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

Abstract. We propose a novel optimization-based paradigm for 3D human model fitting on images and scans. In contrast to existing approaches that directly regress the parameters of a low-dimensional statistical body model (*e.g.* SMPL) from input images, we train an ensemble of per vertex neural fields network. The network predicts, in a distributed manner, the vertex descent direction towards the ground truth, based on neural features extracted at the current vertex projection. At inference, we employ this network, dubbed LVD, within a gradient-descent optimization pipeline until its convergence, which typically occurs in a fraction of a second even when initializing all vertices into a single point. An exhaustive evaluation demonstrates that our approach is able to capture the underlying body of clothed people with very different body shapes, achieving a significant improvement compared to state-of-the-art. LVD is also applicable to 3D model fitting of humans and hands, for which we show a significant improvement to the SOTA with a much simpler and faster method. Code is released at `https://www.iri.upc.edu/people/ecorona/lvd/`

1 Introduction

Fitting 3D human models to data (single images / video / scans) is a highly ambiguous problem. The standard approach to overcome this is by introducing statistical shape priors [5,45,85] controlled by a reduced number of parameters. Shape recovery then entails at estimating these parameters from data. There exist two main paradigms for doing so.

On the one side, optimization-based methods iteratively search for the model parameters that best match available image cues, like 2D keypoints [58,11,6,38], silhouettes [42,77] or dense correspondences [29]. On the other side, data-driven regression methods for mesh recovery leverage deep neural networks to directly predict the model parameters from the input [35,29,59,19,26,4,54]. In between these two streams, there are recent approaches that build hybrid methods combining optimization-regression schemes [87,34,38,79].

Regardless of the inference method, optimization or regression, and input modality, 2D evidence based on the entire image, keypoints, silhouettes, point-clouds, all these previous methods aim at estimating the parameters of a low-dimensional model (typically based on SMPL [45]). However, as we will show in

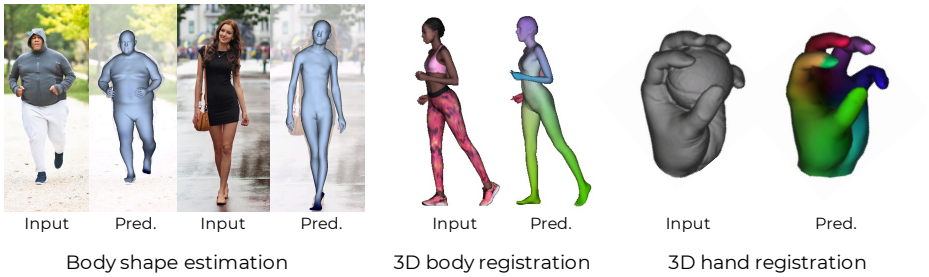

Fig. 1. Learned Vertex Descent (LVD) is a novel optimization strategy in which a network leverages local image or volumetric features to iteratively predict per-vertex directions towards an optimal body/hand surface. The proposed approach is directly applicable to different tasks with minimal changes on the network, and we show it can fit a much larger variability of body shapes than previous state-of-the-art. The figure depicts results on the three tasks where we have evaluated LVD: body shape reconstruction from a single image, and 3D fitting of body and hand scans.

the experimental section, these models struggle in capturing detailed body shape, specially for morphotypes departing from the mean (overweight or skinny people) or when the person is wearing loose clothing. We hypothesize that this is produced by two main reasons: 1) the models induce a bias towards the mean shape; and 2) the mapping from local image / pointcloud features to *global shape* parameters is highly non-linear. This makes optimization-based approaches prone to get stuck at local minima and have slow run times. Global shape regression methods lack the error-feedback loop of optimization methods (comparing the current estimate against image / scan input), and hence exhibit an even more pronounced bias towards mean shapes. Overcoming this problem would require immense amounts of training data, which is infeasible for 3D bodies.

To recover more detail, recent works regress or optimize a set of displacements on top of SMPL global shape [3,1,4,10,57], local surface elements [46] or points [48]. Like us, [39] by-pass the regression of global shape parameters and regress model vertices direclty. However, similar to displacement-based methods [1,4], the proposed regression scheme [39] predicts the position of all points in one single pass and lacks an error-feedback loop. Hence, these methods regress a global shape in one pass based on global image features and also suffer from bias towards the mean. Works based on implicit surfaces [17,71,72] address these limitations by making point-wise distributed predictions. Being more local, they require less training data. However, these methods do not produce surfaces with a coherent parameterization (e.g. SMPL vertices), and hence control is only possible with subsequent model fitting, which is hard if correspondences are not known [8,9,33,31].

In this paper, we propose a significantly different approach to all prior model fitting methods. Inspired by classical model-based fitting, where image gradients drive the direction of vertices and in turn global shape parameters, we propose to iteratively learn where 3D vertices should move based on neural features. For that purpose, we devise a novel data-driven optimization in which an ensemble of per-vertex neural fields is trained to predict the optimal 3D vertex displacement

towards the ground-truth, based on local neural features extracted at the current vertex location. We dub this network LVD, from 'Learned Vertex Descent'. At inference, given an input image or scan, we initialize all mesh vertices into a single point and iteratively query LVD to estimate the vertex displacement in a gradient descent manner.

We conduct a thorough evaluation of the proposed learning-based optimization approach. Our experiments reveal that LVD combines the advantages of classical optimization and learning-based methods. LVD captures off-mean shapes significantly more accurately than all prior work, unlike optimization approaches it does not suffer from local minima, and converges in just 6 iterations. We attribute the better performance to the *distributed per-vertex predictions* and to the *error feedback loop* – the current vertex estimate is iteratively verified against the image evidence, a feature present in all optimization schemes but missing in learning-based methods for human shape estimation.

We demonstrate the usefulness of LVD for the tasks of 3D human shape estimation from images, and 3D scan registration(see Fig 1). In both problems, we surpass existing approaches by a considerable margin.

Our key contributions can be summarized as follows:

- A novel learning-based optimization where vertices *descent* towards the correct solution according to learned neural field predictions. This optimization is fast, does not require gradients and hand-crafted objective functions, and is not sensitive to initialization.
- We empirically show that our approach achieves state-of-the-art results in the task of human shape recovery from a single image.
- The LVD formulation can be readily adapted to the problem of 3D scan fitting. We also demonstrate state-of-the-art results on fitting 3D scans of full bodies and hands.
- By analysing the variance of the learned vertex gradient in local neighborhoods we can extract uncertainty information about the reconstructed shape. This might be useful for subsequent downstream applications that require confidence measures on the estimated body shape.

2 Related work

2.1 Parametric models for 3D body reconstruction

The *de-facto* approach for reconstructing human shape and pose is by estimating the parameters of a low-rank generative model [45,58,85,69], being SMPL [45] or SMPL-X [58] the most well known. We next describe the approaches to perform model fitting from images.

Optimization. Early approaches on human pose and shape estimation from images used optimization-based approaches to estimate the model parameters from 2D image evidence. Sigal *et al* [77] did so for the SCAPE [5] human model, and assuming 2D input silhouettes. Guan *et al* [28], combined silhouettes with manually annotated 2D skeletons. More recently, the standard optimization relies on

2D skeletons [11,58,79], estimated by off-the-shelf and robust deep methods [14]. This is typically accompanied by additional pose priors to ensure anthropomorphism of the retrieved pose [11,58]. Subsequent works have devised approaches to obtain better initialization from image cues [38], more efficient optimization pipelines [79], focused on multiple people [24] or extended the approach to multi-view scenarios [43,24].

