., Xin, X., Jia, W., Xu, S.: Renovating parsing r-cnn for accurate multiple human parsing. In: ECCV (2020)