On Self-Contact and Human Pose

Lea Müller[1], Ahmed A. A. Osman[1], Siyu Tang[2], Chun-Hao P. Huang[1], Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems, Tübingen [2]ETH Zürich

{lea.mueller, ahmed.osman, stang, paul.huang, black}@tuebingen.mpg.de

Abstract

People touch their face 23 times an hour, they cross their arms and legs, put their hands on their hips, etc. While many images of people contain some form of self-contact, current 3D human pose and shape (HPS) regression methods typically fail to estimate this contact. To address this, we develop new datasets and methods that significantly improve human pose estimation with self-contact. First, we create a dataset of 3D Contact Poses (3DCP) *containing SMPL-X bodies fit to 3D scans as well as poses from AMASS, which we refine to ensure good contact. Second, we leverage this to create the* Mimic-The-Pose (MTP) *dataset of images, collected via Amazon Mechanical Turk, containing people mimicking the 3DCP poses with self-contact. Third, we develop a novel HPS optimization method, SMPLify-XMC, that includes contact constraints and uses the known 3DCP body pose during fitting to create near ground-truth poses for MTP images. Fourth, for more image variety, we label a dataset of in-the-wild images with* Discrete Self-Contact (DSC) *information and use another new optimization method, SMPLify-DC, that exploits discrete contacts during pose optimization. Finally, we use our datasets during SPIN training to learn a new 3D human pose regressor, called* TUCH (Towards Understanding Contact in Humans). *We show that the new self-contact training data significantly improves 3D human pose estimates on withheld test data and existing datasets like 3DPW. Not only does our method improve results for self-contact poses, but it also improves accuracy for non-contact poses. The code and data are available for research purposes at* https://tuch.is.tue.mpg.de.

1. Introduction

Self-contact takes many forms. We touch our bodies both consciously and unconsciously [24]. For the major limbs, contact can provide physical support, whereas we touch our faces in ways that convey our emotional state. We perform self-grooming, we have nervous gestures, and we communicate with each other through combined face and hand motions (e.g. "shh"). We may wring our hands when worried, cross our arms when defensive, or put our hands behind our head when confident. A Google search for "sitting person" or "thinking pose" for example, will return images, the majority of which, contain self-contact.

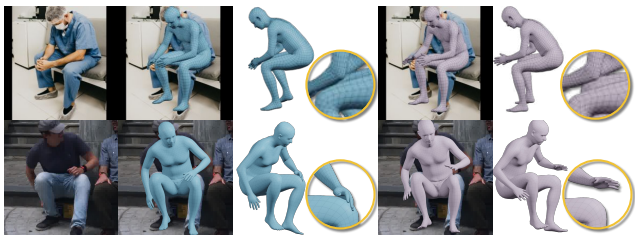

Figure 1. The first column shows images containing self-contact. In blue (left), results of TUCH, compared to SPIN results in violet (right). When rendered from the camera view, the estimated pose may look fine (column two vs. four). However, when rotated, it is clear that training TUCH with self-contact information improves 3D pose estimation (column three vs. five).

Although self-contact is ubiquitous in human behavior, it is rarely explicitly studied in computer vision. For our purposes, self-contact comprises "self touch" (where the hands touch the body) and contact between other body parts (e.g. crossed legs). We ignore body parts that are frequently in contact (e.g. at the crotch or armpits) and focus on contact that is communicative or functional. Our goal is to estimate 3D human pose and shape (HPS) accurately for any pose. When self-contact is present, the estimated pose should reflect the true 3D contact.

Unfortunately, existing methods that compute 3D bodies from images perform poorly on images with self-contact; see Fig. 1. Body parts that should be touching generally are not. Recovering human meshes from images typically involves either learning a regressor from pixels to 3D pose and shape [20, 23], or fitting a 3D model to image features using an optimization method [4, 34, 45, 46]. The learning approaches rely on labeled training data. Unfortunately, current 2D datasets typically contain labeled keypoints or segmentation masks but do not provide any information

about 3D contact. Similarly, existing 3D datasets typically avoid capturing scenarios with self-contact because it complicates mesh processing. What is missing is a dataset with in-the-wild images and reliable data about 3D self-contact.

To address this limitation, we introduce three new datasets that focus on self-contact at different levels of detail. Additionally, we introduce two new optimization-based methods that fit 3D bodies to images with contact information. We leverage these to estimate pseudo ground-truth 3D poses with self-contact. To make reasoning about contact between body parts, the hands, and the face possible, we represent pose and shape with the SMPL-X [34] body model, which realistically captures the body surface details, including the hands and face. Our new datasets then let us train neural networks to regress 3D HPS from images of people with self-contact more accurately than state-of-the-art methods.

To begin, we first construct a *3D Contact Pose (3DCP)* dataset of 3D meshes where body parts are in contact. We do so using two methods. First, we use high-quality 3D scans of subjects performing self-contact poses. We extend previous mesh registration methods to cope with self-contact and register the SMPL-X mesh to the scans. To gain more variety of poses, we search the AMASS dataset [28] for poses with self-contact or "near" self-contact. We then optimize these poses to bring nearby parts into full contact while resolving interpenetration. This provides a dataset of valid, realistic, self-contact poses in SMPL-X format.

Second, we use these poses to collect a novel dataset of images with near ground-truth 3D pose. To do so, we show rendered 3DCP meshes to workers on Amazon Mechanical Turk (AMT). Their task is to *Mimic The Pose (MTP)* as accurately as possible, including the contacts, and submit a photograph. We then use the "true" pose as a strong prior and optimize the pose in the image by extending SMPLify-X [34] to enforce contact. A key observation is that, if we know about self-contact (even approximately), this greatly reduces pose ambiguity by removing degrees of freedom. Thus, knowing contact makes the estimation of 3D human pose from 2D images more accurate. The resulting method, SMPLify-XMC (for SMPLify-X with Mimicked Contact), produces high-quality 3D reference poses and body shapes in correspondence with the images.

Third, to gain even more image variety, we take images from three public datasets [16, 17, 27] and have them labeled with discrete body-part contacts. This results in the *Discrete Self-Contact (DSC)* dataset. To enable this, we define a partitioning of the body into regions that can be in contact. Given labeled discrete contacts, we extend SMPLify to optimize body shape using image features and the discrete contact labels. We call this method SMPLify-DC, for SMPLify with Discrete Self-Contact.