While optimization-based approaches do not require images with 3D annotation for training and achieve relatively good registration of details to 2D observations, they tend to suffer from the non-convexity of the problem, being slow and falling into local minima unless provided with a good initialization and accurate 2D observations. In this work, we overcome both these limitations. From one side we only use as input a very coarse person segmentation and image features obtained with standard encoder-decoder architectures. And from the other side, the learned vertex displacements help the optimizer to converge to good solutions (empirically observed) in just a few iterations. On the downside, our approach requires 3D training data, but as we will show in the experimental section, by using synthetic data we manage to generalize well to real images.

Regression. Most current approaches on human body shape recovery consider the direct regression of the shape and pose parameters of the SMPL model [29,6,37,39,59,26,40,70,75,76,18]. As in optimization-based methods, different sorts of 2D image evidence have been used, *e.g.* keypoints [42], keypoints plus silhouette [60] or part segmentation maps [54]. More recently, SMPL parameters have been regressed directly from entire images encoded by pre-trained deep networks (typically ResNet-like) [35,29,59,19,26]. However, regressing the parameters of a low-dimensional parametric model from a single view is always a highly ambiguous problem. This is alleviated by recent works that explore the idea of using hybrid approaches combining optimization and regression [87,34,38,79,51,44]. Very recently, [40] proposed regressing a distribution of parameters instead of having a regression into a single pose and shape representation. In any event, all these works still rely on representing the body shape through low-rank models.

We argue that other shape representations are necessary to model body shape details. This was already discussed in [39], which suggested representing the body shape using all vertices of a template mesh. We will follow the same spirit, although in a completely different learning paradigm. Specifically [39] proposed regressing all points of the body mesh in one single regression pass of a Graph Convolutional Network. This led to noisy outputs that required from post-processing step to smooth the results by fitting the SMPL model. Instead, we propose a novel optimization framework, that leverages on a pre-learned prior that maps image evidence to vertex displacements towards the body shape. We will show that despite its simplicity, this approach surpasses by considerable margins all prior work, and provides smooth while accurate meshes without any post-processing.

2.2 Fitting scans

Classical ICP-based has been used for fitting SMPL with no direct correspondences [62,12,13,64,32,25] or for registration of garments [63,10,41]. Integrating additional knowledge such as pre-computed 3D joints, facial key points [2] and body part segmentation [8] significantly improves the registration quality but these pre-processing steps are prone to error and often require human supervision. Other works initialize correspondences with a learned regressor [80,66,27] and later optimize model parameters. Like us, more recent methods also propose predicting correspondences [9] or body part labels [8] extracted via learnt features. Even though we do not explicitly propose a 3D registration method, LVD is a general algorithm that predicts parametric models. By optimizing these predictions without further correspondences, we surpass other methods that are explicitly designed for 3D registration.

2.3 Neural fields for parametric models

Neural fields [21,67,55,20,90,81] have recently shown impressive results in modeling 3D human shape [15,89,23,55,22,53]. However, despite providing the level of detail that parametric models do not have, they are computationally expensive and difficult to integrate within pose-driven applications given the lack of correspondences. Recent works have already explored possible integrations between implicit and parametric representations for the tasks of 3D reconstruction [33,84], clothed human modeling [73,46,47], or human rendering [61].

We will build upon this direction by framing our method in the pipeline of neural fields. Concretely, we will take the vertices of an unfit mesh and use image features to learn their optimal displacement towards the optimal body shape.

3 Method

We next present our new paradigm for fitting 3D human models. For clarity, we will describe our approach in the problem of 3D human shape reconstruction from a single image. Yet, the formulation we present here is generalizable to the problem of fitting 3D scans, as we shall demonstrate in the experimental section.

3.1 Problem formulation

Given a single-view image $\mathbf{I} \in \mathbb{R}^{H \times W}$ of a person our goal is to reconstruct his/her full body. We represent the body using a 3D mesh $\mathbf{V} \in \mathbb{R}^{N \times 3}$ with N vertices. For convenience (and compatibility with SMPL-based downstream algorithms) the mesh topology will correspond to that of the SMPL model, with $N = 6.890$ vertices and triangular connectivity (13.776 faces). It is important to note that our method operates on the vertices directly and hence it is applicable to other models (such as hands [69]). In particular, we do not use the low dimensional pose and shape parameterizations of such models.

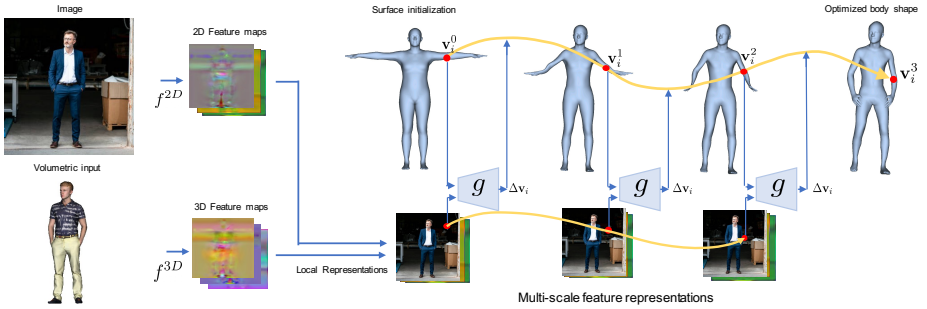

Fig. 2. LVD is a novel framework for estimation of 3D human body where local features drive the direction of vertices iteratively by predicting a per-vertex neural field. At each step t, g takes an input vertex $\mathbf{v}_i^t$ with its corresponding local features, to predict the direction towards its groundtruth position. The surface initialization here follows a T-Posed body, but the proposed approach is very robust to initialization.

3.2 LVD: Learning Vertex Descent

We solve the model fitting problem via an iterative optimization approach with learned vertex descent. Concretely, let $\mathbf{v}_i^t$ be the i-th vertex of the estimated mesh $\mathbf{V}$ at iteration t. Let us also denote by $\mathbf{F} \in \mathbb{R}^{H' \times W' \times F}$ the pixel aligned image features, and by $\mathbf{f}_i$ the F-dimensional vector of the specific features extracted at the projection of $\mathbf{v}_i^t$ on the image plane.

We learn a function $g(\cdot)$ that given the current 3D vertex position, and the image features at its 2D projection, predicts the magnitude and direction of steepest descent towards the ground truth location of the i-th vertex, which we shall denote as $\hat{\mathbf{v}}_i$. Formally:

$$g : (\mathbf{v}_i^t, \mathbf{f}_i) \mapsto \Delta \mathbf{v}_i \, . \tag{1}$$

where $\Delta \mathbf{v}_i \in \mathbb{R}^3$ is a vector with origin at $\mathbf{v}_i^t$ and endpoint at the ground truth $\hat{\mathbf{v}}_i$. In practice, during training, we will apply a component-wise clipping to the ground truth displacements with threshold λ. This stabilizes convergence during the first training iterations.

