

Real-time RGBD-based Extended Body Pose Estimation

Renat Bashirov^{1*} Anastasia Ianina^{1*} Karim Iskakov^{1,2} Yevgeniy Kononenko¹
Valeriya Strizhkova² Victor Lempitsky^{1,2} Alexander Vakhitov¹

¹ Samsung AI Center – Moscow, Russia ² Skolkovo Institute of Science and Technology, Russia

Abstract

We present a system for real-time RGBD-based estimation of 3D human pose. We use parametric 3D deformable human mesh model (SMPL-X) as a representation and focus on the real-time estimation of parameters for the body pose, hands pose and facial expression from Kinect Azure RGB-D camera. We train estimators of body pose and facial expression parameters. Both estimators use previously published landmark extractors as input and custom annotated datasets for supervision, while hand pose is estimated directly by a previously published method. We combine the predictions of those estimators into a temporally-smooth human pose. We train the facial expression extractor on a large talking face dataset, which we annotate with facial expression parameters. For the body pose we collect and annotate a dataset of 56 people captured from a rig of 5 Kinect Azure RGB-D cameras and use it together with a large motion capture AMASS dataset. Our RGB-D body pose model outperforms the state-of-the-art RGB-only methods and works on the same level of accuracy compared to a slower RGB-D optimization-based solution. The combined system runs at 25 FPS on a server with a single GPU. The code will be available at saic-violet.github.io/rgbd-kinect-pose

1. Introduction

A decade ago, realtime human pose estimation using RGBD Kinect sensor became a landmark achievement for computer vision [22, 38]. Over the subsequent decade, the focus in human pose estimation has however shifted onto RGB sensors [8, 9]. Also, while the original Kinect-based approach and many subsequent RGB-based works aimed at skeleton joints estimation, most recent trend is to estimate *extended* pose description that includes face expression and hands pose [19, 35].

Here, we argue that despite all the progress in RGB-

based pose estimation, the availability of depth can still be of great use for the pose estimation task. We therefore build an RGBD-based system for realtime pose estimation that uses a modern representation for extended pose (SMPL-X [35]) involving face and hands pose estimation. To build the system, we adopt a simple fusion approach, which uses pretrained realtime estimators for body, face, and hands poses. We then convert the outputs of these estimators into a coherent set of SMPL-X parameters.

To train our system, we collect a dataset of 56 people using a calibrated rig of five Kinect sensors. We then establish “ground truth” poses using slow per-frame optimization-based fitting process that accurately matches multi-view observations. We also fit the deformable head mesh to the videos from the large-scale VoxCeleb2 dataset [12]. The result of this fitting is then used as a ground truth for the learnable components of our system.

To recover the body pose, we train a neural network that converts the stream of depth-based skeleton estimates (as provided by the Kinect API [6]) into a stream of SMPL pose parameters (angles). We use a specific (residual) parameterization, which asks the network to predict corrections to the angles of the Kinect Skeleton. In the comparison, we observe that tracking accuracy of such depth-based system considerably exceeds the accuracy of the state-of-the-art RGB-based methods [11, 25, 24], validating the usefulness of the depth channel for human pose estimation. Furthermore, we compare the performance of our feed-forward network with the depth-based baseline that performs per-frame optimization of pose parameters. We observe that the feed-forward network achieves same accuracy and much higher speed.

In addition to the body parameters inferred from depth channel, we estimate the face and the hand parameters from the RGB color stream, since the effective resolution of the depth channel at the camera-to-body distances typical for fullbody tracking is not sufficient for these tasks. We use the outputs of the pretrained MinimalHand system [50] to perform SMPLX-compatible (MANO [37]) hand pose estimation. We also estimate the face keypoints using the Me-

*equal contribution

diaPipe face mesh model [21], and train a feedforward network that converts the keypoint coordinates into the SMPL-X compatible (FLAME [26]) face parameters. The MANO and FLAME estimates are fused with the body pose to complete the extended pose estimation.

Our resulting system thus uses the depth channel to estimate the body pose, and the color channel to estimate hands and face poses. It runs at 25 frames-per-second on a modern desktop with a single 2080TI GPU card, and provides reliable extended body tracking. We demonstrate that same as ten years back, RGBD-based pose estimation still strikes a favourable balance between the simplicity of the setup (single sensor is involved, no extrinsic calibration is needed), and the accuracy of pose tracking (in particular, higher accuracy is attainable compared to monocular RGB-based estimation).

2. Related work

Classic methods approach human pose estimation as sparse keypoint localization in color [36, 14, 13] or RGB-D [47, 38] images. In [43] the authors use dense per-pixel prediction of surface coordinates in RGB-D images. Deep network-based methods are significantly more accurate in such a direct regression, e.g. [44, 34, 40]. Some methods estimate the body joints in 3D from single [29] or multiple RGB views [18]. RGBD-based 3D joint prediction using deep learning is proposed in [30, 32, 16], and in this work we build on a recent commercial system [6] of this kind. In [49], Zhou et al. use known constant bone lengths and predict joint angles of a body kinematic model, for the first time proposing a deep architecture which estimates the body pose assuming the body shape is known in the form of bone lengths. Subsequent research [31, 41, 42, 5] focuses on incorporating structural knowledge into pose estimation.