Given the MTP and DSC datasets, we finetune a recent HPS regression network, SPIN [23]. When we have 3D reference poses, i.e. for MTP images, we use these as though they were ground truth and do not optimize them in SPIN. When discrete contact annotations are available, i.e. for DSC images, we use SMPLify-DC to optimize the fit in the SPIN training loop. Fine-tuning SPIN on MTP and DSC significantly improves accuracy of the regressed poses when there is contact (evaluated on 3DPW [43]). Surprisingly, the results on non-self-contact poses also improve, suggesting that (1) gathering accurate 3D poses for in-the-wild images is beneficial, and (2) that self-contact can provide valuable constraints that simplify pose estimation.

We call our regression method *TUCH* (Towards Understanding Contact in Humans). Figure 1 illustrates the effect of exploiting self-contact in 3D HPS estimation. By training with self-contact, TUCH significantly improves the physical plausibility.

In summary, the key contributions of this paper are: (1) We introduce TUCH, the first HPS regressor for self-contact poses, trained end-to-end. (2) We create a novel dataset of 3D human meshes with realistic contact (3DCP). (3) We define a "Mimic The Pose" MTP task and a new optimization method to create a novel dataset of in-the-wild images with accurate 3D reference data. (4) We create a large dataset of images with reference poses that use discrete contact labels. (5) We show in experiments that taking self-contact information into account improves pose estimation in two ways (data and losses), and in turn achieves state-of-the-art results on 3D pose estimation benchmarks. (6) The data and code are available for research purposes.

2. Related Work

3D pose estimation with contact. Despite rapid progress in 3D human pose estimation [19, 20, 23, 31, 34, 40, 45], and despite the role that self-contact plays in our daily lives, only a handful of previous works discuss self-contact. Information about contact can benefit 3D HPS estimation in many ways, usually by providing additional physical constraints to prevent undesirable solutions such as interpenetration between limbs.

Body contact. Lee and Chen [25] approximate the human body as a set of line segments and avoid collisions between the limbs and torso. Similar ideas are adopted in [3, 10] where line segments are replaced with cylinders. Yin et al. [48] build a pose prior to penalize deep interpenetration detected by the Open Dynamics Engine [41]. While efficient, these stickman-like representations are far from realistic. Using a full 3D body mesh representation, Pavlakos et al. [34] take advantage of physical limits and resolve interpenetration of body parts by adding an interpenetration loss. When estimating multiple people from an image, Zanfir et al. [49] use a volume occupancy exclusion loss to prevent penetration. Still, other work has exploited

textual and ordinal descriptions of body pose [35, 36]. This includes constraints like "Right hand above the hips". These methods, however, do not consider self-contact.

Most similar to us is the work of Fieraru et al. [8], which utilizes discrete contact annotations between people. They introduce contact signatures between people based on coarse body parts. This is similar to how we collect the DSC dataset. Contemporaneous with our work, Fieraru et al. [9] extend this to self-contact with a 2-stage approach. They train a network to predict "self-contact signatures", which are used for optimization-based 3D pose estimation. In contrast, TUCH is trained end-to-end to regress body pose with contact information.

World contact. Multiple methods use the 3D scene to help estimate the human pose. Physical constraints can come from the ground plane [44, 49], an object [13, 21, 22], or contextual scene information [11, 47]. Li et al. [26] use a DNN to detect 2D contact points between objects and selected body joints. Narasimhaswamy et al. [32] categorize hand contacts into self, person-person, and object contacts and aim to detect them from in-the-wild images. Their dataset does not provide reference 3D poses or shape.

All the above works make a similar observation: human pose estimation is not a stand-alone task; considering additional physical contact constraints improves the results. We go beyond prior work by addressing self-contact and showing how training with self-contact data improves pose estimation overall.

3D body datasets. While there are many datasets of 3D human scans, most of these have people standing in an "A" or "T" pose to explicitly minimize self-contact [38]. Even when the body is scanned in varied poses, these poses are designed to avoid self-contact [2, 6, 7, 37]. For example, the FAUST dataset has a few examples of self-contact and the authors identify these as the major cause of error for scan processing methods [5]. Recently, the AMASS [28] dataset unifies 15 different optical marker-based motion capture (mocap) datasets within a common 3D body parameterization, offering around 170k meshes with SMPL-H [39] topology. Since mocap markers are sparse and often do not cover the hands, such datasets typically do not explicitly capture self-contact. As illustrated in Table 1, none of these datasets explicitly addresses self-contact.

Pose mimicking. Our Mimic-The-Pose dataset uses the idea that people can replicate a pose that they are shown. Several previous works have explored this idea in different contexts. Taylor et al. [42] crowd-source images of people in the same pose by imitation. While they do not know the true 3D pose, they are able to train a network to match images of people in similar poses. Marinoiu et al. [29] motion capture subjects reenacting a 3D pose from a 2D image. They found that subjects replicated 3D poses with a mean joint error of around 100mm. This is on par with existing 3D pose regression methods, pointing to people's ability to approximately recreate viewed poses. Fieraru et al. [9] ask subjects to reproduce contact from an image in a lab setting. They manually annotate the contact, whereas our MTP task is done in people's homes and SMPLify-XMC is used to automatically optimize the pose and contact.

Name	Meshes	Meshes with self-contact
3DCP Scan (ours)	190	188
3D BodyTex [1]	400	3
SCAPE [2]	70	0
Hasler et al. [12]	520	0
FAUST [5]	100/ 400	20/ 140

Table 1. Existing 3D human mesh datasets with the number of poses and the number of contact poses identified by visual inspection. 3DCP Scan is the scan subset of 3DCP (see Section 4). FAUST (train/test) includes scans with self-contact, i.e. 20 in the training and 140 in the test set. However, in FAUST the variety is low as each subject is scanned in the same 10/20 poses, whereas in 3DCP Scan each subject does different poses.

3. Self-Contact

An intuitive definition of contact between two meshes, e.g. a human and an object, is based on intersecting triangles. Self-contact, however, must be formulated to exclude common, but not functional, triangle intersections, e.g. at the crotch or armpits. Intuitively, vertices are in self-contact if they are close in Euclidean distance (near zero) but distant in geodesic distance, i.e. far away on the body surface.

