PanoMan: Sparse Localized Components–based Model for Full Human Motions

YUPAN WANG, GUIQING LI, HUIQIAN ZHANG, XINYI ZOU, YUXIN LIU, and YONGWEI NIE,
South China University of Technology, China

Parameterizing Variations of human shapes and motions is a long-standing problem in computer graphics and vision. Most of the existing methods only deal with a specific kind of motion, such as body poses, facial expressions, or hand gestures. We propose PanoMan (sParse locAlized compoNents based mOdel for full huMAn motioNs) to handle shape variation and full-motion across body, face, and hand in a unified framework. Like previous approaches, we factor shape variation into principal components to obtain a human shape space that approximates the shape of arbitrary identity. We then analyze sparse localized components in terms of relative edge length and dihedral angle to capture full motions of body poses, facial expressions, and hand gestures. The final piece of our model is a multilayer perceptron (MLP) that fits the residual between the ground truth and the aforementioned two-level approximation. As an application, we employ the discrete-shell deformation to drive the model to fit sparse constraints such as joint positions and surface feature points. We thoroughly evaluate PanoMan on body, face, and hand motion benchmarks as well as scanned data. The existing skinning-based techniques suffer from joint collapsing when encountering twisting motion of joints. Experiments show that PanoMan can capture all kinds of full human motions with high quality and is easier than the state-of-the-art models in recovering poses with wide joint twisting and complex hand gestures.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Shape modeling**; **Mesh models**; **Animation**; **Motion capture**;

Additional Key Words and Phrases: Human parametric model, PCA, sparse localized components, multilayer perceptrons

ACM Reference format:
Yupan Wang, Guiqing Li, Huiqian Zhang, Xinyi Zou, Yuxin Liu, and Yongwei Nie. 2021. PanoMan: Sparse Localized Components–based Model for Full Human Motions. *ACM Trans. Graph.* 40, 2, Article 19 (April 2021), 17 pages.
https://doi.org/10.1145/3447244

This work was partially supported by NSFC (61972160, 62072191, 61572202) and NSF of Guangdong, China (2021A1515012301, 2019A1515010860).
Authors' addresses: Y. Wang, (currently affiliated with) NetEase Games AI Lab; G. Li (corresponding author), H. Zhang, X. Zou, Y. Liu, and Y. Nie, School of Computer Science and Engineering, South China University of Technology, 510006, China; emails: 394013938@qq.com, ligq@scut.edu.cn, {791771249, 461423987, 465367868}@qq.com, nieyongwei@scut.edu.cn.

0730-0301/2021/04-ART19
https://doi.org/10.1145/3447244

1 INTRODUCTION

Human body shapes exhibit very significant variations across different individuals. Meanwhile, individual motions show complexity and diversity. This motivates the investigation of parameterizing body shapes and motions using a set of shape parameters and pose parameters, respectively.

Initial efforts in this regard may date back to the PCA (principal component analysis) models for representing variation of face [Blanz and Vetter 1999] or body [Allen et al. 2003] shapes. Since then, a variety of representations have been developed to parameterize human dynamic surface geometries such as body shapes and poses [Anguelov et al. 2005; Chen et al. 2013; Loper et al. 2015], facial shapes and expressions [Blanz and Vetter 1999; Bouaziz et al. 2013; Cao et al. 2014; Li et al. 2017], and hand gestures [Tkach et al. 2016]. Some of these models were further enhanced to more faithfully describe the effects of soft-tissue dynamics on motion bodies by either regressing the residuals of soft-issue deformation extracted from captured dynamic geometries [Loper et al. 2015; Pons-Moll et al. 2015] or physically simulating nonlinear elastic deformations of volumetric bodies [Kadleček et al. 2016; Kim et al. 2017; Xu and Barbič 2016].

Shortly afterwards, researchers realized that dealing with individual parts of the whole body independently is hard to meet the requirement of real-world applications in human computer interaction and virtual reality. SMPL+H [Romero et al. 2017] fuses motions of bodies and hands into one model. The Frankenstein model [Joo et al. 2018] is the first effort, which stitches SMPL [Loper et al. 2015], the bilinear facial model [Cao et al. 2014], and an artist-defined hand rig, to describe human dynamic geometries with full motions. The SMPL-X model [Pavlakos et al. 2019] integrates SMPL [Loper et al. 2015], FLAME [Li et al. 2017], and MANO [Romero et al. 2017] into a unified framework that is then trained using 3D scans to capture the natural correlations between the shape of bodies, faces, and hands. These models are directly built on SMPL, which is actually a linear blending skinning (LBS) or a dual-quaternion blend skinning (DQS) mounted with shape and pose control parameters. It is well known that LBS usually exhibits joint-collapsing for large twisting, while DQS tends to suffer joint-bulging for large bending [Jacobson and Sorkine 2011; Kim and Han 2014; Le and Hodgins 2016]. This is making harder the job of using SMPL-based models to reconstruct some extreme human dynamic geometries (see Figure 1), requiring careful tuning to fit these parametric models to data.

We propose PanoMan to unify the representation of body poses, facial expressions, and hand gestures together. It utilizes sparse localized components based on edge lengths and dihedral angles [Liu et al. 2019; Wang et al. 2017] to uniformly depict

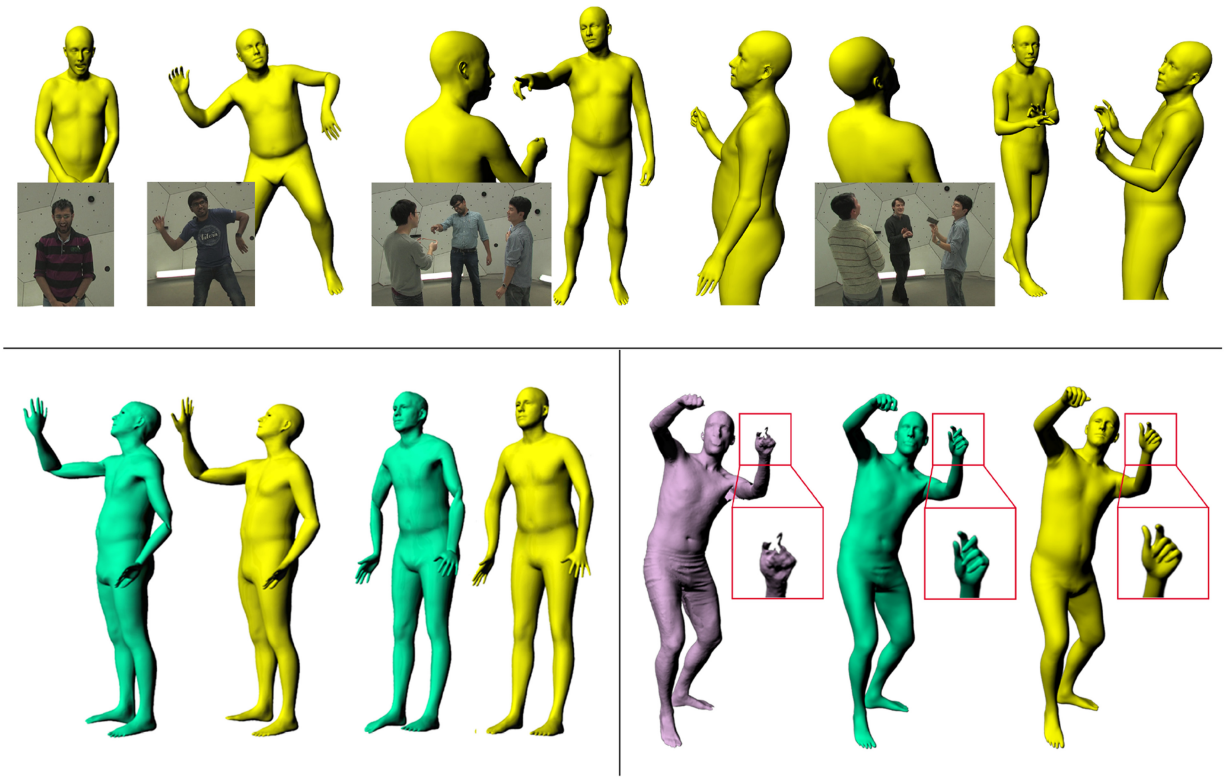

Fig. 1. Top row: Our parametric representation can simultaneously capture human body pose, facial expression and hand gesture very well. Bottom row: The left two examples demonstrate that SMPL (green) results in serious joint collapse for large twisting motion while our approach (yellow) still yields reasonable results; the rightmost compares PanoMan with SMPL+H (green). In the zoom-in, it can be observed that PanoMan better reconstructs the geometry of the data missing area.