Parametric body models [27, 35, 19, 46] are rich and higher level representations of body geometry separated into person-specific shape and pose. In this work we rely on a recent SMPL-X model [35] which integrates body, face and hand models into a unified framework, while alternatives are [46] or [19]. The most accurate methods for the estimation of the body model parameters were computationally intensive and offline [7, 35, 46, 45]. Recent research features feedforward deep architectures achieving real-time prediction of the body model parameters [20, 25, 11]. In particular, [20] proposes to estimate shape and pose parameters of the SMPL model from a single RGB image, while recent works, e.g. SPIN [25], show significant accuracy improvements in this direction. ExPose [11] is a real-time capable deep network predicting the whole set of SMPL-X parameters. VIBE [24] is a fast video-based SMPL shape and pose prediction method, while [46] shows a similar approach for the newer GHUM model. To the best of our knowledge RGB-D based parametric body model tracking

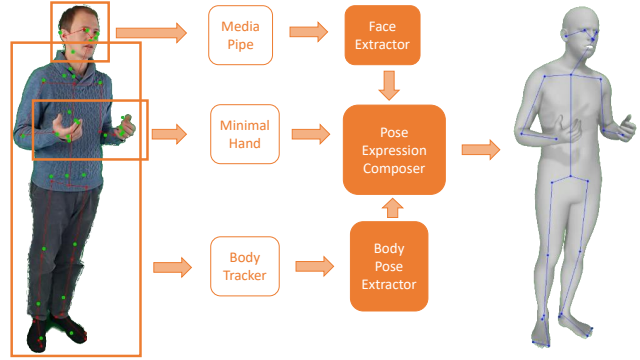


Figure 1. The real-time system predicts the SMPL-X [35] pose and expression from the RGB-D frame. We rely on the Kinect Body Tracker [6], the MediaPipe face mesh [21] and the hand pose regressor [50] (empty boxes). We propose the *body pose extractor* to estimate the body joint angles from the 3D landmarks, the *face extractor* to estimate the facial expression and jaw pose, and the *pose expression composer* to integrate the predictions into a single set of consistent SMPL-X parameters (filled boxes).

has not been tackled in previously published works.

For the body models, the recent research on body pose forecasting [4, 3] relies on a large motion capture dataset [28] with SMPL annotations. However, these approaches deal with prediction of noise-free poses rather than extraction of the high-level pose representation in the form of angles from the lower-level representation, as in this work. Another closely related field is skeletal retargeting, where researchers find ways of transforming the skeletons between different animated characters, trying to keep the pose semantics intact, e.g. [2, 1].

This work proposes a sequence-based method to estimate poses of the SMPL-X model, which assumes knowledge of the body shape as opposed to VIBE [24] and [48], and uses a sequence of RGB-D frames as input.

3. Methods

We use the SMPL-X (Skinned Multi-Person Linear - eXpressive) format for extended pose [35], which represents the shape of a human body using parametric deformable mesh model. The mesh vertex positions M are determined by a set of body shape β , pose Θ and facial expression ψ parameters as follows:

$$\begin{aligned}
 M(\beta, \Theta, \psi) &= W(T_p(\beta, \Theta, \psi), J(\beta), \Theta, \mathcal{W}) \\
 T_p(\beta, \Theta, \psi) &= \bar{T} + B_S(\beta; \mathcal{S}) + B_E(\psi; \mathcal{E}) + B_P(\Theta; \mathcal{P}).
 \end{aligned}
 \tag{1}$$

Here, $M(\beta, \Theta, \psi)$ is the posed human mesh, $T_p(\beta, \Theta, \psi)$ encodes the deformed mesh in a default body pose, W is a linear blend skinning function with vertex-joint assignment weights \mathcal{W} and joint coordinates $J(\beta)$; the mesh T_p is expressed as a sum of the vector of mean vertex coordinates \bar{T} summed with the blend-shape function B_S , the

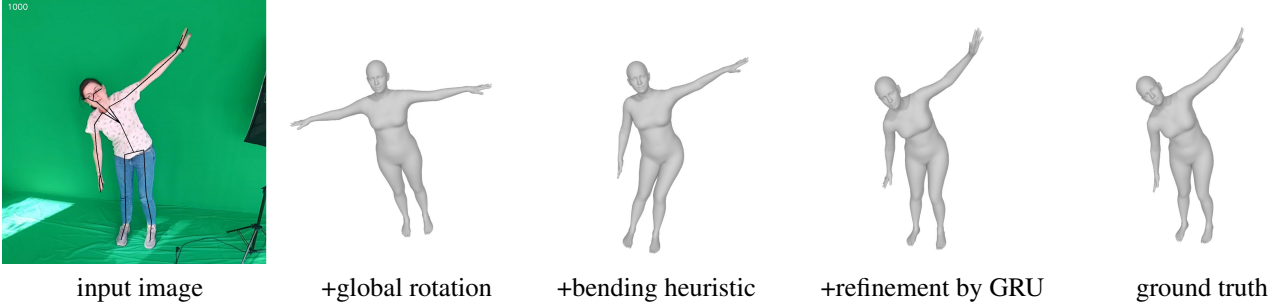


Figure 2. Our body pose extractor has 4 steps. Step 1: align the SMPL mesh with a global rotation. Step 2: apply learning-free bending heuristic to the limbs and the neck. Steps 3 and 4: refine the joint rotations with a GRU network, refine global rotation and translation.

displacements corresponding to the facial expression B_E and the pose B_P , where \mathcal{S} , \mathcal{E} , \mathcal{P} denote the mesh deformation bases due to shape, expression and pose, see [35] for the full definition of the model. The pose can be decomposed as $\Theta = [\theta^T, \xi^T, \zeta^T]^T$ into the parts corresponding to the hands ξ , the face ζ and the remaining body joints θ . The SMPL-X model is an “amalgamation” of the SMPL body model [27], the FLAME face/head model [26], and the MANO hand model [37] that had been proposed earlier.

3.1. System Overview

The proposed system regresses the sequence of the pose-expression pairs (Θ_i, ψ_i) from a sequence of RGB-D frames, see Fig. 1. As a first step of processing the frame i it extracts the vector of 3D body landmark predictions \mathbf{j}_i using the Kinect Body Tracker API [6]. For simplicity, we assume that they are defined in the RGB camera coordinates, in reality we use the intrinsics provided with the device for the coordinate system alignment. We predict the body pose sequence $\{\theta_i\}_{i=1}^N$, where N is the total number of frames, from a sequence of 3D landmarks $\{\{x_{i,k}\}_{k=1}^m\}_{i=1}^N$ using our *body pose extractor* module, where $x_{i,k}$ is the k -th 3D landmark for the i -th frame. We then crop the face and process it with MediaPipe face mesh predictor [21], and later extract the facial expression ψ_i and the pose of the jaw ζ_i using our *facial extractor* module. We crop the hands and predict the hand pose ξ_i in the SMPL-X (MANO) format using the MinimalHand [50] method. The components of the model are then combined together, and temporal smoothing is applied: Slerp Quaternion interpolation is used for body, jaw and hands rotations, and exponential temporal filter is used for face expression.