Definition 3.1. Given a mesh M with vertices M_V, we define two vertices v_i and $v_j \in M_V$ to be in *self-contact*, if (i) $\|v_i - v_j\| < t_{eucl}$, and (ii) $geo(v_i, v_j) > t_{geo}$, where t_{eucl} and t_{geo} are predefined thresholds and $geo(v_i, v_j)$ denotes the geodesic distance between v_i and v_j. We use shape-independent geodesic distances precomputed on the neutral, mean-shaped SMPL and SMPL-X models.

Following this definition, we denote the set of vertex pairs in self-contact as $M_C := \{(v_i, v_j) | v_i, v_j \in M_V$ and v_i, v_j satisfy Definition 3.1$\}$. M is a *self-contact mesh* when $|M_C| > 0$. We further define an operator $\mathcal{U}(\cdot)$ that returns a set of unique vertices in M_C, and an operator $f_g(\cdot)$ that takes v_i as input and returns the Euclidean distance to the nearest v_j that is far enough in the geodesic sense. Formally, $\mathcal{U}(M_C) = \{v_0, v_1, v_2, \ldots, v_n\}$, where $\forall v_i \in \mathcal{U}(M_C)$, $\exists v_j \in \mathcal{U}(M_C)$, such that $(v_i, v_j) \in M_C$. $f_g(v_i) := \min_{v_j \in M_G(v_i)} \|v_i - v_j\|$, where $M_G(v_i) := \{v_j | geo(v_i, v_j) > t_{geo}\}$.

We further cluster self-contact meshes into distinct types. To that end, we define self-contact signatures $\mathbf{S} \in \{0, 1\}^{K \times K}$; see [9] for a similar definition. We first segment the vertices of a mesh into K regions R_k, where

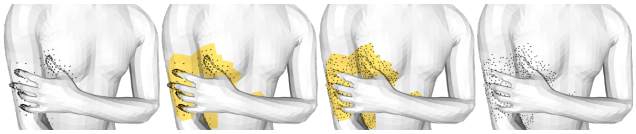

Figure 2. Visualization of the function $HD(X)$, that maps from mesh vertices to mesh surface points. First, a SMPL-X mesh with vertices in contact highlighted. Second, in yellow, all faces containing a vertex in contact are selected. Then, all points lying on a face containing a vertex in contact are selected from M_P, denoted as M_D. M_P is a fixed set of mesh surface points that are regressed from mesh vertices. Note that in image one and two the finger vertices are denser than the arm and chest vertices, in contrast to the more uniform density in images three and four.

$R_k \cap R_l = \emptyset$ for $k \neq l$ and $\bigcup_{k=1}^{K} R_k = M_V$. We use fine signatures to cluster self-contact meshes from AMASS (see Sup. Mat.) and rough signatures (see Fig. 6) for human annotation.

Definition 3.2. Two regions R_k and R_l are in contact if $\exists(v_i, v_j) \in M_C$, such that $v_i \in R_k$ and $v_j \in R_l$ holds. If R_k and R_l are in contact, $\mathbf{S}_{kl} = \mathbf{S}_{lk} = 1$. $M_\mathbf{S}$ denotes the contact signature for mesh M.

To detect self-contact, we need to be able to quickly compute the distance between two points on the body surface. Vertex-to-vertex distance is a poor approximation of this due to the varying density of vertices across the body. Consequently, we introduce HD SMPL-X and HD SMPL to efficiently approximate surface-to-surface distance. For this, we uniformly, and densely, sample points, $M_P \in \mathbb{R}^{P\times 3}$ with $P = 20,000$ on the body. A sparse linear regressor $\mathcal{P}$ regresses M_P from the mesh vertices M_V, $M_P = \mathcal{P}M_V$. The geodesic distance $geo_{HD}(p_1, p_2)$ between $p_1 \in M_P$ and $p_2 \in M_P$ is approximated via $geo(m, n)$, where $m = \arg\min_{v\in M_V} \|v - p_1\|$ and $n = \arg\min_{v\in M_V} \|v - p_2\|$. In practice, we use mesh surface points only when contact is present by following a three-step procedure as illustrated in Fig. 2. First, we use Definition 3.1 to detect vertices in contact, M_C. Then we select all points in M_P lying on faces that contain vertices in M_C, denoted as M_D. Last, for $p_i \in M_D$ we find the closest mesh surface point $\min_{p_j\in M_D} \|p_i - p_j\|$, such that $geo_{HD}(p_i, p_j) > t_{geo}$. With $HD(X) : X \subset M_V \rightarrow M_D \subset M_P$ we denote the function that maps from a set of mesh vertices to a set of mesh surface points. As the number of points, P, increases, the point-to-point distance approximates the surface-to-surface distance.

4. Self-Contact Datasets

Our goal is to create datasets of in-the-wild images paired with 3D human meshes as pseudo ground truth. Unlike traditional pipelines that collect images first and then annotate them with pose and shape parameters [18, 43], we take the opposite approach. We first curate meshes with self-contact and then pair them with images through a novel pose mimicking and fitting procedure. We use SMPL-X to create the 3DCP and MTP dataset to better fit contacts between hands and bodies. However, to fine-tune SPIN [23], we convert MTP data to SMPL topology, and use SMPLify-DC when optimizing with discrete contact.

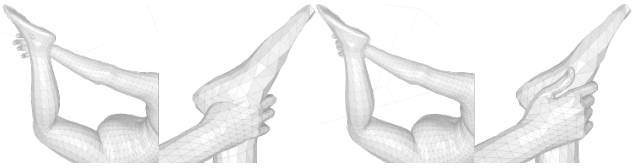

Figure 3. Self-contact optimization. Column 1 and 2: a pose selected from AMASS with near self-contact (between the fingertips and the foot) and interpenetration (thumb and foot). Column 3 and 4: after self-contact optimization, all fingers are in contact with the foot and interpenetration is reduced.

4.1. 3D Contact Pose (3DCP) Meshes

We create 3D human meshes with self-contact in two ways: with 3D scans and with motion capture data.