motions of different body parts. Our key observation is that the sparse localized components are superior in characterizing different scales of motion owing to their locality. In Figure 1, the top row shows some reconstruction results mixing different scale motions, while the bottom row demonstrates that PanoMan outperforms the state-of-the-art methods in terms of recovering complex poses. We believe that this is an important supplement to the skeleton skinning–based parametric models such as SMPL and its variants.

Like previous parametric models, PanoMan is also data-driven. Therefore, we need to build a database of human dynamic geometries, which is synthesized from several independent recent datasets. There is a rest shape as well as a group of geometries with different body poses, facial expressions, and hand gestures for each subject of the database. All shapes have been carefully aligned and represented as meshes with the same connectivity. We directly employ 10 PCA bases of SMPL [Loper et al. 2015] to describe shape variation among different identities (subjects/people). For each identity, we compute the relative edge length and dihedral angle vector of a variety of body poses, facial expressions, and hand gestures. We then conduct the sparse localized decomposition over the vectors from all identities to yield a set of sparse localized components. Finally, we introduce a correction term, which is defined as a function of motion blending coefficients to further improve the representation accuracy. Our contributions are as follows:

- A parametric model named PanoMan is built to encode full human body dynamic geometries, which consists of three parts: a set of PCA bases for blending human body shape, a set of sparse localized components for representing full-scale motions across body poses, facial expressions, and hand gestures, and a correction term for reducing the reconstruction error caused by components truncation;
- The correction term is estimated by using a multilayer perceptron (MLP) that is responsible to regress the reconstruction error of training data in terms of blending (motion) coefficients of the sparse localized components; relative edge lengths and dihedral angles are introduced to reduce the impact of edge variation across different identities in the sparse localized decomposition;
- A training dataset is synthesized from a set of heterogeneous body, face, and hand databases. Although the dataset does not contain samples simultaneously involving arbitrary body poses, facial expressions, and hand gestures, experiments show that it is sufficient for our scenarios;
- A framework is developed to apply PanoMan to recover 3D dynamic geometries with full motion of body, face, and hands

from 3D sparse constraints; a multilayer perceptron is again trained to estimate initial blending coefficients of sparse localized components with respect to skeleton motion parameters for solving the reconstruction framework; experiments show that our framework is able to achieve high performance both in recovery of poses with wide joint twisting and in 3D reconstruction accuracy.

2 RELATED WORK

A great deal of literature has been contributed to investigate the parametric representations of body poses, facial expressions, and hand gestures, separately. Some of them only concern shape variation of individuals, while others further involve their motions. Here, we mainly discuss statistical parametric models and then briefly review some approaches about the sparse localized components. More details about body and facial parametric representations can be found in Orvalho et al. [2012] and Cheng et al. [2018], respectively.

3D deformable models. Allen et al. built the first PCA model for human bodies [Allen et al. 2003] using CAESAR scans. Although trying to use the motion data and skinning weights of a pose to animate the parameterized ones, they did not couple the two ingredients together to describe the variation of both shape and motion.

An early method achieving this is SCAPE [Anguelov et al. 2005] using deformation gradients of triangles to capture the pose change. It decomposes the deformation of each triangle into shape deformation, pose-related deformation, and deformation correction. Jain et al. [2010] greatly reduced the complexity of SCAPE by replacing per-triangle transformation with skeletal skinning to serve their video editing goal, which does not require much details. Pishchulin et al. named this method as simplified SCAPE (S-SCAPE) [Pishchulin et al. 2017]. Hirshberg et al. [2012] addressed the convergence issue of SCAPE by introducing BlendSCAPE, which views the transformation of triangles as the linear blending of rigid patches. Some methods improved SCAPE model in either accuracy [Chen et al. 2013, 2019] or time cost [Chen et al. 2016] while Dyna [Pons-Moll et al. 2015] extends BlendSCAPE to capture secondary motions of soft-tissues. Though SCAPE has been widely used in motion capture [Loper et al. 2014] and 3D body reconstruction [Cheng et al. 2016; Song et al. 2016; Zheng et al. 2014], decoupling shape and pose deformations makes SCAPE difficult to characterize subtle difference across individuals doing the same pose [Chen et al. 2013]. As corrected components are learned from examples by linear regression, SCAPE based methods may results in serious visual artifacts [Weber et al. 2007] similar to skeletal skinning.

Though designed for deforming general objects, some general data-driven deformation approaches [Gao et al. 2017, 2016; Heeren et al. 2018; Tan et al. 2018a] show strong ability to capture complicated articulated poses such as human body and finger motions. Nevertheless, they do not simultaneously encode shape variations of human body geometries in a unified framework.

Skeleton skinning based models. SMPL (Skinned Multi-Person Linear Model) [Loper et al. 2015] employs PCA bases of vertex coordinates to capture shape variation and leverages a skeleton skinning technique to represent different poses. The model is faster and more straightforward compared than SCAPE-based model, though it can be viewed as an evolution of the S-SCAPE [Jain et al. 2010]. Compared to the latter, it not only trains the skinning weights from scanned data but also provides a motion correction with respect to shape difference. It has been widely used in many fields in recent years, such as motion capture and 3D reconstruction from depth maps [Bhatnagar et al. 2020; Yu et al. 2018; Zhang et al. 2017; Zheng et al. 2020], images [Alldieck et al. 2019a; Bogo et al. 2016; Pavlakos et al. 2018], or 2D/3D videos [Alldieck et al. 2018; Lassner et al. 2017; Pons-Moll et al. 2017]. Alldieck et al. [2019b] even augmented SMPL using two UV maps (normal map and vector displacement map) to capture detailed shape geometries such as cloth and skin wrinkles.

In view of the success of SMPL in representing personalized human body geometries, Li et al. [2017] transplanted it to parameterize head shapes and motions including facial expressions, while Romero et al. [2017] built a statistical hand model called MANO by decomposing hand motion data into PCA bases. Combining the body SMPL and hand MANO together renders SMPL+H, which represents human body poses and hand gestures simultaneously. However, Kim et al. extended SMPL to VSMPL (a volumetric SMPL) for further parameterizing secondary motions of soft-tissues.

Inspired by the part-based stitched puppet model by Zuffi and Black [2015], which codes the variants of segmented parts and combines them together, Joo et al. [2018] proposed the Frankenstein model to represent the motion of face, hand, and body. They employed SMPL to characterize body and hand variations but leverage the bilinear model to express facial expressions. Noticing Frankenstein is not fully realistic, Pavlakos et al. seamlessly fused SMPL+H and FLAME [Li et al. 2017] into a so-called SMPL-X model [Pavlakos et al. 2019] that is recently employed to build body motion datasets for grasping objects [Rong et al. 2020; Taheri et al. 2020]. Though being able to recover expressive results for conventional poses, these approaches suffer from the same issue as SMPL, i.e., making it harder to avoid candy-wrapper or bulging effects when applied to reconstruct poses with large joint rotation.