3.2. Body Pose Extractor

The aim of the body pose extractor is to predict the body pose θ given a known SMPL-X body shape vector β and the 3D body landmarks $\{x_k\}_{k=1}^m$ extracted by the Body Tracker [6] observed at a certain moment of time. The pose extractor should work for an arbitrary shape of a person, while in the public domain to the best of our knowledge there is no RGB-D dataset with sufficient variability

of human shapes. We achieve generalization to an arbitrary human shape by two means. Firstly, using the landmark-anchored vertices precomputed for the Body Tracker landmarks, we are able to leverage a large and diverse motion capture dataset [4] for learning our models. Secondly, we propose a specific residual rotation-based architecture designed to abstract from a particular human shape. Apart from the temporal connections inside the architecture, the whole pipeline estimates the pose independently for each frame. We therefore omit the temporal indices.

We assume that the Azure Kinect 3D landmark locations $\{x_k\}_{k=1}^m$ are given in the coordinate frame of the RGB camera. For each x_k we create an additional vertex in the SMPL-X mesh: a total of 32 vertices are added to 10475 SMPL-X vertices. For the SMPLX-X mesh aligned with the Azure Kinect skeleton, these additional vertices should coincide with the corresponding $\{x_k\}_{k=1}^m$. Each additional vertex is created to minimize the distance to the x_k over a training dataset as explained at the end of the Section 3.4. We call these additional vertices *landmark-anchored* vertices. They were created during AzurePose dataset annotation. As the body shape and the pose varies, landmark-anchored vertices follow the SMPL-X skinning equations (1).

For a landmark x_k , we denote the position of the corresponding landmark-anchored vertex in the SMPL-X pose θ as $\hat{x}_k(\theta)$. Our goal is essentially to find the body pose θ that aligns landmark-anchored vertices $\{\hat{x}_k(\theta)\}_{k=1}^m$ on the SMPL-X body mesh with the corresponding observed landmarks $\{x_k\}_{k=1}^m$. The body pose extractor takes the observed landmarks and outputs the body pose, while SMPL-X inference takes the body pose and outputs the landmark-anchored vertices, this way the body pose extractor “inverts” the inference of SMPL-X. For body pose extractor we use feed-forward computations without inverse kinematics-like optimizations.

To achieve this, we propose a four-step approach (Figure 2). Firstly, we find the global (rigid) rotation of the body mesh in the default SMPL pose. We define the vertical direction v between the ‘chest’ and ‘pelvis’ landmarks and the horizontal direction w between the left and right shoulder

	simple				complex				avg			
	AUC↑	Euler↓	joint-ang↓	Pos↓	AUC↑	Euler↓	joint-ang↓	Pos↓	AUC↑	Euler↓	joint-ang↓	Pos↓
single-frame												
SMPLify-X [35]	0.735	2.680	0.324	0.079	0.651	2.224	0.474	0.136	0.693	2.452	0.399	0.108
SPIN [25]	0.802	3.018	0.298	0.058	0.756	1.920	0.342	0.074	0.779	2.469	0.32	0.066
ExPose-X [11]	0.781	3.345	0.305	0.064	0.746	2.6	0.369	0.078	0.764	2.972	0.337	0.071
Ours MLP	0.853	2.878	0.240	0.040	0.777	1.552	0.296	0.068	0.815	2.215	0.268	0.054
multi-frame												
VIBE [24]	0.823	2.744	0.271	0.052	0.753	1.763	0.323	0.076	0.788	2.254	0.297	0.064
Ours RNN	0.870	2.988	0.233	0.038	0.791	1.846	0.287	0.063	0.83	2.417	0.26	0.05
Ours RNN-SPL	0.877	3.050	0.226	0.036	0.810	1.735	0.260	0.057	0.844	2.392	0.243	0.046

Table 1. Comparison with single- and multi-frame RGB baselines on AzurePose (simple/complex). For all metrics except AUC, lower is better. Our RGBD-based methods are more accurate in most metrics (except Euler angles measured for local rotations in simple sequences).

landmarks, while the corresponding directions in the mesh with the default pose are \hat{v} and \hat{w} , and all the direction vectors have a unit norm. We find the rotation matrix R , such that $v = R\hat{v}$, $v \times w = \frac{1}{\|\hat{v} \times \hat{w}\|} R(\hat{v} \times \hat{w})$ which aligns v and \hat{v} , and the plane defined by v and w with the one defined by \hat{v} , \hat{w} . Such a rotation is unique given that v and w are not collinear (which they are not by construction). It aligns the coordinate system of the root joint with the 3D landmark skeleton.

In the second step, we apply a learning-free bending heuristic to obtain an initial pose estimate θ^0 . As Azure Kinect skeleton topology roughly matches SMPL-X skeleton topology, and some of the landmark locations of Azure Kinect skeleton are close to the joints of SMPL-X skeleton, we therefore can set a subset of rotations of SMPL-X joints to match the rotations of Azure Kinect skeleton. To do that we compute two types of bone vectors for Azure Kinect skeleton: the observed bone vectors $b_k = x_k - x_{p(k)}$, where a function $p(k)$ returns the index of the parent of a landmark k in the kinematic tree, and the estimated bone vectors defined by the current estimate of the pose θ : $\hat{b}_k(\theta) = \hat{x}_k(\theta) - \hat{x}_{p(k)}(\theta)$. For each landmark-anchored vertex \hat{x}_k we define a *dominating* SMPL-X joint $s(k)$ as a joint with the maximal weight for this particular vertex in the linear SMPL deformable model. For two vectors a and b , denote as $R(a, b)$ the matrix of minimal rotation between a and b , which we define as a rotation with the axis $\frac{1}{\|a \times b\|} a \times b$ which aligns the vectors a and b , such that $\frac{1}{\|a\|} Ra = \frac{1}{\|b\|} b$. We then process a subset of bones \mathcal{D} corresponding to the limbs and the neck, each bone from $b_k \in \mathcal{D}$ having a unique $s(p(k))$, and for each bone $b_k \in \mathcal{D}$, we set the rotation matrix of $s(p(k))$ to $R(\hat{b}_k(\theta), b_k)$. We denote the resulting pose θ^0 . The positions of the vertices in SMPL mesh are affected by multiple joints, and not all of the bones are in \mathcal{D} . Thus the pose θ^0 does not result in collinearity of all the bone vectors $\{b_k\}$ and $\{\hat{b}_k(\theta^0)\}$, and the goal of the next step is to achieve a more precise alignment.