3DCP Scan. We scan 6 subjects (3 males, 3 females) in self-contact poses. We then register the SMPL-X mesh topology to the raw scans. These registrations are obtained using Co-Registration [14], which iteratively deforms the SMPL-X template mesh V to minimize the *point-to-plane* distance between the scan points $S \in \mathbb{R}^{N\times 3}$, where N is the number of scan points and the template points $V \in \mathbb{R}^{10375\times 3}$. However, registering poses with self-contact is challenging. When body parts are in close proximity, the standard process can result in interpenetration. To address this, we add a self-contact-preserving energy term to the objective function. If two vertices v_i and v_j are in contact according to Definition 3.1, we minimize the *point-to-plane* distance between triangles including v_i and the triangular planes including v_j. This term ensures that body parts that are in contact remain in contact; see Sup. Mat. for details.

3DCP Mocap. While mocap datasets are usually not explicitly designed to capture self-contact, it does occur during motion capture. We therefore search the AMASS dataset for poses that satisfy our self-contact definition. We find that some of the selected meshes from AMASS contain small amounts of self-penetration or near contact. Thus, we perform *self-contact optimization* to fix this while encouraging contact, as shown in Fig. 3; see Sup. Mat. for details.

4.2. Mimic-The-Pose (MTP) Data

To collect in-the-wild images with near ground-truth 3D human meshes, we propose a novel two-step process (see Fig. 4). First, using meshes from 3DCP as examples, workers on AMT are asked to mimic the pose as accurately as possible while someone takes their photo showing the full

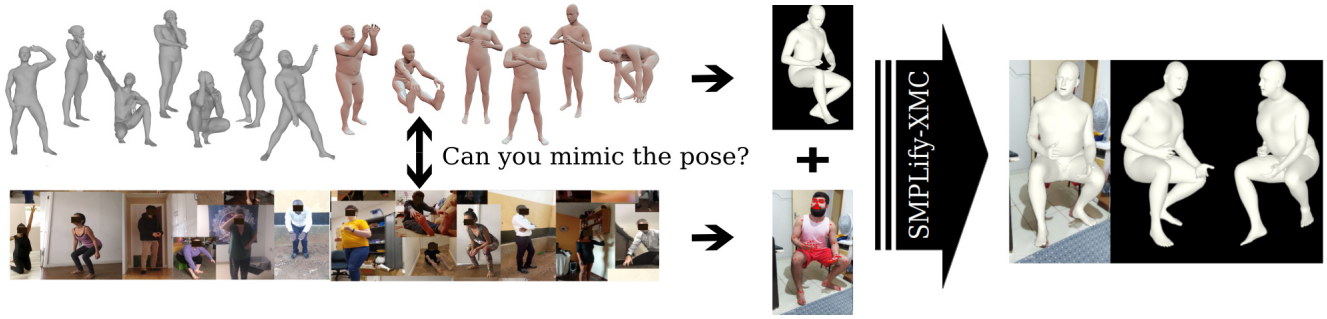

Figure 4. Mimic-The-Pose (MTP) dataset. MTP is built via: (1) collecting many 3D meshes that exhibit self-contact. In grey, new 3D scans in self-contact poses, in brown self-contact poses optimized from AMASS mocap data. (2) collecting images in the wild, by asking workers on AMT to mimic poses and contacts. (3) the presented meshes are refined via SMPLify-XMC to match the image features.

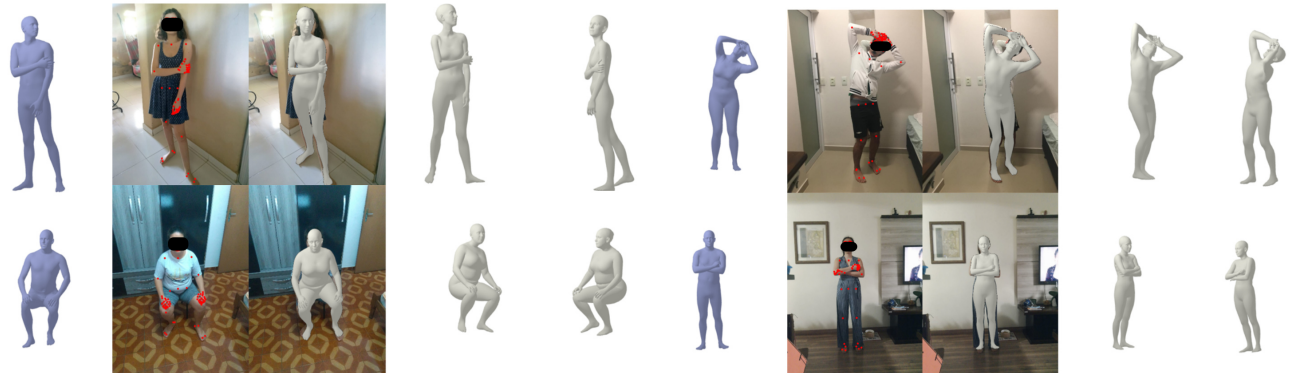

Figure 5. MTP results. Meshes presented to AMT workers (blue) and the images they submitted with OpenPose keypoints overlaid. In grey, the pseudo ground-truth meshes computed by SMPLify-XMC.

body (the *mimicked pose*). Mimicking poses may be challenging for people when only a single image of the pose is presented [29]. Thus, we render each 3DCP mesh from three different views with the contact regions highlighted (the *presented pose*). We allot 3 hours time for ten poses. Participants also provide their height and weight. All participants gave informed consent for the capture and the use of their imagery. Please see Sup. Mat. for details.

SMPLify-XMC. The second step applies a novel optimization method to estimate the pose in the image, given a strong prior from the presented pose. The presented pose $\tilde{\theta}$, shape $\tilde{\beta}$, and gender is not mimicked perfectly. To obtain pseudo ground-truth pose and shape, we adapt SMPLify-X [34], a multi-stage optimization method, that fits SMPL-X pose θ, shape β, and expression ψ to image features starting from the mean pose and shape. We make use of the presented pose $\tilde{\theta}$ in three ways: first, to initialize the optimization and solve for global orientation and camera; second, it serves as a pose prior; and third its contact is used to keep relevant body parts close to each other. We refer to this new optimization method as SMPLify-XMC.