We learn the vertex descent function $g(\cdot)$ using a joint ensemble of per-vertex neural field networks, which we describe in Sect. 3.3. Once this mapping is learned, we can define the following update rule for our learned optimization:

$$\mathbf{v}_i^{t+1} = \mathbf{v}_i^t + \Delta \mathbf{v}_i \, . \tag{2}$$

The reconstruction problem then entails iterating over Eq. 2 until the convergence of $\Delta \mathbf{v}_i$. Fig 2 depicts an overview of the approach.

Note that in essence we are replacing the standard gradient descent rule with a learned update that is locally computed at every vertex. As we will empirically demonstrate in the results section, despite its simplicity, the proposed approach allows for fast and remarkable convergence rates, typically requiring only 4 to 6 iterations no matter how the mesh vertices are initialized.

Uncertainty estimation. An interesting outcome of our approach is that it allows estimating the uncertainty of the estimated 3D shape, which could be useful in downstream applications that require a confidence measure. For estimating the uncertainty of a vertex $\mathbf{v}_i$, we compute the variance of the points after perturbing them and letting the network converge. After this process, we obtain the displacements $\Delta\mathbf{x}_j^i$ between perturbed points $\mathbf{x}_j$ and the mesh vertex $\mathbf{v}_i$ predicted initially. We then define the uncertainty of $\mathbf{v}_i$ as:

$$U(\mathbf{v}_i) = \text{std}(\{\mathbf{x}_j + \Delta\mathbf{x}_j^i\}_{j=1}^M) \ . \tag{3}$$

In Figs. 1 and 4 we represent the uncertainty of the meshes in dark blue. Note that the most uncertain regions are typically localized on the feet and hands.

3.3 Network architecture

The LVD architecture has two main modules, one that is responsible of extracting local image features and the other of learning the optimal vertices' displacement.

Local features. Following recent approaches [71,72], the local features $\mathbf{F}$ are learned with an encoder-decoder Hourglass network trained from scratch. Given a vertex $\mathbf{v}_i^t = (x_i^t, y_i^t, z_i^t)$ and the input image $\mathbf{I}$, these features are estimated as:

$$f : (\mathbf{I}, \pi(\mathbf{v}_i^t), z_i^t) \mapsto \mathbf{f}_i \ , \tag{4}$$

where $\pi(\mathbf{v})$ is a weak perspective projection of $\mathbf{v}$ onto the image plane. We condition $f(\cdot)$ with the depth z_i^t of the vertex to generate depth-aware local features. A key component of LVD is Predicting vertex displacements based on local features, which have been shown to produce better geometric detail, even from small training sets [71,72,16]. Indeed, this is one of our major differences compared to previous learning approaches for human shape estimation relying on parametric body models. These methods learn a mapping from a full image to global shape parameters (two disjoint spaces), which is hard to learn, and therefore they are unable to capture the local details. This results in poor image overlap between the recovered shape and the image as can be seen in Fig. 1.

Network field. In order to implement the function $g(\cdot)$ in Eq. 1 we follow recent neural field approaches [50,56] and use a simple 3-layer MLP that takes as input the current estimate of each vertex $\mathbf{v}_i^t$ plus its local F-dimensional local feature $\mathbf{f}_i$ and predicts the displacement $\Delta\mathbf{v}_i$.

3.4 Training LVD

Training the proposed model entails learning the parameters of the functions $f(\cdot)$ and $g(\cdot)$ described above. For this purpose, we will leverage a synthetic dataset of images of people under different clothing and body poses paired with the corresponding SMPL 3D body registrations. We will describe this dataset in the experimental section.

In order to train the network, we proceed as follows: Let us assume we are given a ground truth body mesh $\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_N]$ and its corresponding image $\mathbf{I}$. We then randomly sample M 3D points $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$, using a combination of points uniformly sampled in space and points distributed near the surface. Each of these points, jointly with the input image $\mathbf{I}$ is fed to the LVD model which predicts its displacement w.r.t. all ground truth SMPL vertices. Then, the loss associated with $\mathbf{x}_i$ is computed as:

$$\mathcal{L}(\mathbf{x}_i) = \sum_{j=1}^{N} \|\Delta\mathbf{x}_i^j - \hat{\Delta}\mathbf{x}_i^j\|_1 \ , \tag{5}$$

where $\Delta\mathbf{x}_i^j$ is the predicted displacement between $\mathbf{x}_i$ and $\hat{\mathbf{v}}_j$ and $\hat{\Delta}\mathbf{x}_i^j$ the ground truth displacement. $\|\cdot\|_1$ is the L1 distance. Note that by doing this, we are teaching our network to predict the displacement of any point in space to all vertices of the mesh. We found that this simple loss was sufficient to learn smooth but accurate body prediction. Remarkably, no additional regularization losses enforcing geometry consistency or anthropomorphism were required.

The reader is referred to the Supplemental Material for additional implementation and training details.

3.5 Application to 3D scan registration

The pipeline we have just described can be readily applied to the problem of fitting the SMPL mesh to 3D scans of clothed people or fitting the MANO model [69] to 3D scans of hands. The only difference will be in the feature extractor $f(\cdot)$ of Eq. 4, which will have to account for volumetric features. That is, if $\mathbf{X}$ is a 3D voxelized input scan, the feature extractor for a vertex $\mathbf{v}_i$ will be defined as:

$$f^{3D} : (\mathbf{X}, \mathbf{v}_i) \mapsto \mathbf{f}_i \ , \tag{6}$$

where again, $\mathbf{f}_i$ will be an F-dimensional feature vector. For the MANO model, the number of vertices of the mesh is $N = 778$. In the experimental section, we will show the adaptability of LVD to this scan registration problem.

4 Connection to classical model based fitting

Beyond its good performance, we find the connection of LVD to classical optimization based methods interesting, and understanding its relationship can be important for future improvements and extensions of LVD. Optimization methods for human shape recovery optimize model parameters to match image features such as correspondences [28,11,58], silhouettes [77,78]. See [65] for an in-depth discussion of optimization-based model based fitting.

Optimization based. These methods minimize a scalar error $e(\mathbf{p}) \in \mathbb{R}$ with respect to human body parameters $\mathbf{p}$. Such scalar error is commonly obtained from a sum of squares error $e = \mathbf{e}(\mathbf{p})^T\mathbf{e}(\mathbf{p})$. The error vector $\mathbf{e} \in \mathbb{R}^{dN}$ contains

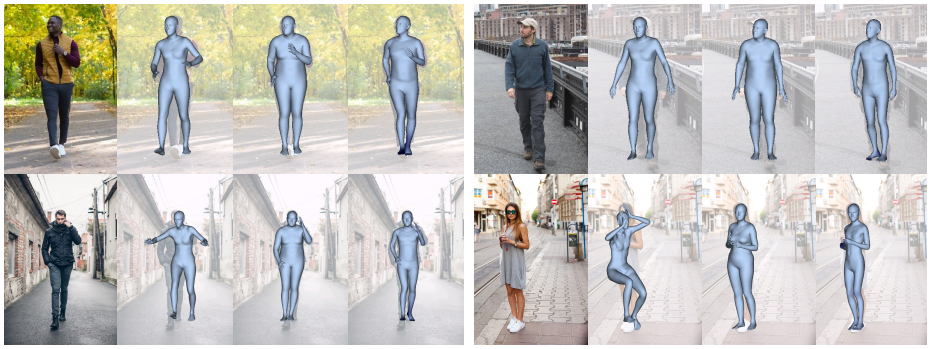

Fig. 3. Comparison of LVD to body shape estimation baselines. As a first baseline, we train a ResNet [30] (using the same data as for LVD) to predict the SMPL parameters. This approach fails to generalize to novel poses and shapes. We also compare LVD to Sengupta *et al* [76], which perform well on real images, even though the predicted shapes do not fit perfectly the silhouettes of the people. See also quantitative results in Table 1.