Instead of using skeletal motion parameters, we capture the pose variation via sparse localized components of edge lengths and dihedral angles. As our model is essentially a thin-shell deformation method, it is easier to avoid joint collapse. In addition, we employ an MLP to directly fit the reconstruction residual with respect to blending coefficients of the components, while SMPL modifies the rest pose to achieve the goal. This makes our approach easier to control error.

Parametric models of human faces. Blendshape is the most popular representation for early facial animation [Williams 1990]. It creates vertex displacement bases for each of the facial feature points and represents the animation of a personalized neutral face by blending these bases. However, this model does not capture shape variation. On the contrary, 3DMM (3D morphable model) by Blanz and Vetter [1999] focuses on capturing shape variation but does not support expressions, which consists of a set of PCA bases extracted from 3D facial models from different subjects. Amberg et al. [2008] enhanced the model by training it with a new dataset much larger than the original one.

Weise et al. [2011] designed a realtime performance-based facial animation system by combining Blendshape and 3DMM. Bouaziz et al. [2013] built an adaptive DEM (Dynamic Expression Model) upon 3DMM to simultaneously capture facial shape and expression variations. Cao et al. [2014] developed FaceWarehouse to learn a bilinear face model, which is similar to Chen et al. [2013], to represent the variations of face shapes and expressions. Garrido et al. [2016] established a comprehensive facial model based on 3DMM and DEM with a layer of vertex-based displacement details. FLAME [Li et al. 2017] extends SMPL with additional expression blendshapes to model head pose, facial shape, and expression, which is more accurate than previous representations. Based on FLAME, Sanyal et al. [2019] trained a RingNet without 3D supervision to recover 3D facial shape and expression from a single image. Our approach makes use of exactly the same type of components as body motion to describe facial expression.

Parametric models of human hands. Most parametric models for human hands were proposed for hand motion tracking [Mueller et al. 2018; Qian et al. 2014; Sun et al. 2015] instead of precise geometry reconstruction. The skeletal skinning technique is often used [Ballan et al. 2012; Romero et al. 2017; Sharp et al. 2015; Sridhar et al. 2015; Taylor et al. 2016]. However, implicit representations built on skeletal meshes are also popular [Tkach et al. 2016, 2017], which can provide more superior tracking performance. However, most of aforementioned approaches are not statistics based except for MANO (hand part of SMPL+H) in Romero et al. [2017]. Zhang et al. [2019] regressed the MANO parameters from a single image by using an end-to-end deep neural network. Again, our method employs sparse localized components to capture hand gesture and therefore maintaining consistency in encoding three parts of animation.

Sparse localized decomposition. Sparse localized decomposition [Neumann et al. 2013] imposes locality on principal components to achieve local control of facial animations. It fails to work with articulated rotational motions. Huang et al. [2014] addressed the issue by performing the decomposition on deformation gradients [Sumner and Popović 2004]. As deformation gradients are neither invariant nor localized up to rotation, this method cannot handle motions with global or large rotation. The articulated-motion-aware sparse localized components by Wang et al. [Liu et al. 2019; Wang et al. 2017] are obtained by analyzing the variations of edge lengths and dihedral angles (LAs). The authors showed that these quantities possess good locality and linearity up to rotation. Sassenm et al. [2020] introduced sparse principal geodesic analysis on the representation of LAs to better capture the articulation of the input shapes. Tan et al. [2018b] used mesh-based autocoders to extract sparse localized components. The idea is also extended to analyze localized vibration modes [Brandt and Hildebrandt 2017]. These methods only describe motion variations from a specific object instead of behaviors of a group of identities such as humanity.

3 OVERVIEW

Our work consists of two parts: building a parametric model, i.e., PanoMan, for human body dynamic geometries with full motion based on our synthesized dataset, and reconstructing human dynamic geometries by fitting PanoMan to sparse 3D surface constraints.

As shown in Figure 2, we reuse existing human motion databases including CAESAR, FacewareHouse, SPML+H to synthesize our training dataset, which consists of a group of full motion human body meshes with the same connectivity for building PanoMan (Section 4.4). First, we directly adopt the average shape and the 10 PCA bases of SMPL to parameterize the shape of different identities (see Section 4.2). Following, the sparse localized decomposition is employed to extract motion bases from a set of relative edge length and dihedral angle (RLA) vectors of arbitrary poses (Section 4.3).

Given joint positions or feature points of a specific pose of arbitrary identity, reconstructing its PanoMan representation can be reformulated as a discrete-shell deformation problem. In such a framework, vertices and parameters of PanoMan of the deformed mesh are jointly optimized, as we cannot express the vertex constraint as the function of the parameters. If joint positions are known, then we convert them into vertex constraints by linear regression.

Notations. In our setting, all human shapes and poses are represented using a triangular mesh template with fixed connectivity. Denote the template by $\mathcal{M} = \langle V, E, F \rangle$, where V, E and F are vertex, edge and facet sets, respectively. We always denote the element number of set S by $|S|$, and henceforth have $E \subset \{1, 2, \cdots, |V|\} \times \{1, 2, \cdots, |V|\}$ and $F \subset \{1, 2, \cdots, |V|\}^3$. Furthermore, $|V|$, E, and F remain the same for all shapes except for vertex positions. Without loss of generality, we also use V to represent the $3|V|$-dimensional vector concatenating all vertex coordinates sequentially.

4 PANOMAN MODEL

As we have shown in Figure 2, PanoMan tries to use two groups of parameters to precisely capture shape variation and full pose change of human. As mentioned in the overview, we first introduce relative edge lengths and dihedral angles (Section 4.1) and then construct PanoMan (Section 4.2). Following that, synthesis of training data is discussed (Section 4.3). Training details are given in the end (Section 4.4).

4.1 RLA Representation of Meshes

Lengths and dihedral angles of all edges of a mesh can uniquely determine the shape of the mesh up to a rigid transformation [Winkler et al. 2010], where the dihedral angle of an edge is the inner included angle of its two adjacent triangles. We call the vector concatenating all these edge lengths and dihedral angles together an LA vector of the mesh. By performing the sparse localized decomposition [Neumann et al. 2013] over the LA vectors of animated meshes, Wang et al. [2017] extracted a set of components being aware of local rotations. Observing that the change of edge length and dihedral angle is relatively independent, Liu et al. [2019] further improved the approach by decoupling two types of quantities and introducing adaptive support for components to interpret different scales of motion.