In the first stage, we optimize body shape β, camera Π (rotation, translations, and focal length), and body global orientation θ_g, while the pose θ is initialized as $\tilde{\theta}$ and stays constant; see Sup. Mat. for a description of the first stage.

In the second and third stage, we jointly optimize θ, β, and Π to minimize

$$\begin{aligned}\mathcal{L}(\theta,\beta,\Pi) =& E_J + \lambda_{m_h} E_{m_h} + \lambda_{\tilde{\theta}}\mathcal{L}_{\tilde{\theta}} + \\ & \lambda_M \mathcal{L}_M + \lambda_{\tilde{C}}\mathcal{L}_{\tilde{C}} + \lambda_S \mathcal{L}_S.\end{aligned} \tag{1}$$

E_J denotes the same re-projection loss as specified in [34][1]. We use the standard SMPLify-X priors for left and right hand E_{m_h}. While the pose prior in [34] penalizes deviation from the mean pose, here, $\mathcal{L}_{\tilde{\theta}}$ is an L2-Loss that penalizes deviation from the presented pose. The measurements loss $\mathcal{L}_M$ takes ground-truth height and weight into account; see Sup. Mat. for details. The term $\mathcal{L}_{\tilde{C}}$ acts on $\tilde{M}_C$, the vertices in self-contact on the presented mesh. To ensure the desired self-contact, one could seek to minimize the distances between vertices in contact, e.g. $||v_i - v_j||$, $(v_i, v_j) \in \tilde{M}_C$. However, with this approach, we observe slight mesh distortions, when presented and mimicked contact are different. Instead, we use a term that encourages every vertex in $\tilde{M}_C$ to be in contact. More formally,

$$\mathcal{L}_{\tilde{C}} = \frac{1}{|\mathcal{U}(\tilde{M}_C)|} \sum_{v_i \in \mathcal{U}(\tilde{M}_C)} \tanh(f_g(v_i)). \tag{2}$$

[1] We denote loss terms defined in prior work as E while ours as $\mathcal{L}$.

The third stage actives $\mathcal{L}_S$ for fine-grained self-contact optimization, which resolves interpenetration while encouraging contact. The objective is $\mathcal{L}_S = \lambda_C\mathcal{L}_C + \lambda_P\mathcal{L}_P + \lambda_A\mathcal{L}_A$. Vertices in contact are pulled together via a contact term $\mathcal{L}_C$; vertices inside the mesh are pushed to the surface via a pushing term $\mathcal{L}_P$, and $\mathcal{L}_A$ aligns the surface normals of two vertices in contact.

To compute these terms, we must first find which vertices are inside, $M_I \subset M_V$, or in contact, $M_C \subset M_V$. M_C is computed following Definition 3.1 with $t_{geo} = 30$cm and $t_{eucl} = 2$cm. The set of inside vertices M_I is detected by generalized winding numbers [15]. SMPL-X is not a closed mesh and thus complicating the test for penetration. Consequently, we close it by adding a vertex at the back of the mouth. In addition, neighboring parts of SMPL and SMPL-X often intersect, e.g. torso and upper arms. We identify such common self-intersections and filter them out from M_I. See Sup. Mat. for details. To capture fine-grained contact, we map the union of inside and contact vertices onto the HD SMPL-X surface, i.e. $M_D = HD(M_I \cup M_C)$, which is further segmented into an inside M_{D_I} and outside $M_{D_I^\complement}$ subsets by testing for intersections. The self-contact objectives are defined as

$$\mathcal{L}_C = \sum_{p_i \in M_{D_I^\complement}} \alpha_1 \tanh\left(\frac{f_g(p_i)}{\alpha_2}\right)^2,$$

$$\mathcal{L}_P = \sum_{p_i \in M_{D_I}} \beta_1 \tanh\left(\frac{f_g(p_i)}{\beta_2}\right)^2,$$

$$\mathcal{L}_A = \sum_{(p_i,p_j) \in M_{D_C}} 1 + \langle N(p_i), N(p_j) \rangle.$$

f_g denotes the function that finds the closest point $p_j \in M_D$. M_{D_C} is the subset of vertices in contact in M_D. We use $\alpha_1 = \alpha_2 = 0.005$, $\beta_1 = 1.0$, and $\beta_2 = 0.04$ and visualize the contact and pushing functions in the Sup. Mat. Fig. 5 shows examples of our pseudo ground-truth meshes.

4.3. Discrete Self-Contact (DSC) Data

Images in the wild collected for human pose estimation normally come with 2D keypoint annotations, body segmentation, or bounding boxes. Such annotations lack 3D information. Discrete self-contact annotation, however, provides useful 3D information about pose. We use $K = 24$ regions and label their pairwise contact for three publicly available datasets, namely Leeds Sports Pose (LSP), Leeds Sports Pose Extended (LSPet), and DeepFashion (DF). An example annotation is visualized in Fig. 6. Of course, such labels are noisy because it can be difficult to accurately determine contact from an image. See Sup. Mat. for details.

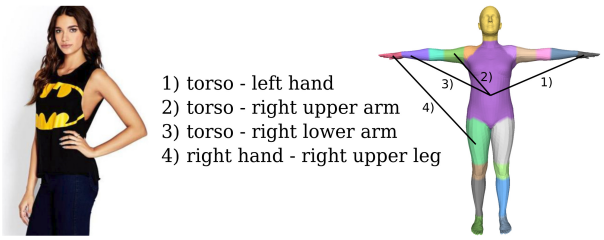

Figure 6. DSC dataset. Image with discrete contact annotation on the left. Right: DSC signature with $K = 24$ regions.

4.4. Summary of the Collected Data

Our 3DCP human mesh dataset consists of 190 meshes containing self-contact from 6 subjects, 159 SMPL-X bodies fit to commercial scans from AGORA [33], and 1304 self-contact optimized meshes from mocap data. From these 1653 poses, we collect 3731 mimicked pose images from 148 unique subjects (52 female; 96 male) for MTP and fit pseudo ground-truth SMPL-X parameters. MTP is diverse in body shapes and ethnicities. Our DSC dataset provides annotations for 30K images.