the d dimensional residuals for the N vertices of a mesh, which typically correspond to measuring how well the projected $i-th$ vertex in the mesh fits the image evidence (*e.g*, matching color of the rendered mesh vs image color). To minimize e one can use gradient descent, Gauss-Newton or Levenberg-Marquadt (LM) optimizer to find a descent direction for human parameters $\mathbf{p}$, but ultimately the direction is obtained from local image gradients as we will show. Without loss of generality, we can look at the individual residual incurred by one vertex $\mathbf{e}_i \in \mathbb{R}^d$, although bear in mind that an optimization routine considers all residuals simultaneously (the final gradient will be the sum of individual residual gradients or step directions in the case of LM type optimizers). The gradient of a single residual can be computed as

$$\nabla_{\mathbf{p}} e_i = \frac{\partial(\mathbf{e}_i^T \mathbf{e}_i)}{\partial \mathbf{p}} = 2 \left[\frac{\partial \mathbf{e}_i}{\partial \mathbf{v}_i} \frac{\partial \mathbf{v}_i}{\partial \mathbf{p}} \right]^T \mathbf{e}_i \tag{7}$$

where the matrices that play a critical role in finding a good direction are the error itself $\mathbf{e}_i$, and $\frac{\partial \mathbf{e}_i}{\partial \mathbf{v}_i}$ which is the Jacobian matrix of the i-th residual with respect to the i-th vertex (the Jacobian of the vertex w.r.t. to parameters $\mathbf{p}$ is computed from the body model and typically helps to restrict (small) vertex displacements to remain within the space of human shapes). When residuals are based on pixel differences (common for rendering losses and silhouette terms) obtaining $\frac{\partial \mathbf{e}_i}{\partial \mathbf{v}_i}$ requires computing image gradients via finite differences. Such classical gradient is only meaningful once we are close to the solution.

Learned Vertex Descent. In stark contrast, our neural fields compute a learned vertex direction, with image features that have a much higher receptive field than a classical gradient. This explains why our method converges much faster and more reliably than classical approaches. To continue this analogy, our

Fig. 4. SMPL reconstruction on images on-the-wild. For each method, we show the reconstruction in posed and canonical space. While previous works focus on pose estimation, they are prone to generate always an average body shape. In contrast, LVD generates a much richer distribution of body shapes as shown in the right-most column.

network learns to minimize the following objective error (for a single vertex)

$$e_i^{LVD} = \mathbf{e}_i^T \mathbf{e}_i = (\mathbf{v}_i - \mathbf{v}_i^{gt})^T (\mathbf{v} - \mathbf{v}_i^{gt}) \tag{8}$$

whose vertex gradient $\nabla_{\mathbf{v}_i} e_i^{LVD}$ points directly to the ground truth vertex $\mathbf{v}_i^{gt}$. In fact, our LVD is trained to learn the step size as well as the direction. What is even more remarkable, and surprising to us, is that we do not need a body model to constraint the vertices. That is, during optimization, we do not need to compute $\frac{\partial \mathbf{v}_i}{\partial \mathbf{p}}$, and project the directions to the space of valid human shapes. Since LVD has been learned from real human shapes, it automatically learns a prior, making the model very simple and fast during inference.

5 Experiments

We next evaluate the performance of LVD in the tasks of 3D human reconstruction from a single image and 3D registration. Additionally, we will provide empirical insights about the convergence of the algorithm and its shape expressiveness compared to parametric models.

Data. We use the RenderPeople, AXYZ and Twindom datasets [68,7,82], which consist of 767 3D scans. We first obtain SMPL registrations and manually an-

Table 1. Single-view SMPL estimation from LVD and baselines [58,38,70,19,40,76] in the BUFF Dataset [88]. The experiments take into account front, side and back views from the original scans and show that LVD outperforms all baselines in all scenarios and metrics except for back views. *We also report the results of PIFu, although note that this is a model-free approach in contrast to ours and the rest of the baselines, which recover the SMPL model. All errors are reported in mm.

	Vertex-to-Vertex				Vertex-to-Surface				V2V	V2S
Viewing angle:	0°	90°	180°	270°	0°	90°	180°	270°	Avg.	Avg.
*PIFu [71]	36.71	40.55	72.57	39.23	35.16	39.04	71.38	38.51	47.18	46.05
SMPL-X [58]	41.30	77.03	61.40	92.50	40.00	75.68	60.27	91.30	68.07	66.81
SPIN [38]	31.96	42.10	53.93	44.54	30.68	40.87	52.86	43.30	43.13	41.92
FrankMocap [70]	27.24	43.33	47.36	42.36	25.70	41.93	46.15	40.85	40.07	38.66
ExPose [19]	26.07	40.83	54.42	44.34	24.61	39.60	53.23	43.16	41.41	40.15
ProHMR [40]	39.55	49.26	55.42	46.03	38.42	48.18	54.41	44.88	47.56	46.47
Sengupta [76]	27.70	51.10	**40.11**	53.28	25.96	49.77	**38.80**	52.03	43.05	41.64
LVD	**25.44**	**38.24**	54.55	**38.10**	**23.94**	**37.05**	53.55	**36.94**	**39.08**	**37.87**

notate the correct fits. Then, we perform an aggressive data augmentation by synthetically changing body pose, shape and rendering several images per mesh from different views and illuminations. By doing, this we collect a synthetic dataset of $\sim 600k$ images which we use for training and validation. Test will be performed on real datasets. Please see Suppl. Mat. for more details about the construction of this dataset.

5.1 3D Body shape estimation from a single image

We evaluate LVD in the task of body shape estimation and compare it against Sengupta *et al* [76], which uses 2D edges and joints to extract features that are used to predict SMPL parameters. We also compare it against a model that estimates SMPL pose and shape parameters given an input image. We use a pre-trained ResNet-18 [30] that is trained on the exact same data as LVD. This approach fails to capture the variability of body shapes and does not generalize well to new poses. We attribute this to the limited amount of data (only a few hundred 3D scans), with every image being a training data point, while in LVD every sampled 3D point counts as one training example. Figure 3 shows qualitative results on in-the-wild images. The predictions of LVD also capture the body shape better than those of Sengupta *et al* [76] and project better to the silhouette of the input person.