In the third step, we use machine learning to get a more

accurate alignment. We compute the *residual* minimal rotations $R_k = R(\hat{b}_k(\theta^0), b_k)$ aligning every estimated bone direction $\hat{b}_k(\theta^0)$ with the observed one b_k , in the coordinate frame of the joint $s(p(k))$. Concatenating $l = m - 1$ rotation matrices R_k , we obtain a $9 \times l$ -dimensional rotation residual vector. This vector serves as an input to a deep network, together with concatenated rotation matrices generated by the initialization heuristic from θ^0 , in Table 3 we compare this approach to other possible inputs. The goal of this network is thus to refine the estimates produced by the initialization bending heuristic.

We use Gated Recurrent Unit (GRU) architecture [10] for this refinement network. The network thus predicts incremental rotations \tilde{R}_k for each joint of the SMPL-X model, and the resulting θ^1 is obtained as a composition of these rotations with θ^0 : $R_k = R_k^0 \tilde{R}_k^1$, where R_k^1 is the predicted rotation from the k -th joint to its parent, R_k^0 is the corresponding rotation decoded from the initialization θ^0 . We encode the obtained predictions into the vector θ^1 .

We have tried two variants of the architecture. The first variant utilizes GRU architecture with two layers and hidden size equal to 1000. In the second variant we apply a structured prediction layer (SPL) [4] with hidden size 64 to GRU outputs. We use dense SPL which means that while making prediction for a joint we take into account all its ancestors in the kinematic tree instead of using only one parent joint. Also in this architecture we use dropout with rate 0.5. The networks are trained by minimizing the l^1 loss on concatenated rotation matrices using Adam optimizer [23] with learning rate 0.0001. We decided to use the recurrent architecture that uses temporal context from previous frames to predict the corrections. This allows to improve temporal stability (slightly). In the single-frame experiments below, we replace the GRU network with a simple multi-layer perceptron (MLP) with 5 layers (each layer consists of 512 neurons).

In the fourth and final step, we modify the global location and orientation of the resulting mesh through Pro-

	simple				complex				avg			
	AUC↑	Euler↓	joint-ang↓	Pos↓	AUC↑	Euler↓	joint-ang↓	Pos↓	AUC↑	Euler↓	joint-ang↓	Pos↓
SMPLify-RGBD-Online	0.897	2.9	0.227	0.03	0.841	1.698	0.277	0.047	0.869	2.299	0.252	0.039
SMPLify-RGBD	0.905	1.821	0.141	0.027	0.887	1.281	0.163	0.033	0.896	1.551	0.152	0.03
Ours-RNN	0.899	2.846	0.177	0.029	0.829	1.514	0.237	0.051	0.864	2.18	0.207	0.04
Ours-RNN-SPL	0.896	2.819	0.189	0.029	0.837	1.361	0.233	0.048	0.867	2.09	0.211	0.039

Table 2. Comparison with multi-frame RGB-D baselines on AzurePose (simple/complex), the slow offline optimization-based method is the most accurate. The proposed Ours-RNN-SPL method performs on a similar level of accuracy as the online optimization-based method, but is 2.5 times faster. The feedforward network inside our approach thus serves as an efficient approximation of the optimization process.

crustes analysis between the two sets of 3D points $\{x_k\}_{k=1}^m$ and $\{\hat{x}_k(\theta_i)\}_{k=1}^m$. This step further improves the alignment accuracy.

One may question whether our multi-stage alignment procedure can be replaced with a simpler one, or if the shape body parameters β can help at the fine-tuning prediction stage. In the experiments, we therefore provide an ablation study that compares our full procedure described in this sections with baselines.

3.3. Face and hands

For each input frame, we crop the RGB image regions corresponding to the face and the hands. For each hand we select a Body Tracker landmark corresponding to a hand, define a cube with a center in a selected landmark and a side of $0.3m$, project its vertices to image plane, and determine bounding box from projected points. To extract face region we project Body Tracker 3D face landmarks to the image, calculate minimal square bounding box and expand it by a factor of two.

On the face region, we run the MediaPipe face mesh predictor [21], which outputs 468 2.5D landmarks and rotated bounding box in real time. Absolute X, Y -coordinates of 2.5D landmarks are normalized on width and height of the rotated bounding box correspondingly, while relative Z -coordinate is scaled with a constant value of $1/256$. Normalized landmarks are used to predict jaw pose and facial expression of the SMPL-X head (FLAME) model. Landmarks are first passed through a seven-layer MLP with *Linear-BatchNorm-ReLU* [17, 33] blocks to extract 32-dimensional feature vector. Then this feature vector is fed into two linear layers, which predict jaw pose and facial expression separately. The model is trained on a large dataset of video sequences depicting talking humans, which we annotated with SMPL-X parameters as described in the next section. During training we use two losses: Mean Squared Error (MSE) of facial expressions and Mean Per-Joint Position Error (MPJPE) of SMPL-X 3D face keypoints. Additionally, we increase the weight of the mouth 3D keypoints in the MPJPE loss to make the predicted mouth more responsive (see section 4 for ablation study). All face predictors are trained with ADAM optimizer [23] with a learning rate 0.0005.

On the hands regions, we run a method [50], which outputs SMPL-X compatible set of pose parameters.