5. TUCH

Finally, we train a regression network that has the same design as SPIN [23]. At each training iteration, the current regressor estimates the pose, shape, and camera parameters of the SMPL model for an input image. Using ground-truth 2D keypoints, an optimizer refines the estimated pose and shape, which are used, in turn, to supervise the regressor. We follow this regression-optimization scheme for DSC data, where we have no 3D ground truth. To this end, we adapt the in-the-loop SMPLify routine to account for discrete self-contact labels, which we term SMPLify-DC. For MTP images, we use the pseudo ground truth from SMPLify-XMC as direct supervision with no optimization involved. We explain the losses of each routine below.

Regressor. Similar to SPIN, the regressor of TUCH predicts pose, shape, and camera, with the loss function:

$$L_R = E_J + \lambda_\theta E_\theta + \lambda_\beta E_\beta + \lambda_C\mathcal{L}_C + \lambda_P\mathcal{L}_P. \quad (3)$$

E_J denotes the joint re-projection loss. $\mathcal{L}_P$ and $\mathcal{L}_C$ are self-contact loss terms used in $\mathcal{L}_S$ in SMPLify-XMC, where $\mathcal{L}_P$ penalizes mesh intersections and $\mathcal{L}_C$ encourages contact. Further, E_θ and E_β are L2-Losses that penalize deviation from the pseudo ground-truth pose and shape.

Optimizer. We develop SMPLify-DC to fit pose θ_{opt}, shape β_{opt}, and camera Π_{opt} to DSC data, taking ground-truth keypoints and contact as constraints. Typically, in human mesh optimization methods the camera is fit first, then the model parameters follow. However, we find that this can distort body shape when encouraging contact. Therefore, we optimize shape and camera translation first, using

Figure 7. Initial wrong contact (left) from the regressor is fixed by SMPLify-DC after 5 (middle) and 10 (right) iterations.

the same camera fitting loss as in [23]. After that, body pose and global orientation are optimized under the objective

$$L_O(\theta) = E_J + \lambda_\theta E_\theta + \lambda_C \mathcal{L}_C + \lambda_P \mathcal{L}_P + \lambda_D \mathcal{L}_\mathcal{D}. \quad (4)$$

The discrete contact loss, $\mathcal{L}_\mathcal{D}$, penalizes the minimum distance between regions in contact. Formally, given a contact signature $\mathbf{S}$ where $\mathbf{S}_{ij} = \mathbf{S}_{ji} = 1$ if two regions R_i and R_j are annotated to be in contact, we define

$$\mathcal{L}_\mathcal{D} = \sum_{i=1}^{K} \sum_{j=i+1}^{K} \mathbf{S}_{ij} \min_{v \in R_i, u \in R_j} ||v - u||^2.$$

Given the optimized pose θ_{opt}, shape β_{opt}, and camera Π_{opt}, we compute the re-projection error and the minimum distance between the regions in contact. When the re-projection error improves, and more regions with contact annotations are closer than before, we keep the optimized pose as the current best fit. When no ground truth is available, the current best fits are used to train the regressor.

We make three observations: (1) The optimizer is often able to fix incorrect poses estimated by the regressor because it considers the ground-truth keypoints and contact (see Fig. 7). (2) Discrete contact labels bring overall improvement by helping resolve depth ambiguity (see Fig. 8). (3) Since we have mixed data in each mini-batch, the direct supervision of MTP data improves the regressor, which benefits SMPLify-DC by providing better initial estimates.

Implementation details. We initialize our regression network with SPIN weights [23]. For SMPLify-DC, we run 10 iterations per stage and do not use the HD operator to speed up the optimization process. For the 2D re-projection loss, we use ground-truth keypoints when available and, for MTP and DF images, OpenPose detections weighted by confidence. From DSC data we only use images where the full body is visible and ignore annotated region pairs that are connected in the DSC segmentation (see Sup. Mat.).

6. Evaluation

We evaluate TUCH on the following three datasets: **3DPW** [43], **MPI-INF-3DHP** [30], and **3DCP Scan**. This latter dataset consists of RGB images taken during the 3DCP Scan scanning process. While TUCH has never seen these images or subjects, the contact poses were mimicked in creation of MTP, which is used in training.

We use standard evaluation metrics for 3D pose, namely Mean Per-Joint Position Error (MPJPE) and the Procrustes-aligned version (PA-MPJPE), and Mean Vertex-to-Vertex Error (MV2VE) for shape and contact. Tables 2 and 3 summarize the results of TUCH on 3DPW and 3DCP Scan. Interestingly, TUCH is more accurate than SPIN on 3DPW. See Sup. Mat. for results of fine-tuning EFT.

	MPJPE		PA-MPJPE	
	3DPW	MI	3DPW	MI
SPIN [23]	96.9	105.2	59.2	**67.5**
EFT [18]	-	-	**54.2**	68.0
TUCH	**84.9**	**101.2**	55.5	68.6

Table 2. Evaluation on 3DPW and MPI-INF-3DHP (MI). Bold numbers indicate the best result; units are *mm*. We report the EFT result denoted in their publication when 3DPW was not part of the training data. Please note that SPIN is trained on MI, but we do not include MI in the fine-tuning set. MI contains mostly indoor lab sequences (100% train, 75% test), while DSC and MTP contain only in-the-wild images. This domain gap likely explains the decreased performance in PA-MPJPE.

	MPJPE	PA-MPJPE	MV2VE
SPIN [23]	79.7	50.6	95.7
EFT [18]	71.4	48.3	83.9
TUCH	**69.5**	**42.5**	**81.5**

Table 3. Evaluation on 3DCP Scan. Numbers are in *mm*. Note that in contrast to TUCH, this version of SPIN did not see poses in the MTP dataset during training. Please see Table 5 and the corresponding text for an ablation study.

	MPJPE				PA-MPJPE			
	contact	no contact	unclear	total	contact	no contact	unclear	total
SPIN	100.2	95.5	96.7	96.9	59.1	61.7	55.7	59.2
TUCH	**85.1**	**86.6**	**81.9**	**84.9**	**54.1**	**58.6**	**51.2**	**55.5**

Table 4. Evaluation of TUCH for contact classes in 3DPW. Numbers are in *mm*. See text.