RLA vectors. The aforementioned animated meshes are usually about the motion of a specific object. Mesh frames in our setting come from different identities and therefore edge length variation

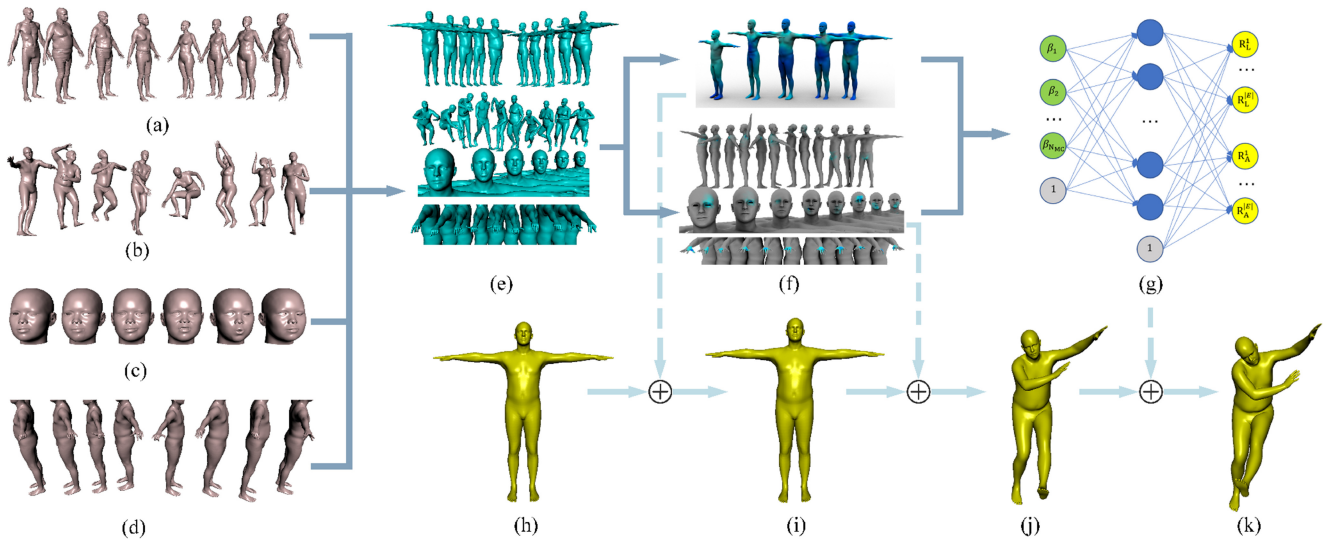

Fig. 2. The pipeline of computing the parametric model for human dynamic geometries of full motion: (a–d) show, respectively, the shape, body pose, face expression, and hand gesture databases that we used to compose our training data (e); (f) show some shape PCA bases and sparse localized components for poses, expressions, and gestures extracted from the shape and motion databases in (e); (g) an MLP learned from (e) and (f); (h) average shape $\overline{V}$; (i) specific shape $B_S(\alpha) = \overline{V} + S\alpha$; (j)$B_P(\alpha, \beta) = g(M_L\beta_L, M_A\beta_A, B_S(\alpha))$, a specific pose of (i); (k) $B_P(\alpha, \beta) = g(X_L(\alpha, \beta), X_A(\alpha, \beta), B_S(\alpha))$, correction of (j). All stages are discussed in Section 4.

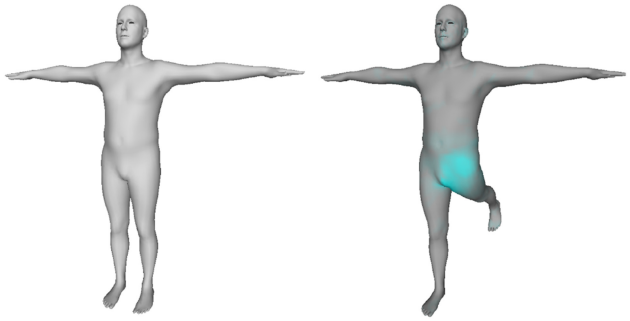

Fig. 3. Locality of RL and RA change in articulated motions: (a) the rest pose; (b) the deformed shape with local RL and RA change.

not only accounts for motion but also occurs when shape changes. To cancel the shape impact on edge length variation, we introduce relative edge length (RL) and relative dihedral angle (RA). Specifically, let $r(V)$ be a rest pose of a specific identity and V be its new pose. $\forall e = (i, j) \in E$, we denote its length and dihedral angle in $r(V)$ by $l_r(e)$ and $a_r(e)$ separately, and in V by $l(e)$ and $a(e)$ separately. RL and RA of e are then defined as

$$\hat{l}(e) = \frac{l(e)}{l_r(e)} - 1, \hat{a}(e) = a(e) - a_r(e). \tag{1}$$

Concatenating all $\hat{l}(e)$ and $\hat{a}(e)$ together, respectively, leads to the so-called RL vector $X_L(r(V), V)$ and RA vector $X_A(r(V), V)$ of V (with respect to $r(V)$). Both vectors are of $|E|$ dimension. Similar to LA vectors, RL and RA vectors are also articulated-motion-aware (see Figure 3).

Shape reconstruction from RL and RA vectors. Given rest pose $r(V)$ and the RL and RA vectors of V, we can easily obtain $l(e)$ and $a(e)$ from Equation (1). The reconstruction Algorithm described in Wang et al. [2017] can then be used to recover V. Denoting the algorithm by g yields $V = g(X_L(r(V), V), X_A(r(V), V), r(V))$. As g will play the initialization role in the reconstruction algorithm of Section 5.3, we present more details in Appendix A.

4.2 Formulation of PanoMan

We do not need to acquire frames simultaneously containing complex pose, expression, and gesture for building PanoMan. It is actually sufficient to collect a dataset in which each mesh only includes one of the body pose, facial expression, and hand gesture due to the locality of sparse localized components. Specifically, our dataset consists of three parts:

$$\mathcal{H} = \mathcal{H}_B \cup \mathcal{H}_F \cup \mathcal{H}_H,$$

where $\mathcal{H}_B$, $\mathcal{H}_F$, and $\mathcal{H}_H$, respectively, denote subsets for extracting pose, expression, and gesture sparse localized components. All poses in $\mathcal{H}$ have been aligned with template $\mathcal{M}$. Its construction will be described in Section 4.3.

PanoMan is made up of three modules: a group of PCA bases that capture the shape variation across different identities, a set of sparse localized components that are used to describe human motion, and a correction term that is responsible for compensating the motion residuals. This section focuses on how to assemble three parts together while their computation will be detailed in Section 4.4.

Shape PCA. We directly take the 10 PCA bases, which are available as open source of SMPL, to capture human shape variation in our setting. Let $\overline{V}$ be the mean shape and $S = [S_1, \cdots, S_{10}]$ be the matrix of 10 $3|V|$-dimensional bases (S_i a column vector). The rest shape of arbitrary human can then be approximated by

$$B_S(\alpha) = \overline{V} + S\alpha, \tag{2}$$

where $\alpha = [\alpha_1, \cdots, \alpha_{10}]^T$ is the vector of blending weights. In this way, $B_S(\alpha)$ can not only capture body shape but also approach facial and hand shapes of arbitrary identity as the training set includes the corresponding data.

Motion sparse localized decomposition. Our motion decomposition works in the spaces of RL and RA vectors of $\mathcal{H}_B, \mathcal{H}_F$, and $\mathcal{H}_H$, respectively. In fact, we totally conducted six times of independent decomposition.