Even though our primary goal is not pose estimation, we also compare LVD against several recent state-of-the-art model-based methods [58,38,70,19,40] on the BUFF dataset, which has 9612 textured scans of clothed people. We have uniformly sampled 480 scans and rendered images at four camera views. Table 1 summarizes the results in terms of the Vertex-to-Vertex (V2V) and Vertex-to-Surface (V2S) distances. The table also reports the results of PIFu [71], although we should take this merely as a reference, as this is a model-free approach, while the rest of the methods in the Table are model-based. Figure 4 shows qualitative

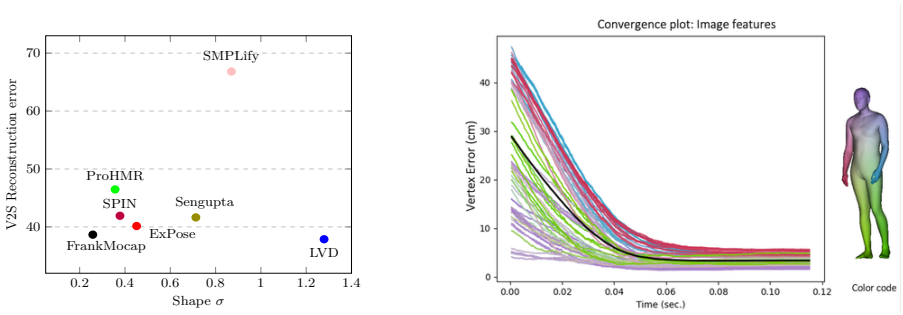

Fig. 5. *Left:* Variability of predicted body shape parameters (x-axis) with respect to vertex error (y-axis, lower is better) for works that fit SMPL to images. Previous approaches have mostly focused on the task of pose estimation. LVD, instead, aims to represent a more realistic distribution of predicted body shapes. *Right:* Convergence analysis of the proposed optimization, showing the distance from each SMPL vertex to the groundtruth scan during optimization, averaged for 200 examples of the BUFF dataset. The first iteration also includes the time to obtain the local representations used during the rest of the optimization. Each line color encodes a different body region and the black line shows the average error of all vertices.

results on in-the-wild images. With this experiment, we want to show that previous works on pose and shape estimation tend to predict average body shapes. In contrast, our approach is able to reconstruct high-fidelity bodies for different morphotypes. It should be noted that our primary goal is to estimate accurate body shape, and our training data does not include extreme poses. Generalizing LVD to complex poses will most likely require self-supervised frameworks with in the wild 2D images like current SOTA [35,40,19,38] , but this is out of the scope of this paper, and leave it for future work.

Finally, it is worth to point that some of the baselines [58,70,38,19] require 2D keypoint predictions, for which we use the publicly available code of OpenPose [14]. In contrast, LVD requires coarse image segmentations to mask the background out because we trained LVD on renders of 3D scans without background. In any event, we noticed that our model is not particularly sensitive to the quality of input masks, and can still generate plausible body shapes with noisy masks (see Supp. Mat.).

5.2 Shape expressiveness and convergence analysis

We further study the ability of all methods to represent different body shapes. For this, we obtain the SMPL shape for our model and pose estimation baselines in Tab. 1 and fit the SMPL model with 300 shape components. We then calculate the standard deviation σ_2 of the second PCA component, responsible for the shape diversity. Fig. 5 (left) depicts the graph of shape σ_2 vs. V2S error. It is clearly shown that LVD stands out in its capacity to represent different shapes. In contrast, most previous approaches have a much lower capacity to recover different body shapes, with a σ_2 value 3 times smaller than ours.

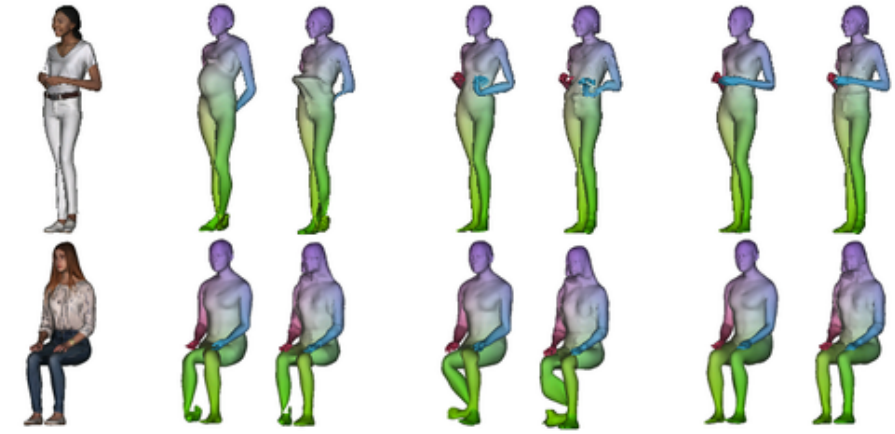

Fig. 6. SMPL and SMPL+D registration of 3D scans from LVD in comparison to LoopReg and IP-Net.

Table 2. Evaluation on SMPL and SMPL+D registration on the RenderPeople Dataset [68]. The initial SMPL estimation from LVD is already very competitive against baselines [9,8]. By using these predictions as initialization for SMPL/SMPL+D registration, we obtain $\sim$ 28.4% and $\sim$ 37.7% relative improvements with respect to the second-best method[8] in joint and SMPL vertex distances respectively.

		Forward pass	SMPL Registration				SMPL+D Registration			
		LVD	No corresp.	LoopReg [9]	IP-Net [8]	LVD	No corresp.	LoopReg [9]	IP-Net [8]	LVD
SMPL error	Joint [cm]	5.89	16.6	9.33	3.60	**2.53**	16.6	9.32	3.63	**2.60**
	Vertex [cm]	6.27	21.3	12.2	5.03	**3.00**	21.3	12.3	5.20	**3.24**
Recons. to Scan	V2V [mm]	8.98	12.51	10.35	8.84	**8.16**	1.45	1.43	1.21	**1.14**
	V2S [mm]	6.61	10.53	8.19	6.61	**5.87**	0.72	0.69	0.53	**0.47**
Scan to Recons.	V2V [mm]	12.6	16.92	14.27	12.25	**11.31**	8.53	8.01	7.22	**6.88**
	V2S [mm]	9.31	13.75	10.49	8.45	**7.43**	4.22	3.47	2.78	**2.44**

We also perform an empirical convergence analysis of LVD. Fig. 5 (right) plots the average V2V error (in mm) vs time, computed when performing shape inference for 200 different samples of the BUFF dataset. Note that the optimization converges at a tenth of a second using a GTX 1080 Ti GPU. The total computation time is equivalent to 6 iterations of our algorithm. The color-coded 3D mesh on the side of Fig. 5 (right) shows in which parts of the body the algorithm suffers the most. These areas are concentrated on the arms. Other regions that hardly become occluded, such as torso or head have the lowest error. The average vertex error is represented with a thicker black line.

Finally, we measure the sensitiveness of the convergence to different initializations of the body mesh. We randomly sampled 1K different initializations from the AMASS dataset [49] and analized the deviation of the converged reconstructions, obtaining a standard deviation of the SMPL surface vertices of only $\sigma = 1.2$mm across all reconstructions. We credit this robustness to the dense supervision during training, which takes input points from a volume on

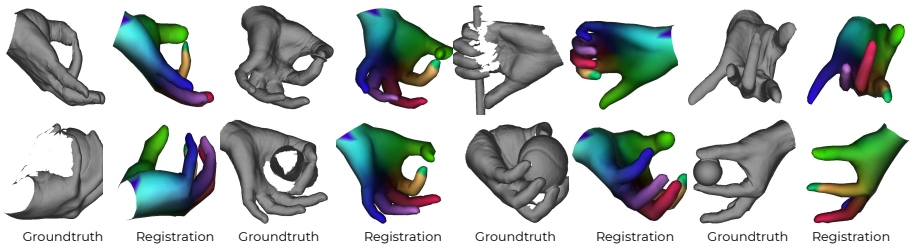

Fig. 7. Registration of MANO from input pointclouds. We include more visuals and qualitative comparisons with baselines in Supplementary Material.