We further evaluate our results w.r.t. contact. To this end, we divide the 3DPW test set into subsets, namely for $t_{geo} = 50$cm: *self-contact* ($t_{eucl} < 1$cm), *no self-contact* ($t_{eucl} > 5$cm), and *unclear* ($1\text{cm} < t_{eucl} < 5\text{cm}$). For 3DPW we obtain 8752 *self-contact*, 16752 *no self-contact*, and 9491 *unclear* poses. Table 4 shows a clear improvement on poses with contact and unclear poses compared to a smaller improvement on poses without contact.

To further understand the improvement of TUCH over SPIN, we break down the improved MPJPE in 3DPW *self-contact* into the pairwise body-part contact labels defined in the DSC dataset. Specifically, for each contact pair, we search all poses in 3DPW *self-contact* that have this particular self-contact. We find a clear improvement for a large number of contacts between two body parts, frequently between arms and torso, or e.g. left hand and right elbow, which is common in arms-crossed poses (see Fig. 9).

TUCH incorporates self-contact in various ways: annotations of training data, in-the-loop fitting, and in the re-

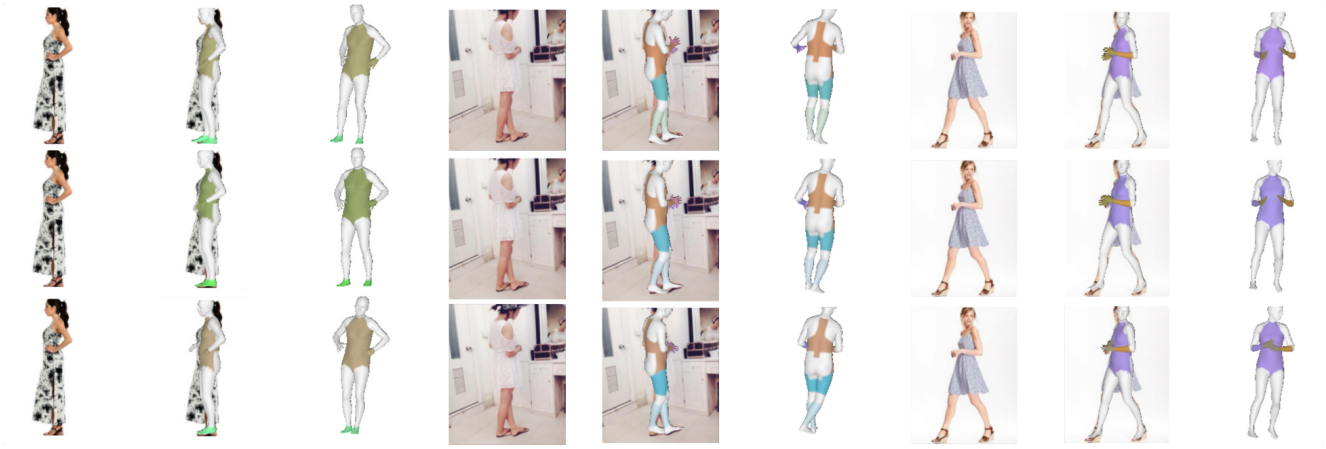

Figure 8. Impact of discrete self-contact labels in human pose estimation. Body parts labeled in contact are shown in the same color. First row shows an initial SPIN estimate, second row the SMPLify fit, third row the SMPLify-DC fit after 20 iterations.

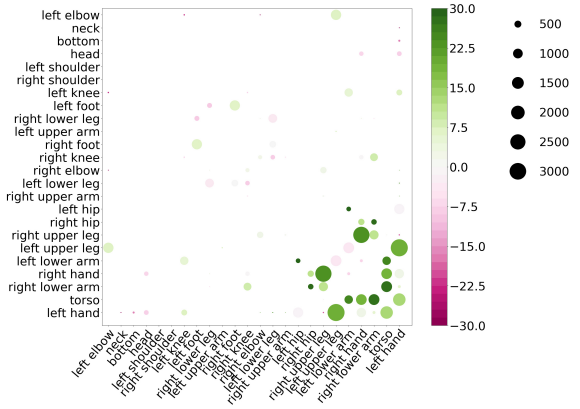

Figure 9. Average MPJPE difference (SPIN - TUCH), evaluated on the *self-contact* subset of 3DPW. The axes show labels for the DSC regions. Green indicates that TUCH has a lower error than SPIN on average across all poses with the corresponding regions in contact. The circle size represents the number of images per region. Regions with small circle sizes are less common.

gression loss. We evaluate the impact of each in Table 5. S+ is SPIN but it sees MTP+DSC images in fine-tuning and runs standard in-the-loop SMPLify with no contact information. S++ is S+ but uses pseudo ground truth computed with SMPLify-XMC on MTP images; thus self-contact is used to generate the data but nowhere else. S+ vs. SPIN suggests that, while poses in 3DCP Scan appear in MTP, just seeing similar poses for training and testing does not yield improvement. S+ vs. TUCH is a fair comparison as both see the same images during training. The improved results of TUCH confirm the benefit of using self-contact.

7. Conclusion

In this work, we address the problem of HPS estimation when self-contact is present. Self-contact is a natural, common occurrence in everyday life, but SOTA methods fail to estimate it. One reason for this is that no datasets pairing images in the wild and 3D reference poses exist. To address this problem we introduce a new way of collecting data: we ask humans to mimic presented 3D poses. Then we use our new SMPLify-XMC method to fit pseudo ground-truth 3D meshes to the mimicked images, using the presented pose and self-contact to constrain the optimization. We use the new MTP data along with discrete self-contact annotations to train TUCH; the first end-to-end HPS regressor that also handles poses with self-contact. TUCH uses MTP data as if it was ground truth, while the discrete, DSC, data is exploited during SPIN training via SMPLify-DC. Overall, incorporating contact improves accuracy on standard benchmarks like 3DPW, remarkably, not only for poses with self-contact, but also for poses without self-contact.

	SPIN	S+	S++	TUCH
3DPW	96.9/ 59.2	96.1/ 61.4	85.0/ 56.3	**84.9/ 55.5**
3DCP Scan	82.2/ 52.1	86.9/ 52.3	**74.8**/ 45.7	75.2/ **45.4**
MI	105.2/ 67.5	105.8/ 69.4	103.1/ 69.0	**101.2/ 68.6**

Table 5. MPJPE/PA-MPJPE (mm) to examine the impact of data and algorithm on 3DPW, 3DCP Scan, and MPI-INF-3DHP (MI).