Let us take the RL space of body motion $\mathcal{H}_B$ as example. $\forall V \in \mathcal{H}_B$, we always know its identity as well as the corresponding rest pose $r(V) \in \mathcal{H}_B$. This enables us to compute its RL vector $X_L(r(V), V)$. Performing the adaptive sparse localized decomposition [Liu et al. 2019] on $\{X_L(r(V), V) : V \in \mathcal{H}_B\}$ yields the average

$$\overline{X}_{LB} = \frac{1}{|\mathcal{H}_B|} \sum_{V \in \mathcal{H}_B} X_L(r(V), V),$$

and RL sparse localized components about body poses

$$M_{LB} = \left[M_1^{LB}, \cdots, M_{|M_{LB}|}^{LB}\right].$$

In the same way, we can obtain the average and sparse localized components for blending facial expressions from $\mathcal{H}_F$

$$\overline{X}_{LF} = \frac{1}{|\mathcal{H}_F|} \sum_{V \in \mathcal{H}_F} X_L(r(V), V), M_{LF} = \left[M_1^{LF}, \cdots, M_{|M_{LF}|}^{LF}\right],$$

and the average and sparse localized components for hand gesture from $\mathcal{H}_H$

$$\overline{X}_{LH} = \frac{1}{|\mathcal{H}_H|} \sum_{V \in \mathcal{H}_H} X_L(r(V), V), M_{LH} = \left[M_1^{LH}, \cdots, M_{|M_{LH}|}^{LH}\right].$$

Replacing RL vectors with RA vectors leads to the other three groups of decomposition:

$$\overline{X}_{AB} = \frac{1}{|\mathcal{H}_B|} \sum_{V \in \mathcal{H}_B} X_A(r(V), V), M_{AB} = \left[M_1^{AB}, \cdots, M_{|M_{AB}|}^{AB}\right];$$

$$\overline{X}_{AF} = \frac{1}{|\mathcal{H}_F|} \sum_{V \in \mathcal{H}_F} X_A(r(V), V), M_{AF} = \left[M_1^{AF}, \cdots, M_{|M_{AF}|}^{AF}\right];$$

$$\overline{X}_{AH} = \frac{1}{|\mathcal{H}_H|} \sum_{V \in \mathcal{H}_H} X_A(r(V), V), M_{AH} = \left[M_1^{AH}, \cdots, M_{|M_{AH}|}^{AH}\right].$$

For sake of simplicity, we introduce total edge length component matrix $M_L = [M_{LB}, M_{LF}, M_{LH}]$ and total dihedral angle component matrix $M_A = [M_{AB}, M_{AF}, M_{AH}]$. Now, given motion blending weight vector $\beta = [\beta_L, \beta_A]^T$, where $\beta_L = [\beta_{LB}, \beta_{LF}, \beta_{LH}]^T$ and $\beta_A = [\beta_{AB}, \beta_{AF}, \beta_{AH}]^T$, and β_{YZ} a $|M_{YZ}|$-dimensional column vector for $Y \in \{L, A\}$ and $Z \in \{B, F, H\}$, we can obtain RL and RA vectors of full pose $B_P(\alpha, \beta)$ for specific identity $B_S(\alpha)$ by blending the corresponding components:

$$\begin{cases} X_L(B_S(\alpha), B_P(\alpha, \beta)) = \overline{X}_L + M_L\beta_L, \\ X_A(B_S(\alpha), B_P(\alpha, \beta)) = \overline{X}_A + M_A\beta_A, \end{cases} \tag{3}$$

where

$$\overline{X}_L = \sum_{Z \in \{B,F,H\}} \overline{X}_{LZ}, \overline{X}_A = \sum_{Z \in \{B,F,H\}} \overline{X}_{AZ}.$$

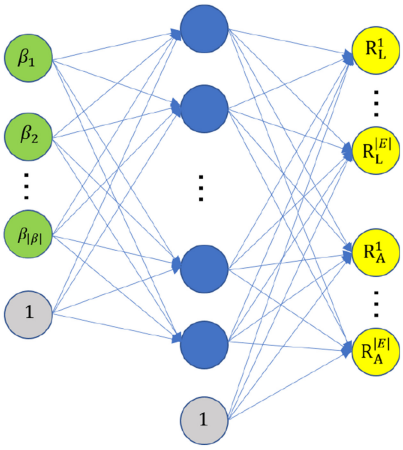

Fig. 4. The MLP for fitting the residual function $R(\beta)$ includes an input layer of motion parameters, a hidden layer with 1,500 nodes, and an output layer of edge length and dihedral angle compensation.

Furthermore, with shape and motion bases at hand, the parameter vectors α (shape) and β (pose) will determine a concrete pose for a specific identity according to g procedure:

$$\begin{aligned} B_P(\alpha, \beta) = g(&X_L(B_S(\alpha), B_P(\alpha, \beta)), \\ &X_A(B_S(\alpha), B_P(\alpha, \beta)), B_S(\alpha)). \end{aligned} \tag{4}$$

Correction MLP. It usually requires a great number of sparse localized components for Equation (4) to achieve high precision. We instead truncate the component set to a proper size and further use an MLP to regress RL and RA errors with respect to pose parameters β as compensation. The MLP contains one hidden layer with 1,500 nodes and the ReLU is taken as the activate function (see Figure 4). Denoting the output of the network as $R(\beta) = [R_A(\beta), R_L(\beta)]^T$, we obtain corrected RL $X_L(\alpha, \beta) = X_L(B_S(\alpha), B_P(\alpha, \beta)) + R_L(\beta)$ and corrected RA $X_A(\alpha, \beta) = X_A(B_S(\alpha), B_P(\alpha, \beta)) + R_A(\beta)$. Equation (4) can then be refined to

$$B_P(\alpha, \beta) = g(X_L(\alpha, \beta), X_A(\alpha, \beta), B_S(\alpha)). \tag{5}$$

4.3 Synthesis of Dataset $\mathcal{H}$

Synthesizing labeled samples is an important way to augment training data. For example, Richardson et al. [2016] learned a ResNet using training samples synthesized by the simplified DEM in Chu et al. [2014] to reconstruct specific 3D facial shapes and expressions. In our setting, it is almost impossible to thoroughly capture human shape and motion data containing body, face, and hand dynamic geometries in one frame. Fortunately, quite a few datasets have been captured for each of the three kinds of data. Reusing these datasets to serve our task is a natural and reasonable choice. We first introduce four databases and then discuss how to build our dataset $\mathcal{H}$.

DFAUST [Bogo et al. 2017]: DFAUST is a dataset of high-resolution 4D real scans of human in motion, captured at 60 FPS. Total 40,000 meshes were acquired for 10 subjects (5 men and 5 women) across a variety of poses. All meshes have been registered with the SMPL template of 6,890 vertices and 13,766 faces. We sample half of the meshes (10,000 for each gender) and refine

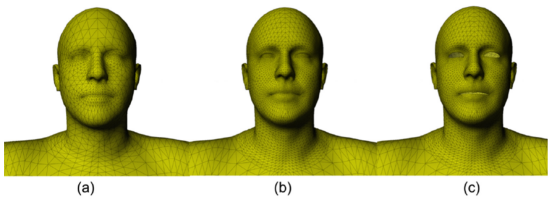

Fig. 5. Refinement of the DFAUST and SURREAL datasets: (a) the original mesh, (b) the refined version, and (c) the refined mesh with eyeball and lip areas removed.

their head region by using butterfly subdivision [Zorin et al. 1996] while keeping the connectivity of the refined meshes compatible. In addition, we remove eyeball regions and the in-between area of lips for conveniently treating facial expressions. This leads to a dataset with all meshes having $|V| = 10{,}945$ vertices, $|E| = 32{,}690$ edges, and $|F| = 21{,}744$ faces, as shown in Figure 5.

SURREAL [Varol et al. 2017]: SURREAL is a large-scale dataset created by synthesizing together a variety of virtual human poses and various real backgrounds [Varol et al. 2017] to obtain images with annotated poses. It makes use of the motion data of the CMU database [CMU 2000] to drive SMPL [Loper et al. 2015] for generating different poses of a specific identity. Total 8,000 mesh models are finally generated for 400 identities (half male and half female) with each having 20 poses. Noting that the mesh template of SURREAL inherits from SMPL, we employ the same strategy to refine the meshes, as shown in Figure 5.

FacewareHouse and SMPL+H: As DFAUST and SURREAL do not involve facial expressions and hand gestures, we need other two databases. FacewareHouse by Cao et al. [2014] consists of head meshes from 200 males and 200 females identities with 40 expressions (1 for neutral expression). SMPL+H [Romero et al. 2017] is a parametric representation for body and hand motion. Its hand dataset [Romero et al. 2017] consists of 2,018 scans from 31 identities but has not been made public. We again use SMPL+H to generate meshes with rest body pose and a variety of hand gestures as our hand motion data.