Table 3. Registration of MANO [69] from input 3D pointclouds of hands.

	MANO Error		Reconstruction to Scan		Scan to Reconstruction	
Method	Joint [cm]	Vertex [cm]	V2V [mm]	V2S [mm]	V2V [mm]	V2S [mm]
No corresp.	6.49	7.05	5.31	5.28	8.06	6.40
IP-Net [8]	1.44	1.73	3.29	3.23	6.17	4.08
LVD	**.76**	**.96**	**2.73**	**2.65**	**5.62**	**3.33**

the 3D space, as well as around the groundtruth body surface. See the Supp. Video for qualitative evidence.

5.3 3D Body Registration

LVD is designed to be general and directly applicable for different tasks. We analyze the performance of LVD on the task of SMPL and SMPL+D registration on 3D point-clouds of humans. This task consists in initially estimating the SMPL mesh (which we do iterating our approach) and then running a second minimization of the Chamfer distance to fit SMPL and SMPL+D. The results are reported in Tab. 2, where we compare against LoopReg [9], IP-Net [8], and also against the simple baseline of registering SMPL with no correspondences starting from a T-Posed SMPL. Besides the V2V and V2S metrics (bi-directional), we also report the Joint error (predicted using SMPL's joint regressor), and the distance between ground truth SMPL vertices and their correspondences in the registered mesh (Vertex distance). Note that again, LVD consistently outperforms the rest of the baselines. This is also qualitatively shown in Fig. 6.

5.4 3D Hand Registration

The proposed approach is directly applicable to any statistical model, thus we also test it in the task of registration of MANO [69] from input point-clouds of hands, some of them incomplete. For this experiment, we do not change the network hyperparameters and only update the number of vertices to predict (778 for MANO). We test this task on the MANO [69] dataset, where the approach

also outperforms IP-Net[8], trained on the same data. Tab. 3 summarizes the performance of LVD and baselines, and qualitative examples are shown in Fig. 7. Note that LVD shows robustness even in situations with partial point clouds.

6 Conclusion

We have introduced Learned Vertex Descent, a novel framework for human shape recovery where vertices are iteratively displaced towards the predicted body surface. The proposed method is lightweight, can work real-time and surpasses previous state-of-the-art in the tasks of body shape estimation from a single view or 3D scan registration, of both the full body and hands. Being so simple, easy to use and effective, we believe LVD can be an important building block for future model-fitting methods. Future work will focus in self-supervised training formulations of LVD for predicting body shape in difficult poses and scenes, and tackling multi-person scenes efficiently.

Acknowledgements This work is supported in part by the Spanish government with the project MoHuCo PID2020-120049RB-I00, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (Emmy Noether Programme, project: Real Virtual Humans) and German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. Gerard Pons-Moll is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. We thank NVIDIA for donating GPUs.

References

1. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single RGB camera. In: CVPR (jun 2019)
2. Alldieck, T., Magnor, M., Bhatnagar, B.L., Theobalt, C., Pons-Moll, G.: Learning to reconstruct people in clothing from a single rgb camera. In: CVPR (2019)
3. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: CVPR (2018)
4. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: ICCV. IEEE (oct 2019)
5. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: shape completion and animation of people. SIGGRAPH (2005)
6. Arnab, A., Doersch, C., Zisserman, A.: Exploiting temporal context for 3d human pose estimation in the wild. In: CVPR. pp. 3395–3404 (2019)
7. Axyz dataset. `https://secure.axyz-design.com/`
8. 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)
9. 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)
10. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: ICCV (2019)

11. 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)
12. Bogo, F., Romero, J., Loper, M., Black, M.J.: Faust: Dataset and evaluation for 3d mesh registration. In: CVPR. pp. 3794–3801 (2014)
13. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic faust: Registering human bodies in motion. In: CVPR. pp. 6233–6242 (2017)
14. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. PAMI **43**(1), 172–186 (2019)
15. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: ICCV (2021)
16. Chibane, J., Pons-Moll, G.: Implicit feature networks for texture completion from partial 3d data. In: ECCV. pp. 717–725. Springer (2020)
17. Chibane, J., Pons-Moll, G., et al.: Neural unsigned distance fields for implicit function learning. NeurIPS (2020)
18. Choutas, V., Müller, L., Huang, C.H.P., Tang, S., Tzionas, D., Black, M.J.: Accurate 3d body shape regression using metric and semantic attributes. In: CVPR. pp. 2718–2728 (2022)
19. 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)
20. Corona, E., Hodan, T., Vo, M., Moreno-Noguer, F., Sweeney, C., Newcombe, R., Ma, L.: Lisa: Learning implicit shape and appearance of hands. arXiv preprint arXiv:2204.01695 (2022)
21. Corona, E., Pumarola, A., Alenya, G., Pons-Moll, G., Moreno-Noguer, F.: Smplicit: Topology-aware generative model for clothed people. In: CVPR. pp. 11875–11885 (2021)
22. Deng, B., Lewis, J.P., Jeruzalski, T., Pons-Moll, G., Hinton, G., Norouzi, M., Tagliasacchi, A.: Nasa neural articulated shape approximation. In: ECCV. pp. 612–628. Springer (2020)
23. Deprelle, T., Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: Learning elementary structures for 3d shape generation and matching. arXiv preprint arXiv:1908.04725 (2019)
24. Dong, Z., Song, J., Chen, X., Guo, C., Hilliges, O.: Shape-aware multi-person pose estimation from multi-view images. In: ICCV. pp. 11158–11168 (2021)
25. Dyke, R.M., Lai, Y.K., Rosin, P.L., Tam, G.K.: Non-rigid registration under anisotropic deformations. Computer Aided Geometric Design **71**, 142–156 (2019)
26. Georgakis, G., Li, R., Karanam, S., Chen, T., Košecká, J., Wu, Z.: Hierarchical kinematic human mesh recovery. In: ECCV. pp. 768–784. Springer (2020)
27. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: 3d-coded: 3d correspondences by deep deformation. In: ECCV. pp. 230–246 (2018)
28. Guan, P., Weiss, A., Balan, A.O., Black, M.J.: Estimating human shape and pose from a single image. In: ICCV. IEEE (2009)
29. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: CVPR (2019)
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
31. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: Animation-ready clothed human reconstruction revisited. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11046–11056 (2021)