3.4. AzurePose dataset and SMPL-X Annotation

For training our models, we collect a dataset of 56 subjects recorded by five synchronous Kinect Azure devices. The maximal horizontal angular parallax between different devices is approximately 90 degrees, the distance to the person from the device is 2-3 meters. Subjects perform a fixed set of motions, with each recording being 5-6 minutes long. As a test set, we collect additionally two recordings of two subjects, with fixed sets of motions, so that the first recording contains basic motions ('simple'), while the other has more occlusions and extreme rotations for several joint ('complex'). The test recordings were taken in a separate session with modified camera geometry to ensure a realistic gap between the train and the test set.

To obtain ground truth SMPL-X poses, we use a slow optimization-based multiview fitting procedure. We thus optimize the SMPL-X parameters to fit the observations. Essentially, we are extending the SMPLify-X [35] method to process multiview synchronous RGB-D sequences. We use the OpenPose landmarks [8, 39] for body, face and hands, as well as 3D landmarks of the Body Tracker [6], and optimize the smooth l^1 regression cost [15]. We perform body pose estimation in the VPoser domain [35], use the l^2 costs for regularizing the estimation of the jaw and eye pose in the angular domain, the hand pose in the MANO [37] parameter domain, the body shape in the SMPL-X domain, face expression in the FLAME [26] domain, the body pose increment between the subsequent frames in the VPoser [35] domain.

We use the explained procedure to obtain SMPL-X annotations on the AzurePose dataset. Since the AzurePose dataset has limited identity and pose variability, we have also built a (semi)-synthetic test set based on AMASS dataset [28], which contains a large variety of tracks in the SMPL format captured with motion capture equipment. We turned the AMASS tracks into synthesized keypoint tracks by taking the coordinates of the landmark-anchored vertices and treating them as Kinect body tracker output. We call the obtained dataset 'AMASS-K'.

To enable feedforward prediction of facial pose and ex-

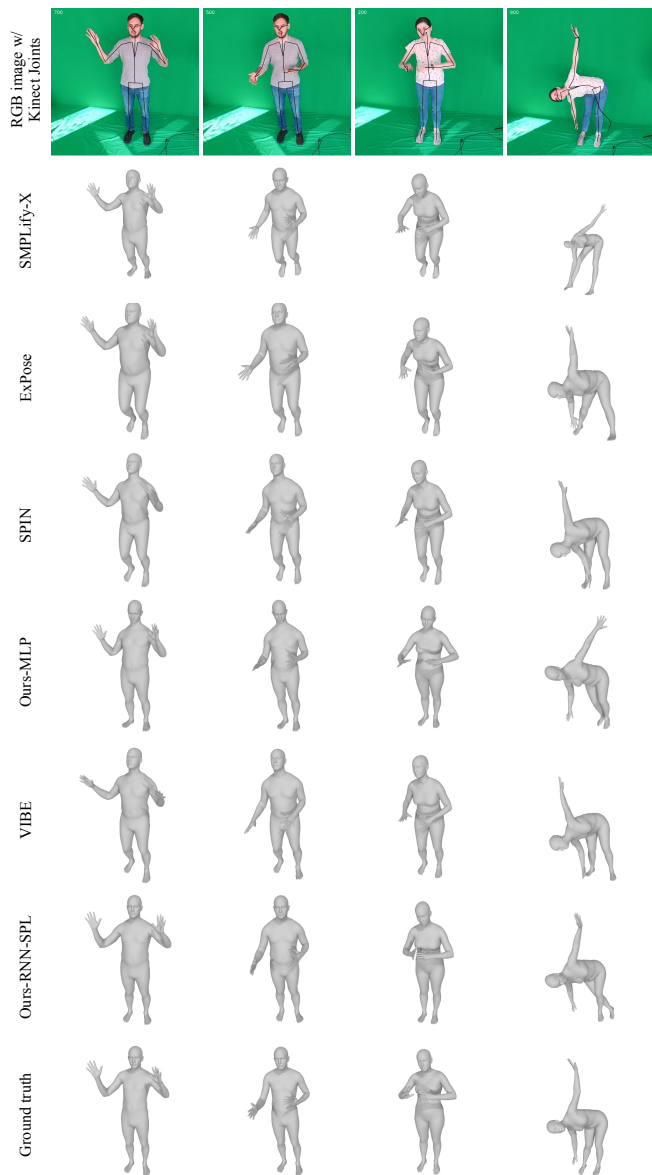


Figure 3. Comparison of the RGB baselines and the proposed methods on AzureTest frames. The use of depth in the proposed methods in most cases leads to more accurate poses. The right-most column shows a failure case where our method cannot recover from a mistake of the Kinect BodyTracker.

pression, we also require a large dataset of face videos annotated with facial parameters of SMPL-X, which are essentially the parameters of the FLAME model. We use a VoxCeleb2 dataset [12] for this purpose. First, we select sequences in which the speaker’s face has a high resolution (face bounding box more than 512×512 pixels). Then OpenPose [8] is used to get face landmarks, and the subset of sequences is filtered again according to landmark’s confidences. Next, we use an offline optimization-based sequence fitting procedure. Specifically, we optimize the shape (shared across sequence), expression, and jaw param-

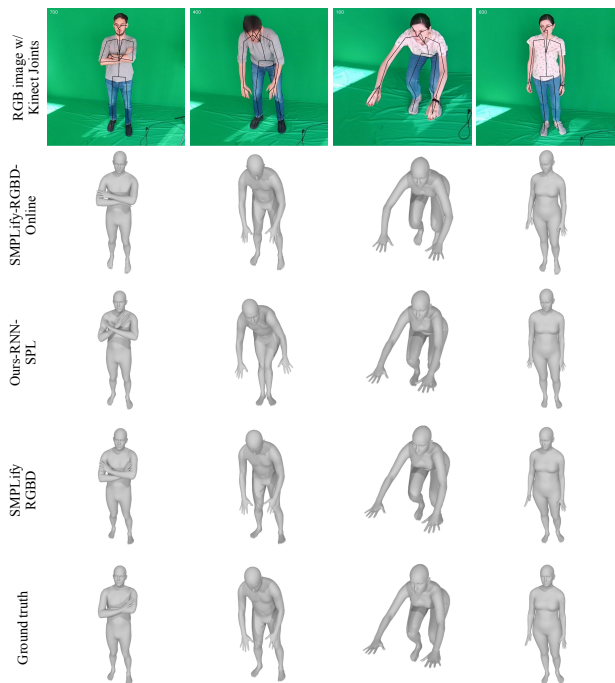


Figure 4. Comparison of the RGB-D baselines and the proposed methods on AzureTest frames. The second column shows an example where our method performs worse than optimization-based approaches. In other examples, the proposed Ours-RNN-SPL method matches the optimization methods closely.

eters of the SMPL-X model to fit the observations. We use the smooth l^1 regression cost for optimization and also add l^2 cost for smoothness regularization.