Our training data $\mathcal{H}$ is built upon the above databases (see Figure 6 for samples). Like previous methods, we also establish our model for male and female, respectively, and therefore synthesize a dataset for each gender. For simplicity, we summarize generation of the male dataset:

- We directly adopt 10 shape bases from SMPL [Loper et al. 2015] with the facial region of all base meshes being refined with Butterfly subdivision;
- Body pose subset $\mathcal{H}_B$ consists of two parts, one from SURREAL and the other from DFAUST; Specifically, we select $200 \times 20 = 4{,}000$ meshes (200 identities with 20 poses) from SURREAL and $5 \times 400 = 2{,}000$ meshes (5 identities and 400 frames) from DFAUST; All meshes are subdivided to have the same connectivity as shape meshes;
- Facial expression subset $\mathcal{H}_F$ includes $100 \times 39 = 3{,}900$ meshes, which are generated by transferring 39 expressions of FacewareHouse [Cao et al. 2014] onto the first 100 identities in SURREAL using deformation transfer [Sumner and Popović 2004];

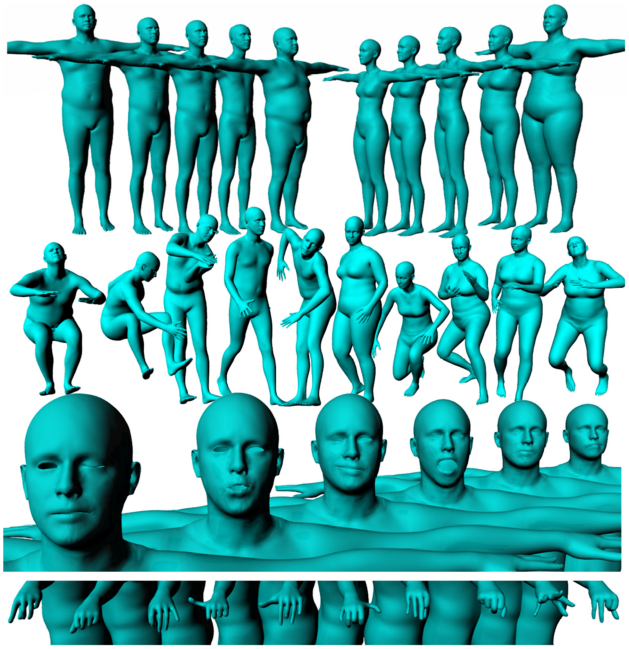

Fig. 6. Some samples of the training dataset: rest shapes of five males and five females (row 1), different poses from different males and females (row 2), facial expressions from the same identity (row 3), and hand gestures of the same identity (row 4).

- Hand gesture subset $\mathcal{H}_F$ has $300 \times 20 = 6{,}000$ meshes (300 identities and 20 poses), which are obtained by sampling the hand pose parametric domain of MANO [Romero et al. 2017] under Gaussian distribution to guarantee the diversity.

It should be emphasized that just copying the facial and hand motions onto rest body poses is sufficient owing to the locality of sparse localized components. In addition, considering that the above training data does not include full motion, we make use of other more complex databases to evaluate our PanoMan and the state-of-the-art models (see Section 6).

4.4 Training

Only motion bases remain to be solved in our training process. This includes two stages: adaptive sparse localized decomposition for extracting motion bases and MLP fitting for remedying the accuracy loss due to making a tradeoff between the component number and approximating error.

Shape PCA bases. In our implementation, we directly employ the average $\overline{V}$ and 10 PCA shape bases of SMPL to capture human shape variation.

Sparse localized decomposition. We perform the adaptive sparse localized decomposition [Liu et al. 2019] on $\mathcal{H}_B$, $\mathcal{H}_F$, and $\mathcal{H}_H$ independently. The decomposition is actually repeated six times to obtain elements in $\{(\overline{X}_{YZ}, M_{YZ}) : Y \in \{L, A\}, Z \in \{B, F, H\}\}$ separately (see Section 4.2). We refer readers to Neumann et al. [2013] and Liu et al. [2019] for more details on the decomposition algorithm.

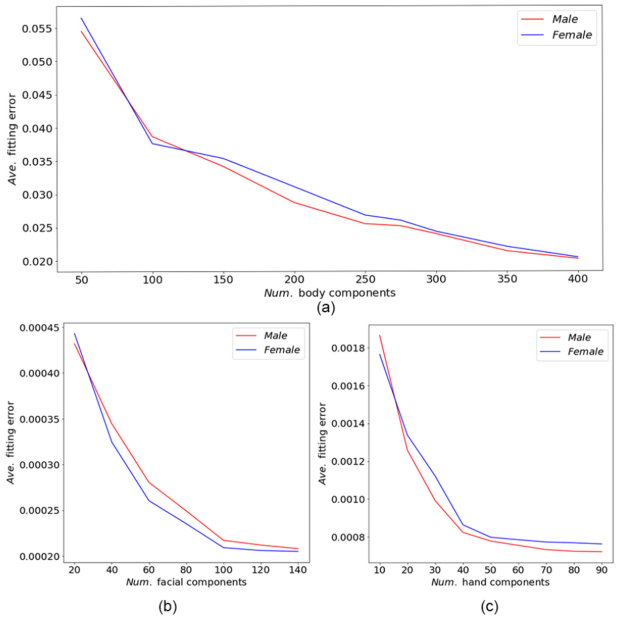

Fig. 7. Average error curves of approximating RLA vectors of $\{V \in \mathcal{H}_B\}$ with respect to numbers of the corresponding sparse localized components for (a) body poses, (b) facial expressions, and (c) hand gestures. Note that we use the same number of components for dihedral angle and edge length. In addition, the analysis is conducted for both male (red) and female (blue).

Increasing components usually improves the expressive capability of the parametric model but also increases the subspace dimensionality and computational complexity in reconstruction. To determine a rational size, we use Equation (4) with different number of sparse localized components to reconstruct our training data and estimate the average reconstruction error. Figure 7(a) indicates that after 250, increasing the body component number does not significantly reduce reconstruction error, therefore, we set $|M_{LB}| = |M_{AB}| = 250$. Similarly, Figures 7(b) and (c), respectively, suggest that $|M_{LF}| = |M_{AF}| = 100$ and $|M_{LH}| = |M_{AH}| = 50$ are good for facial and hand components, separately. Though this leads to $|M_L| = |M_A| = |M_{LB}| + |M_{LF}| + |M_{LH}| = 400$, we have only about $52|E|$ nonzero entries owing to the sparsity of components. Some components are shown in rows 1, 2, and 3 of Figure 8 for body poses, facial expressions, and hand gestures, where the length and dihedral angle of an edge are combined together to compute pseudo color.