32. Hirshberg, D.A., Loper, M., Rachlin, E., Black, M.J.: Coregistration: Simultaneous alignment and modeling of articulated 3d shape. In: ECCV. pp. 242–255. Springer (2012)
33. Huang, Z., Xu, Y., Lassner, C., Li, H., Tung, T.: Arch: Animatable reconstruction of clothed humans. In: CVPR (2020)
34. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. arXiv preprint arXiv:2004.03686 (2020)
35. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
36. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
37. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: CVPR. pp. 5253–5263 (2020)
38. 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)
39. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)
40. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: ICCV. pp. 11605–11614 (2021)
41. Lahner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: ECCV (2018)
42. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: CVPR (2017)
43. Li, Z., Oskarsson, M., Heyden, A.: 3d human pose and shape estimation through collaborative learning and multi-view model-fitting. In: WCACV. pp. 1888–1897 (2021)
44. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR. pp. 1954–1963 (2021)
45. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ToG (2015)
46. Ma, Q., Saito, S., Yang, J., Tang, S., Black, M.J.: Scale: Modeling clothed humans with a surface codec of articulated local elements. In: CVPR. pp. 16082–16093 (2021)
47. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: CVPR. pp. 6469–6478 (2020)
48. Ma, Q., Yang, J., Tang, S., Black, M.J.: The power of points for modeling humans in clothing. In: ICCV. pp. 10974–10984 (2021)
49. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: ICCV (2019)
50. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: CVPR (2019)
51. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: ECCV. pp. 752–768. Springer (2020)
52. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. pp. 483–499. Springer (2016)
53. Niemeyer, M., Mescheder, L., Oechsle, M., Geiger, A.: Occupancy flow: 4d reconstruction by learning particle dynamics. In: CVPR. pp. 5379–5389 (2019)

54. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 3DV. IEEE (2018)
55. Pan, J., Han, X., Chen, W., Tang, J., Jia, K.: Deep mesh reconstruction from single rgb images via topology modification networks. In: ICCV. pp. 9964–9973 (2019)
56. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: CVPR (2019)
57. Patel, C., Liao, Z., Pons-Moll, G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In: CVPR. IEEE (jun 2020)
58. 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)
59. Pavlakos, G., Kolotouros, N., Daniilidis, K.: Texturepose: Supervising human mesh estimation with texture consistency. In: ICCV. pp. 803–812 (2019)
60. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: Learning to estimate 3d human pose and shape from a single color image. In: CVPR. pp. 459–468 (2018)
61. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR. pp. 9054–9063 (2021)
62. Pishchulin, L., Wuhrer, S., Helten, T., Theobalt, C., Schiele, B.: Building statistical shape spaces for 3d human modeling. Pattern Recognition **67**, 276–286 (2017)
63. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: ClothCap: Seamless 4D clothing capture and retargeting. SIGGRAPH **36**(4) (2017)
64. Pons-Moll, G., Romero, J., Mahmood, N., Black, M.J.: Dyna: A model of dynamic human shape in motion. ToG **34**(4), 1–14 (2015)
65. Pons-Moll, G., Rosenhahn, B.: Model-Based Pose Estimation, chap. 9, pp. 139–170. Springer (2011)
66. Pons-Moll, G., Taylor, J., Shotton, J., Hertzmann, A., Fitzgibbon, A.: Metric regression forests for correspondence estimation. IJCV **113**(3), 163–175 (2015)
67. Prokudin, S., Black, M.J., Romero, J.: Smplpix: Neural avatars from 3d human models. In: WCACV. pp. 1810–1819 (2021)
68. Renderpeople dataset. `https://renderpeople.com/`
69. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ToG (2017)
70. 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)
71. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV (2019)
72. Saito, S., Simon, T., Saragih, J., Joo, H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In: CVPR (2020)
73. Saito, S., Yang, J., Ma, Q., Black, M.J.: Scanimate: Weakly supervised learning of skinned clothed avatar networks. In: CVPR. pp. 2886–2897 (2021)
74. 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)
75. Sengupta, A., Budvytis, I., Cipolla, R.: Synthetic training for accurate 3d human pose and shape estimation in the wild. BMVC (2020)

76. Sengupta, A., Budvytis, I., Cipolla, R.: Hierarchical kinematic probability distributions for 3d human shape and pose estimation from images in the wild. In: ICCV. pp. 11219–11229 (2021)
77. Sigal, L., Balan, A., Black, M.: Combined discriminative and generative articulated pose and non-rigid shape estimation. NeurIPS **20**, 1337–1344 (2007)
78. Sminchisescu, C., Triggs, B.: Covariance scaled sampling for monocular 3d body tracking. In: CVPR. vol. 1, pp. I–I. IEEE (2001)
79. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: ECCV. pp. 744–760. Springer (2020)
80. Taylor, J., Shotton, J., Sharp, T., Fitzgibbon, A.: The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In: CVPR. pp. 103–110. IEEE (2012)
81. Tiwari, G., Antic, D., Lenssen, J.E., Sarafianos, N., Tung, T., Pons-Moll, G.: Posendf: Modeling human pose manifolds with neural distance fields. In: European Conference on Computer Vision (ECCV). Springer (October 2022)
82. Twindom dataset. `https://web.twindom.com/`
83. Wu, Y., He, K.: Group normalization. In: ECCV. pp. 3–19 (2018)
84. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Chore: Contact, human and object reconstruction from a single rgb image. In: European Conference on Computer Vision (ECCV). Springer (October 2022)
85. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: CVPR. pp. 6184–6193 (2020)
86. 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)
87. Zanfir, A., Bazavan, E.G., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Neural descent for visual 3d human pose and shape. In: CVPR. pp. 14484–14493 (2021)
88. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: CVPR (2017)
89. Zheng, Z., Yu, T., Liu, Y., Dai, Q.: Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. PAMI (2021)
90. Zhou, K., Bhatnagar, B., Lenssen, J.E., Pons-Moll, G.: Toch: Spatio-temporal object correspondence to hand for motion refinement. arXiv preprint arXiv:2205.07982 (2022)

SUPPLEMENTARY MATERIAL

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 [52] trained from scratch with 4 stacks and batch normalization replaced with group normalization [83]. 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 [74], 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 [36] with $\beta_1 = 0.9$ $\beta_2 = 0.999$. When considering point-clouds as input we train an IF-Net backbone[16] 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 [69], 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.

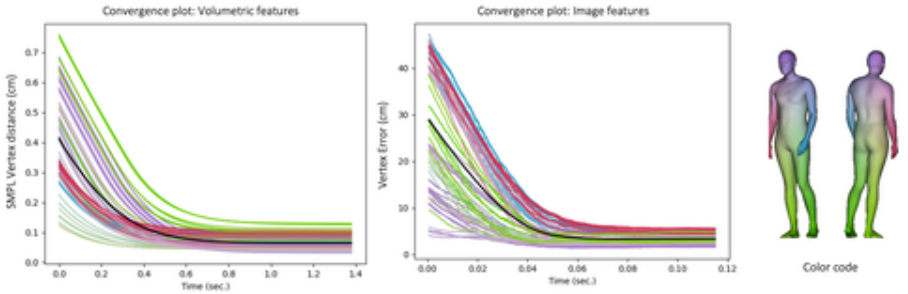

Fig. 1. Convergence plot of the proposed optimization, for voxel-based experiments in comparison to image-based reconstruction. In comparison with the reported results on image-based reconstruction (which also are shown in the main paper), volumetric reconstruction takes almost a second to converge with 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[86]. 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 [68,7,82], 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.