For our tasks, we require the landmark-anchored vertices, for which the average distance between them and corresponding Body Tracker landmarks is minimal. We find the optimal landmark-anchored vertices consistently with linear blend skinning (LBS) transformations (1). Firstly, we find a closest SMPL-X vertex i to a landmark over the SMPL-X annotated dataset, denoting its LBS weights as \mathcal{W}_i . Then, we optimize for the coordinates of a new virtual vertex v_m in a default body pose, fixing the LBS weights of this vertex equal to \mathcal{W}_i . Finally, we add a vertex with weights $\mathcal{W}_m = \mathcal{W}_i$ and coordinates v_m to the model.

4. Experiments

In this section, we evaluate the accuracies of our body pose and face extractors, and report timings of the obtained system. Additional qualitative results and comparisons are available in the **Supplementary video**.

Datasets. Firstly, we use the test set of the AzurePose dataset, collected by ourselves, see Section 3.4. As this dataset has low shape variability (just two people), we also report test results on the hold-out part of the AMASS-K dataset (140 sequences with 1120 frames).

To measure the quality of the face predictor we fil-

Method	Res	Input type				AMASS-K				AzurePose Test (simple/complex)			
		KP	β	Init	Twists	AUC \uparrow	Euler \downarrow	joint-ang \downarrow	Pos \downarrow	AUC \uparrow	Euler \downarrow	joint-ang \downarrow	Pos \downarrow
0	-	+	-	-	-	0.885	1.629	0.211	0.033	0.849	2.187	0.231	0.046
1	-	+	+	-	-	0.859	1.761	0.238	0.041	0.852	2.204	0.237	0.045
2	-	+	-	+	-	0.9	1.492	0.181	0.028	0.863	2.16	0.213	0.04
3	+	-	-	-	+	0.935	3.001	0.139	0.018	0.761	3.897	0.443	0.075
4	+	+	+	+	-	0.897	1.543	0.184	0.029	0.864	2.118	0.215	0.04
5	+	-	-	+	+	0.937	1.264	0.109	0.017	0.864	2.18	0.207	0.04

Table 3. Comparison of our models, with different types of inputs and outputs, see text. A model taking the initial pose θ_i^0 and the minimal bone-aligning rotations, and producing the rotation increments is the best on the shape-diverse AMASS-K dataset.

tered and annotated the test split of VoxCeleb2 dataset [12], which consists of subjects, who don't appear in the train set. For evaluation, we randomly selected 150 annotated test videos with 73 unique subjects. Our face evaluation set covers a wide variety of human face appearance, camera viewpoints, and motions.

Metrics. We use the set of metrics proposed by recent SMPL pose prediction work [4]. We obtain the *Euler angles* in the z, y, x order, choose the solution with the least amount of rotation, and compute a mean over the l^2 norm of the vectors of concatenated angles for all body joints. Next, we compute the *joint angle difference* as the mean l^2 norm of the rotation logarithm for each body joint. Following [4], we compute the Euler angle error on local rotations of joints with respect to their kinematic parents, and the joint angle difference on global rotations between the joint and the body root coordinate systems. We also compute the *positional* mean-squared error over the positions of the joints. Finally, we report the normalized *area under curve (AUC)* for the PCK_ρ values, where PCK_ρ is a probability for a positional error for a certain landmark to be lower than ρ . For face predictor evaluation we use MPJPE, which is an Euclidean distance between the ground-truth and predicted 3D keypoints. For computing MPJPE we use 68 SMPL-X head 3D keypoints. For comparing facial expression vectors we employ Mean Squared Error (MSE).

We run the experiments on a single server with AMD Ryzen Threadripper 1900X 8-Core CPU clocked at 3.8GHz and a NVIDIA GeForce RTX 2080Ti GPU.

Comparison with RGB-based methods. We start with comparing our approach to the recent RGB-only methods, in order to highlight the benefits of using the depth information. We use the recent single frame baselines SMPLify-X [35], SPIN [25], ExPose [11]. In a sequence-based setup, we compare against the state-of-the-art RGB video-based VIBE [24]. Although our method assumes known shape, other methods estimate the shape alongside with the pose.

The results on AzurePose, test in Table 1 indicate that our MLP-based model significantly and consistently outperforms the other methods in a single-frame experiment with respect to all metrics by 10-45%. As our method requires

SPL	Train		AzurePose test			
	AZ	AM-K	AUC \uparrow	Euler \downarrow	joint-ang \downarrow	Pos \downarrow
-	+	-	0.845	2.22	0.237	0.047
+	+	-	0.853	2.286	0.222	0.044
-	+	+	0.864	2.18	0.207	0.04
+	+	+	0.867	2.09	0.211	0.039

Table 4. Comparison of the RNN body pose prediction models with or without the SPL layer, trained either only on the AzurePose-train (AZ), or jointly on AzurePose-train and AMASS-K (AM-K). Joint training on two datasets increases pose prediction accuracy.

the body shape estimate β , we use a shape estimated by SPIN [25] as an input to our method. We note that body shape estimation may also benefit substantially from the depth information, but we leave this for future work and use the RGB-based body shape estimate.

In a multiframe experiment, we obtain the results which are better by 5-30% than the ones of the strongest RGB-based competitor (VIBE [24]) with respect to all the metrics except the local Euler angles on simple sequences, which can be explained by a bias toward complex body poses in the training data. We use β estimated by VIBE.