Regression of MLP correction term. Figure 7 shows that the accuracy of the facial and hand gestures is higher than the accuracy of the body. Therefore, we regress a correction term with respect to body motion weights. Specifically, $\forall V \in \mathcal{H}_B$, we first estimate its edge length and dihedral angle vectors X_L and X_A, and then fit $X_L(B_S(\alpha), B_P(\alpha, \beta))$ to X_L, and $X_A(B_S(\alpha), B_P(\alpha, \beta))$ to X_A. The residual $R(V) = (R_L(V), R_A(V))$ is defined as

$$R_L(V) = X_L - X_L(B_S(\alpha), B_P(\alpha, \beta)),$$

$$R_A(V) = X_A - X_A(B_S(\alpha), B_P(\alpha, \beta)).$$

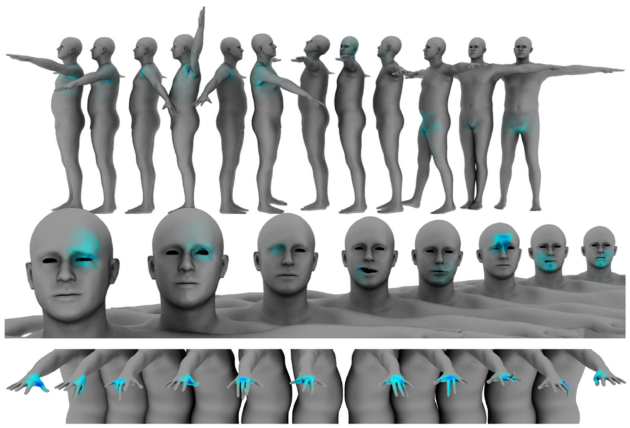

Fig. 8. Illustration of some sparse localized components on the body part (row 1), face area (row 2), and hand regions (row 3). The non-zero region of each component is highlighted in blue, and the darker the color becomes, the greater the magnitude is.

Note that meshes in $\mathcal{H}_B$ do not contain expression and gesture, and $\beta_{LF}, \beta_{AF}, \beta_{LH}$, and β_{AH} in β remain unchanged for all samples in $\mathcal{H}_B$. Henceforth, we feed (β_{LB}, β_{AB}) into the MLP (see Figure 4) to output the corresponding $(R_L(V), R_A(V))$ during training.

5 RECONSTRUCTION BASED ON PANOMAN

Establishing human parametric models enables us to constrain the human shapes and motions within a subspace. Their most important application is to help reconstruct instances stably with sparse conditions. This section proposes a framework to apply PanoMan to reconstruct 3D poses of a specific subject. Given some constraints about the underlying geometry of the specific identity such as vertex positions, triangle orthogonal frames, motion data, and so on, we recover a mesh model with the same connectivity as PanoMan template. We follow the variant [Liu et al. 2019] of discrete shell deformation in Fröhlich and Botsch [2011] to establish our framework under the guidance of PanoMan subspace.

5.1 Reconstruction Framework

We model the reconstruction based on PanoMan via the discrete-shell deformation [Fröhlich and Botsch 2011]. The advantage to do so is twofold: It adds excellent extrapolation ability to our approach and makes direct vertex editing possible [Liu et al. 2019]. Specifically, let $V = [v_1, \cdots, v_{|V|}]$ be the vertex vector to be reconstructed. Denote its edge length vector and dihedral angle vector by $X_L(V)$ and $X_A(V)$, respectively. Assume $B_P(\alpha, \beta)$ be the guidance pose of rest pose $B_S(\alpha)$. Accordingly, $X_L(\alpha)$ and $X_A(\alpha)$ are the corresponding edge length vector and dihedral angle vector of $B_S(\alpha)$ while $X_L(\alpha, \beta)$ and $X_A(\alpha, \beta)$ are the same for $B_P(\alpha, \beta)$. Our reconstruction is cast to minimize the total energy of stretching, bending, and constraint parts

$$\mathcal{E}(V, \alpha, \beta) = \mathcal{E}_S(V, \alpha, \beta) + \mathcal{E}_B(V, \alpha, \beta) + \lambda \mathcal{E}_C(V, \alpha, \beta). \quad (6)$$

The combination of stretching and bending terms enforces the deformation of V falling into the subspace of $B_P(\alpha, \beta)$. The former sustains relative edge lengths

$$\mathcal{E}_S(V, \alpha, \beta) = \|[X_L(V) - X_L(\alpha, \beta)] \oslash X_L(\alpha)\|^2, \quad (7)$$

where $\oslash$ denotes the entry-by-entry division. The latter preserves weighted dihedral angles

$$\mathcal{E}_B(V,\alpha,\beta) = \|[X_A(V) - X_A(\alpha,\beta) \odot X_L(\alpha)] \oslash S_A\|^2, \quad (8)$$

where $\odot$ stands for entry-by-entry multiplication, and S_A is an $|E|$-dimensional vector whose entry i is the area of two triangles adjacent to edge e_i.

The third term, i.e., the constraint energy, in Equation (6), varies depending on applications, which will be discussed later.

5.2 Constraint Energies

In this subsection, we present several reconstruction constraints to show the flexibility of PanoMan.

Editing via shape and motion parameters. In this case, we are actually given new blending coefficients $\{\hat{\alpha}_k : k \in H_{SB} \subseteq \{1,2,\cdots,|S|\}\}$, and $\{\hat{\beta}_k : k \in H_{MC} \subseteq \{1,2,\cdots,|M_L|\}\}$. Hence, the constraint energy includes two items:

$$\mathcal{E}_C(V,\alpha,\beta) = \mathcal{E}_{CSP}(\alpha) + \mathcal{E}_{CMP}(\beta), \quad (9)$$

where

$$\begin{cases} \mathcal{E}_{CSP}(\alpha) = \sum\limits_{k\in H_{SB}} (\alpha_k - \hat{\alpha}_k)^2, \\ \mathcal{E}_{CMP}(\beta) = \sum\limits_{k\in H_{MC}} (\beta_k - \hat{\beta}_k)^2. \end{cases} \quad (10)$$

Vertex handles. In the case of using PanoMan to fit sparse feature points or dense scanned data, a corresponding $\hat{v}_k$ is assumed having been found for each vertex v_k in handle set $H_V (k \in H_V \subseteq \{1,2,\cdots,N_V\})$. The vertex constraint is then formulated as:

$$\mathcal{E}_C(V,\alpha,\beta) = \sum_{k\in H_V} (v_k - \hat{v}_k)^2. \quad (11)$$

Joint positions. Joint positions $J = \{J_1,\cdots,J_{|J|}\}$ imply not only motion data but also some shape information such as bone length. We establish the relationship between a set of selected vertices and the specific joint. Therefore, joint position constraints can be converted into vertex constraints.

Motion data. Our model can combine motion data $\theta = \{\theta_1,\cdots,\theta_{3|J|}\}$ and a specific shape $B_S(\alpha)$ to produce new poses. Taking use of the labeling of our dataset, we regress function $\beta = \beta(\theta)$ by learning an MLP, as shown in Figure 9. This converts the issue of motion data constraint to solving Equation (9).

5.3 Numerical Solution

Compared to the formulations in Fröhlich and Botsch [2011] and Liu et al. [2019], our setting is more complex due to involving two sets of blending parameters as well as a nonlinear MLP. Fortunately, the numerical method used there still works, i.e., we employ the Gauss-Newton method to solve the minimization of Equation (6). Considering that facial expressions and hand gestures are much smaller than those of body poses in scale, we adopt a block coordinate descent strategy to reconstruct facial expression first, and then solve for body pose and hand gesture one after the other.

A good initialization is critical for our task, especially when the subject is ongoing a substantial motion. Fortunately, if being able to "guess" approximate values for α and β, we can then evaluate the LA vectors (α,β) and $A(\alpha,\beta)$. Vertex vector V can henceforth be initialized by using shape recovering approach g described in Appendix A. In our current implementation, we always perform inverse kinematics with SMPL to obtain initial shape blending weights α and motion data θ and then estimate our blending parameters β using the MLP in Figure 9.