Acknowledgments: We thank Tsvetelina Alexiadis and Galina Henz for their help with data collection and Vassilis Choutas for the SMPL-X body measurements and his implementation of Generalized Winding Numbers. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Lea Müller and Ahmed A. A. Osman.

Disclosure: MJB has received research gift funds from Adobe, Intel, Nvidia, Facebook, and Amazon. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, Max Planck. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH.

References

[1] Eman Ahmed, Alexandre Saint, Abd El Rahman Shabayek, Kseniya Cherenkova, Rig Das, Gleb Gusev, Djamila Aouada, and Bjorn Ottersten. A survey on deep learning advances on different 3d data representations. *arXiv preprint arXiv:1808.01462*, 2018. 3

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape Completion and Animation of PEople. *Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 3

[3] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3D pictorial structures revisited: Multiple human pose estimation. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 38(10):1929–1942, 2016. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, volume 9909, pages 561–578, 2016. 1

[5] Federica Bogo, Javier Romero, Matthew Loper, and Michael J. Black. FAUST: Dataset and evaluation for 3D mesh registration. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3794 –3801, 2014. 3

[6] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Dynamic FAUST: Registering human bodies in motion. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5573–5582, 2017. 3

[7] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008. 3

[8] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-Dimensional reconstruction of human interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7212–7221, 2020. 3

[9] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3D human self-contact. In *AAAI Conference on Artificial Intelligence (AAAI)*, Feb. 2021. 3

[10] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real-time human pose tracking from range data. In *European Conference on Computer Vision (ECCV)*, volume 7577, pages 738–751, 2012. 2

[11] Abhinav Gupta, Trista Chen, Francine Chen, Don Kimber, and Larry S Davis. Context and observation driven latent variable model for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 3

[12] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum (CGF)*, volume 28, pages 337–346, 2009. 3

[13] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 3

[14] David A Hirshberg, Matthew Loper, Eric Rachlin, and Michael J Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *European Conference on Computer Vision (ECCV)*, volume 7577, pages 242–255, 2012. 4

[15] Alec Jacobson, Ladislav Kavan, and Olga Sorkine-Hornung. Robust inside-outside segmentation using generalized winding numbers. *Transactions on Graphics (TOG)*, 32(4):1–12, 2013. 6

[16] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *British Machine Vision Conference (BMVC)*, pages 1–11, 2010. 2

[17] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1465–1472, 2011. 2

[18] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. *arXiv:2004.03686*, 2020. 4, 7

[19] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 2

[20] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 1, 2

[21] Vladimir G Kim, Siddhartha Chaudhuri, Leonidas Guibas, and Thomas Funkhouser. Shape2Pose: Human-centric shape analysis. *Transactions on Graphics (TOG)*, 33(4):120, 2014. 3

[22] Hedvig Kjellström, Danica Kragić, and Michael J Black. Tracking people interacting with objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 747–754, 2010. 3

[23] 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)*, pages 2252–2261, 2019. 1, 2, 4, 6, 7

[24] Yen Lee Angela Kwok, Jan Gralton, and Mary-Louise McLaws. Face touching: A frequent habit that has implications for hand hygiene. *Am J Infect Control*, 43(2):112–114, Feb. 2015. 1

[25] Hsi-Jian Lee and Zen Chen. Determination of 3D human body postures from a single view. *Computer Vision, Graphics, and Image Processing (CGIP)*, 30(2):148–168, 1985. 2

[26] Zongmian Li, Jiri Sedlar, Justin Carpentier, Ivan Laptev, Nicolas Mansard, and Josef Sivic. Estimating 3D motion and forces of person-object interactions from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8640–8649, 2019. 3

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016. 2

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, 2019. 2, 3

[29] Elisabeta Marinoiu, Dragos Papava, and Cristian Sminchisescu. Pictorial human spaces: How well do humans perceive a 3d articulated pose? In *International Conference on Computer Vision (ICCV)*, pages 1289–1296, 2013. 3, 5

[30] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 7

[31] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *Transactions on Graphics (TOG)*, 39(4), 2020. 2

[32] Supreeth Narasimhaswamy, Trung Nguyen, and Minh Nguyen. Detecting hands and recognizing physical contact in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 3

[33] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[34] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 5

[35] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7307–7316, 2018. 3

[36] Gerard Pons-Moll, David J. Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2345–2352, 2014. 3

[37] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. Dyna: A model of dynamic human shape in motion. *Transactions on Graphics (TOG)*, 34(4):120:1–120:14, 2015. 3

[38] Kathleen M. Robinette and Hein A. M. Daanen. The caesar project: a 3-d surface anthropometry survey. *Second International Conference on 3-D Digital Imaging and Modeling (Cat. No.PR00062)*, pages 380–386, 1999. 3

[39] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 3

[40] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, 152:1–20, 2016. 2

[41] Russell Smith et al. Open dynamics engine, 2005. 2

[42] Graham W Taylor, Ian Spiro, Christoph Bregler, and Rob Fergus. Learning invariance through imitation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2729–2736, 2011. 3

[43] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume 11214, pages 614–631, 2018. 2, 4, 7

[44] Marek Vondrak, Leonid Sigal, and Odest Chadwicke Jenkins. Dynamical simulation priors for human motion tracking. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(1):52–65, 2012. 3

[45] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966, 2019. 1, 2

[46] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *International Conference on 3D Vision (3DV)*, 2020. 1

[47] Masanobu Yamamoto and Katsutoshi Yagishita. Scene constraints-aided tracking of human body. In *Computer Vision and Pattern Recognition (CVPR)*, pages 151–156, 2000. 3

[48] Kangxue Yin, Hui Huang, Edmond SL Ho, Hao Wang, Taku Komura, Daniel Cohen-Or, and Hao Zhang. A sampling approach to generating closely interacting 3d pose-pairs from 2d annotations. *Transactions on Visualization and Computer Graphics (TVCG)*, 25(6):2217–2227, 2018. 2

[49] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 2, 3