Overall, we observe a consistent and significant increase in the pose estimation accuracy when depth is used in addition to RGB, see also examples in Figure 3. We conclude from this experiment, that RGB-D-based human pose estimation still highly relevant for the tasks requiring high accuracy, robustness, and speed.

Comparison with optimization-based approaches.

Next, we compare our model with the single-view RGB-D methods. One of our baselines is the optimization-based offline method 'SMPLify-RGBD', which is a single view modification of the offline model fitting, see Section 3.4. Another one is 'SMPLify-RGBD-Online', which is a simplification of the former method to allow for the real-time performance, which does not use VPoser, but relies on the covariance-weighted l^2 regularization in the domain of the joint angles as a pose prior, with the covariance matrix computed from the AMASS dataset [28], and uses the l^2 -norm of the SMPL-X joint motion as a discontinuity cost. All the methods in this comparison received the ground truth body

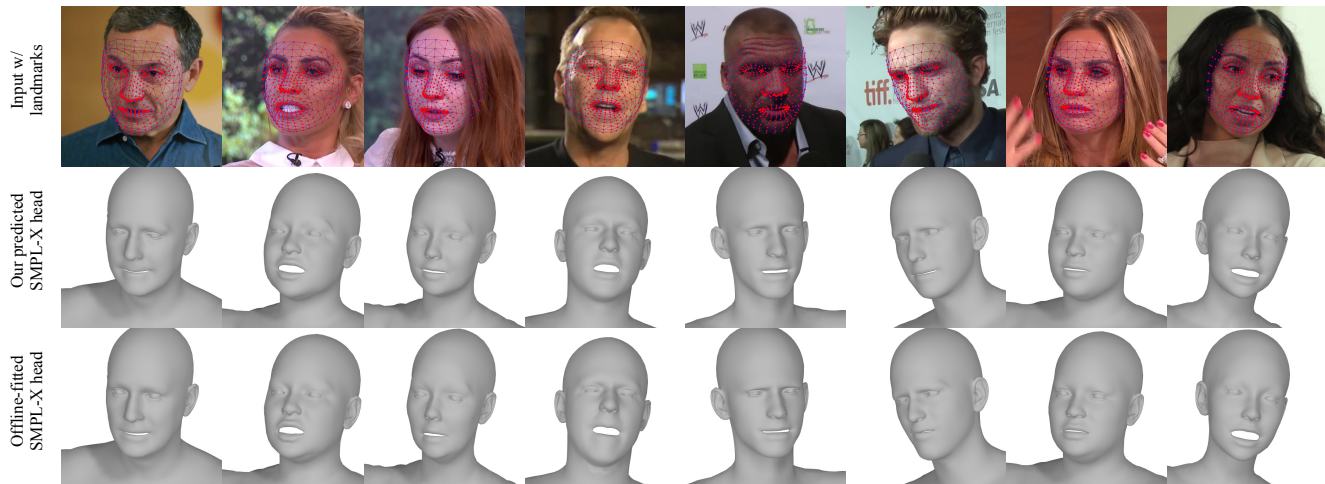


Figure 5. Random samples from VoxCeleb2 test dataset [12] with face model predictions. Top-to-bottom: input image with overlaid face landmarks [21], predicted SMPL-X model and offline-fitted SMPL-X model (used as ground truth during training, see section 3.4).

shape β .

The GRU network in our method serves as a fast replacement for the iterative optimization process, achieving similar accuracy, see Table 2. With our setting, the feedforward system is $2.5\times$ faster than ‘SMPLify-RGBD-Online’ and 25 times faster than ‘SMPLify-RGBD’. Comparison to Table 1 indicates the increase in accuracy from a better estimate of β .

On our desktop, the Body tracker takes 33 ms, MediaPipe [21] face and our face extractor takes 16ms, MinimalHand [50] takes 20 ms, our body pose extractor takes 40ms, and composer filter takes 8 ms. The whole system runs at 25 frames per second on a desktop computer with a single GPU. To compare, SMPLify-X [35], ExPose-X [11], SPIN [25] run at less than 1 FPS, VIBE [24] can run at almost 30 FPS on 2080ti GPU.

Ablation study. In Table 3, we evaluate ablations of our model. For an RNN-based model we consider four types of input: 3D body landmarks (KP), a shape vector β , the initial pose θ_i^0 (Init), and the minimal bone-aligning rotations $R_{i,k}$ (Twists); we compare models predicting rotation increments (Res+), or full rotations (Res-). The AMASS-K dataset has a diverse variety of body shapes, and the best accuracy on this dataset is achieved by a model with highest body shape generalization ability. A model 5 taking θ_i^0 and minimal bone-aligning rotations $R_{i,k}$ and producing rotation increments achieves lowest errors on AMASS-K, and has high accuracy on AzurePose Test. In Table 4 we show, that adding AMASS-K with its diverse poses and shapes to the training set helps pose estimation.

Face fitting evaluation. Next, we move on to the evaluation of the jaw pose and facial expression prediction on VoxCeleb2 test dataset [12]. We compare 3 modifications of our face predictor described in section 3.3 and report 3 metrics: MPJPE of SMPL-X 3D keypoints, MPJPE of SMPL-X mouth 3D keypoints, and MSE of facial expression vectors.

Model	\downarrow MPJPE mm	\downarrow MPJPE (mouth) mm	\downarrow Expression MSE
2D landmarks	2.715	3.623	2.264
2.5D landmarks	2.462	3.487	1.705
2.5D landmarks + mouth loss	2.326	3.395	1.478

Table 5. Face predictor evaluation metrics on VoxCeleb2 test dataset [12] for 3 modifications of our network. Usage of Z -coordinate of input 2.5D landmarks and additional weighing of SMPL-X 3D mouth keypoints in the loss improves metrics.