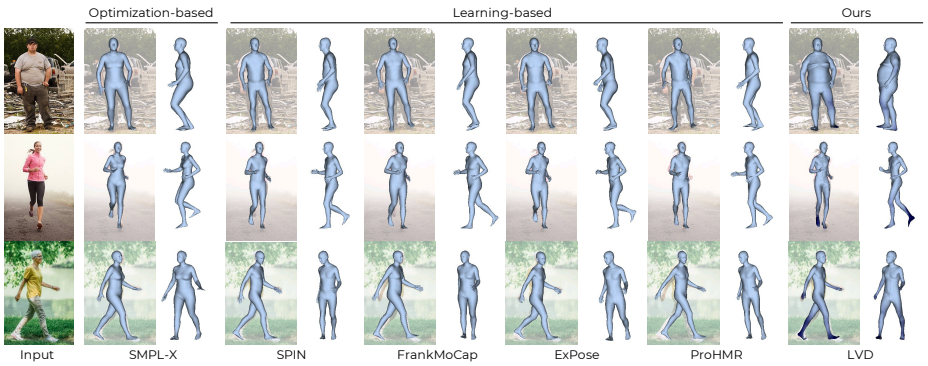

Fig. 4. Qualitative comparisons with more methods. For each method, we show front and side views of the reconstruction.

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 [86] 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 [11,58] are often accurate, but have severe failure cases and are slow. On the other hand, learning based methods [38,70,19,40] 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.

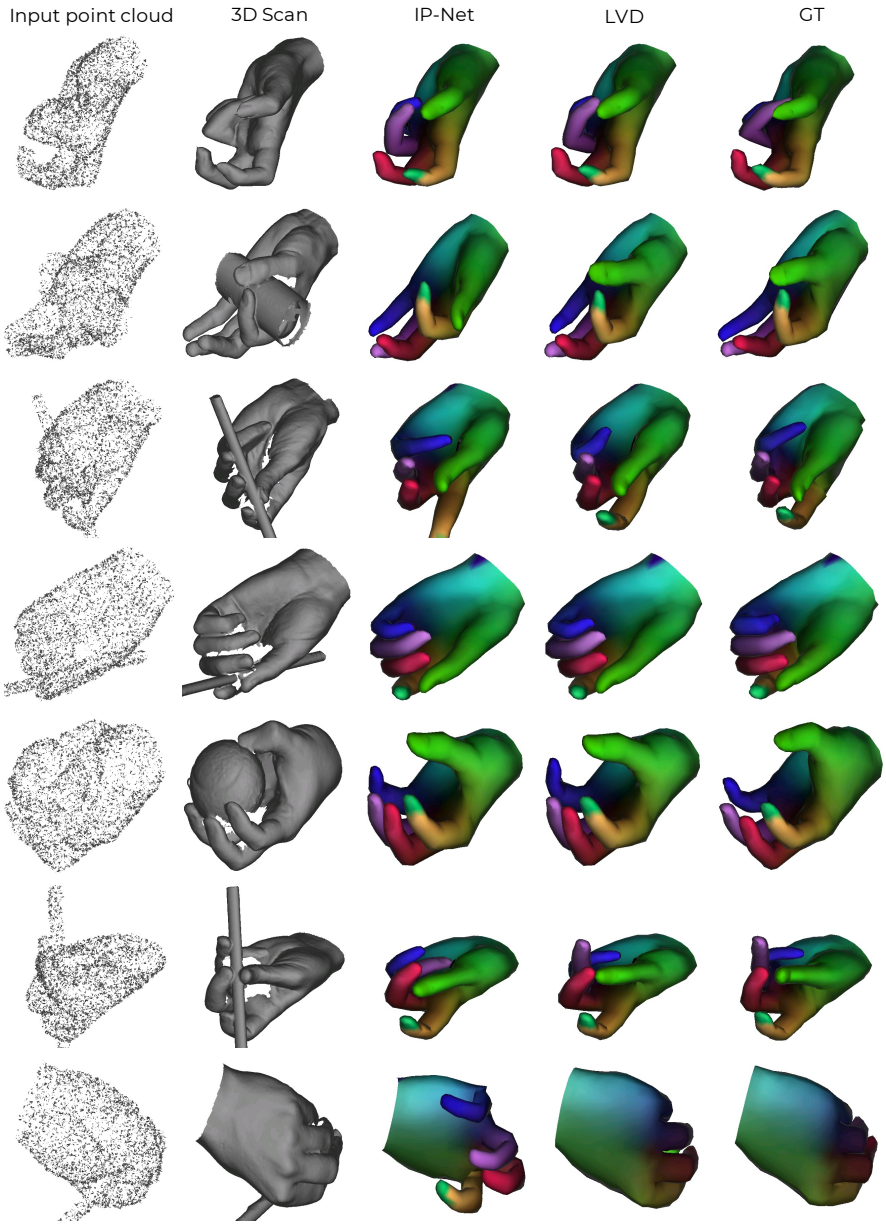

Fig. 6. Registration of 3D Hands using MANO [69]. The input to IP-Net [8] or LVD is the input point cloud in the left column, while the groundtruth 3D scan is shown in the second column. IP-Net performs similarly well in most cases, but is most confused in the presence of other objects or very noisy pointclouds.

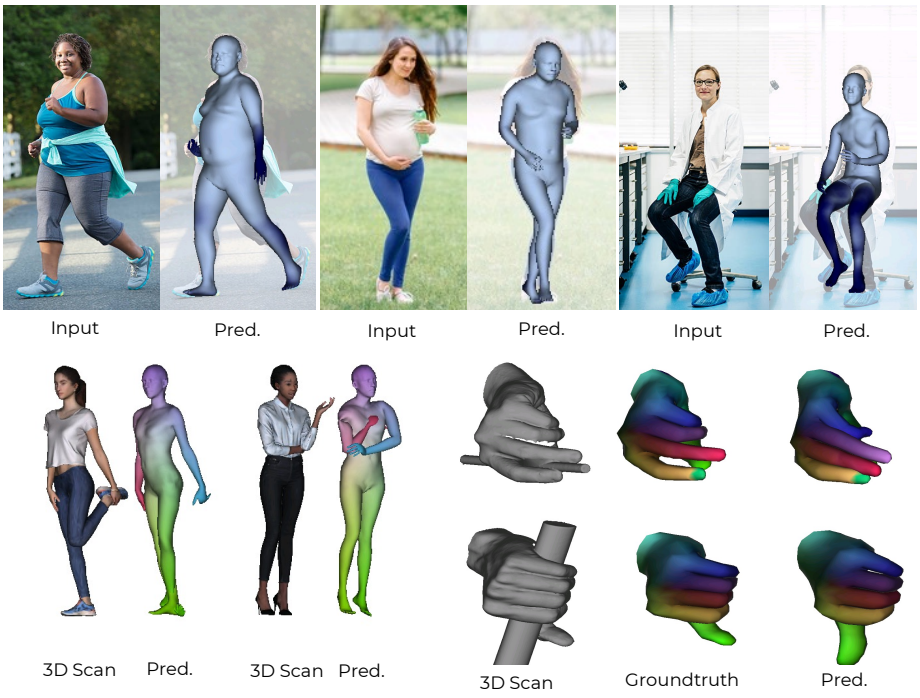

Fig. 7. Failure cases from LVD in body shape estimation from single view images (first row), 3D registration of humans from point clouds (second row - left) and 3D registration from hands (second row - right). See Section 4 for more details.

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 [8]. 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.