Table 5 summarizes the evaluation results. The first row represents evaluation results for the model, which inputs 2D landmarks (X, Y -coordinates). Next, we add Z -coordinate to the input, which significantly improves all the metrics. Our best model (last row) is additionally trained with $3\times$ weight on SMPL-X 3D mouth keypoints in the loss. We find it important to focus the model on the mouth because jaw pose mainly affects the mouth region.

5. Discussion

We have presented the details of the system that estimates the extended pose (including face and hands articulations) from RGBD images or videos in real time. Our systems builds on top of the previously developed components (Kinect pose tracker [6], MediaPipe face tracker [21], MinimalHand tracker [50]) and outputs the result in the popular SMPL-X format [35]. Importantly, we show that depth-based pose estimation still leads to considerable improvement in accuracy (and speed) compared to RGB-only state-of-the-art approaches. This is, of course, hardly surprising. What is perhaps more surprising given increasing availability of excellent RGB-D sensors is the relatively small interest towards depth-based pose estimation and the lack of available frameworks for extended pose estimation from RGB-D video streams. Our system will, hopefully, address this gap and will be useful to multiple applications in such domains as telepresence and human-computer interaction, where both the simplicity of setup and the accuracy of results are important.

References

- [1] K. Aberman, P. Li, D. Lischinski, O. Sorkine-Hornung, D. Cohen-Or, and B. Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62, 2020.
- [2] K. Aberman, R. Wu, D. Lischinski, B. Chen, and D. Cohen-Or. Learning character-agnostic motion for motion retargeting in 2D. *ACM transactions on graphics (TOG)*, 38(4):1–16, 2019.
- [3] E. Aksan, P. Cao, M. Kaufmann, and O. Hilliges. Attention, please: A spatio-temporal transformer for 3D human motion prediction. *arXiv preprint arXiv:2004.08692*, 2020.
- [4] E. Aksan, M. Kaufmann, and O. Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7144–7153, 2019.
- [5] A. Arnab, C. Doersch, and A. Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proc. CVPR*, pages 3395–3404, 2019.
- [6] Azure Kinect Body Tracking SDK. <https://docs.microsoft.com/en-us/azure/kinect-dk/body-sdk-download>. Accessed: 2020-09-28.
- [7] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *Proc. ECCV*, pages 561–578. Springer, 2016.
- [8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
- [9] Y. Chen, Y. Tian, and M. He. Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192:102897, 2020.
- [10] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 2014.
- [11] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black. Monocular expressive body regression through body-driven attention. In *Proc. ECCV*, 2020.
- [12] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.
- [13] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, pages 1–8. IEEE, 2008.
- [14] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. CVPR*, pages 1–8. IEEE, 2008.
- [15] R. Girshick. Fast R-CNN. In *Proc. ICCV*, pages 1440–1448, 2015.
- [16] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *Proc. ECCV*, pages 160–177. Springer, 2016.
- [17] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, 2015.
- [18] K. Isakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. In *Proc. ICCV*, pages 7718–7727, 2019.
- [19] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proc. CVPR*, pages 8320–8329, 2018.
- [20] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *Proc. CVPR*, pages 7122–7131, 2018.
- [21] Y. Kartynnik, A. Ablavatski, I. Grishchenko, and M. Grundmann. Real-time facial surface geometry from monocular video on mobile gpus. *arXiv preprint arXiv:1907.06724*, 2019.
- [22] Kinect gets UK release date. <http://www.bbc.co.uk/newsbeat/article/10996389/kinect-gets-uk-release-date>. Accessed: 2020-09-28.
- [23] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2014.
- [24] M. Kocabas, N. Athanasiou, and M. J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proc. CVPR*, June 2020.
- [25] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proc. CVPR*, pages 2252–2261, 2019.
- [26] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.
- [27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.
- [28] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. AMASS: Archive of motion capture as surface shapes. In *Proc. ICCV*, pages 5442–5451, 2019.
- [29] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proc. ICCV*, pages 2640–2649, 2017.
- [30] A. Martínez-González, M. Villamizar, O. Canévet, and J.-M. Odobez. Residual pose: A decoupled approach for depth-based 3D human pose estimation. In *Proc. IROS*, 2020.
- [31] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single RGB camera. *ACM Transactions on Graphics (TOG)*, 36(4):1–14, 2017.
- [32] G. Moon, J. Yong Chang, and K. Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proc. CVPR*, pages 5079–5088, 2018.
- [33] V. Nair and G. E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proc. ICML*, 2010.
- [34] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *Proc. ECCV*, pages 483–499. Springer, 2016.
- [35] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proc. CVPR*, pages 10975–10985, 2019.

- [36] D. Ramanan. Learning to parse images of articulated objects. In *Proc. NIPS*, 2006.
- [37] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245, 2017.
- [38] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, et al. Efficient human pose estimation from single depth images. *T-PAMI*, 35(12):2821–2840, 2012.
- [39] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proc. CVPR*, 2017.
- [40] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proc. CVPR*, pages 5693–5703, 2019.
- [41] X. Sun, J. Shang, S. Liang, and Y. Wei. Compositional human pose regression. In *Proc. ICCV*, pages 2602–2611, 2017.
- [42] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *Proc. ECCV*, pages 529–545, 2018.
- [43] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *Proc. CVPR*, pages 103–110. IEEE, 2012.
- [44] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *Proc. CVPR*, pages 1653–1660, 2014.
- [45] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [46] H. Xu, E. G. Bazavan, A. Zanfır, W. T. Freeman, R. Sukthankar, and C. Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Proc. CVPR*, pages 6184–6193, 2020.
- [47] M. Ye, X. Wang, R. Yang, L. Ren, and M. Pollefeys. Accurate 3D pose estimation from a single depth image. In *Proc. ICCV*, pages 731–738. IEEE, 2011.
- [48] A. Zanfır, E. G. Bazavan, H. Xu, B. Freeman, R. Sukthankar, and C. Sminchisescu. Weakly supervised 3D human pose and shape reconstruction with normalizing flows. In *Proc. ECCV*, 2020.
- [49] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *Proc. ECCV*, pages 186–201. Springer, 2016.
- [50] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proc. CVPR*, pages 5346–5355, 2020.