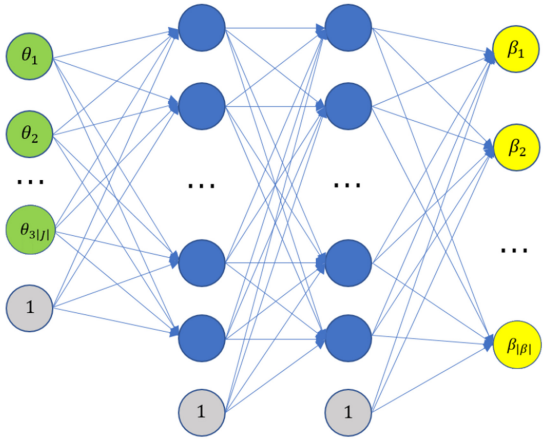

Fig. 9. MLP for regressing parameters β of sparse localized components with respect to motion data θ: The MLP takes motion data as inputs and blending parameters as output; it includes two hidden layers that have 400 and 600 nodes, respectively.

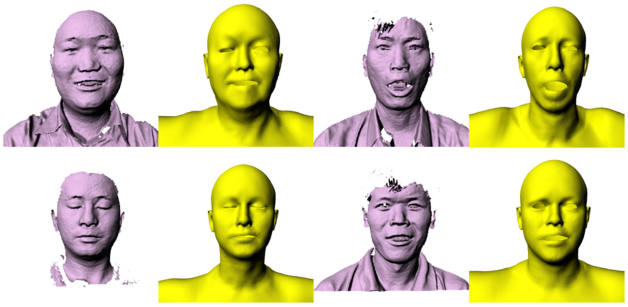

Fig. 10. Fitting PanoMan to the sparse feature points on face point clouds from the CASREAL dataset [Gao et al. 2008] demonstrates that our model can faithfully recover face shapes and expressions from real-world data: Four examples are shown, in which grayish purple images are scanned data and the subsequent yellow ones are fitting results.

6 EVALUATION OF PANOMAN

We evaluate PanoMan from three aspects: ability to represent a variety of motions from three benchmarks, visual and numerical comparison with SMPL-based models (i.e., SMPL, SMPL+H, and SMPL-X), as well as efficiency analysis. All parametric representations are implemented on Visual Studio. As our experimental data are 3D joint positions or feature points, the fitting program for SMPL is adapted from the source codes of SMPLify [Bogo et al. 2016] while those of SMPL+H and SMPL-X are adapted from the source codes of SMPL-X [Pavlakos et al. 2019]. Both programs are originally designed for reconstructing 3D shapes from 2D images.

CASREAL dataset. We first assess PanoMan using single task motions in the CASREAL dataset [Gao et al. 2008]. Figure 10 presents some examples using our model to fit face point clouds of the dataset. Although only 10 shape bases for the whole body are

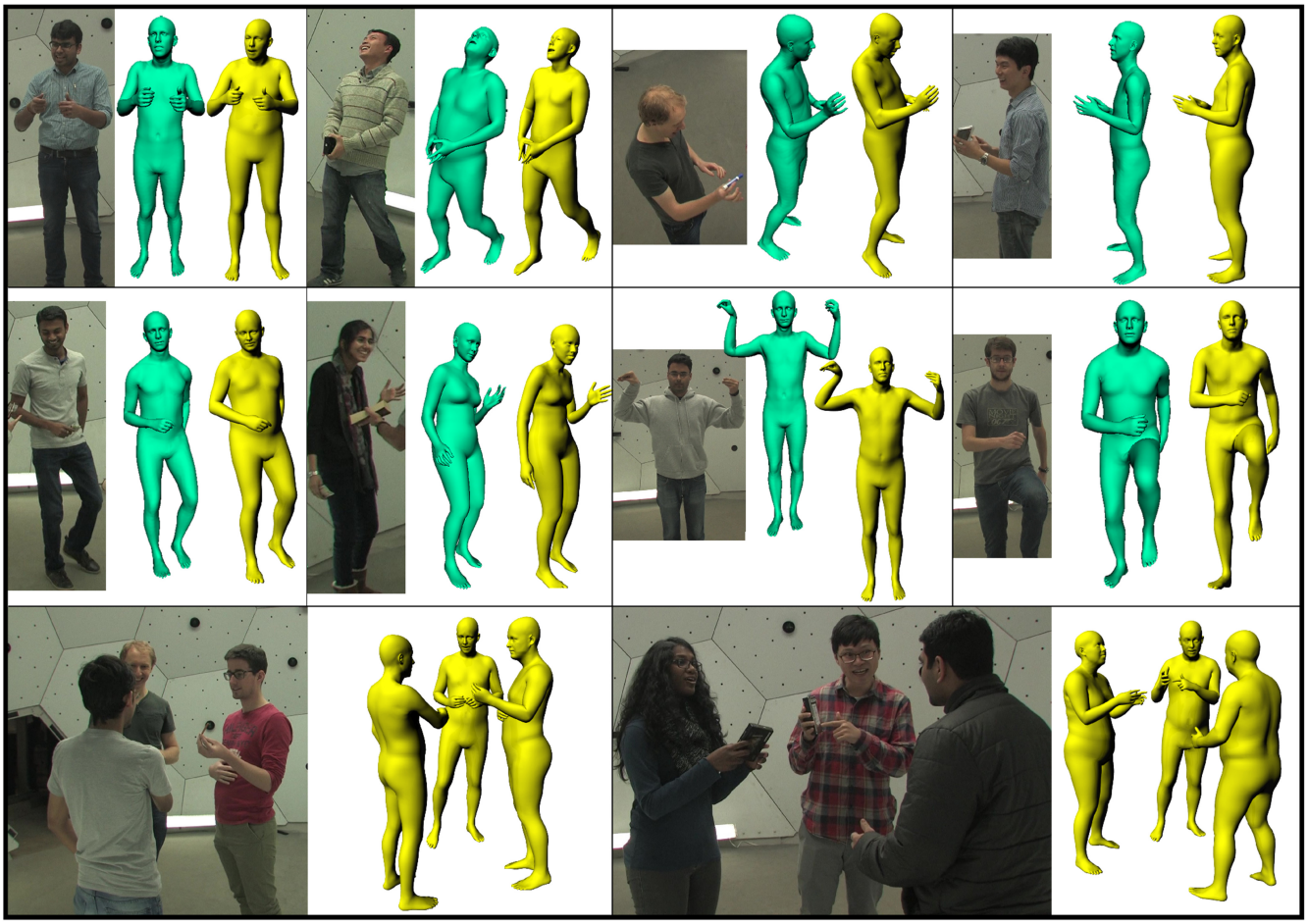

Fig. 11. Reconstruction results on various full motion scenes from the Totalcapture dataset. For each example, the captured scene, the result by SMPL-X [Pavlakos et al. 2019] (except for the bottom row), and ours are depicted in purple, green, and yellow, respectively.

Fig. 12. Reconstruction results on various motion data in the SMPL+H dataset. The scanned data (purple), the result by SMPL+H (green) [Romero et al. 2017], and our result (yellow) are illustrated in each example. In the second example, both methods are even capable of rebuilding crossed fingers.

used, PanoMan has fairly strong ability in recovering face shape and expression.

Totalcapture dataset. This dataset was created by Joo et al. [2018] to capture the 3D motion of multiple people engaged in a social interaction recorded by multiview videos. Besides a point cloud, each identity in a scene is labeled with extra 127 feature points: 15 body joints, 70 facial expression points, and 42 hand gesture joints, among which body skeletons are obtained by the method described in Joo et al. [2019] while face and hand features are estimated by using the method of Simon et al. [2017]. We use the 127 feature points as well as manually labeled 10 points near the torso region in our fitting. The latter plays the role to roughly control the body shape. Figure 1 (the first row) and Figure 11 demonstrate that PanoMan is able to yield visually pleasing results that quite faithfully approximate body poses, facial expressions, as well as hand gestures of the corresponding scenes. As comparison,

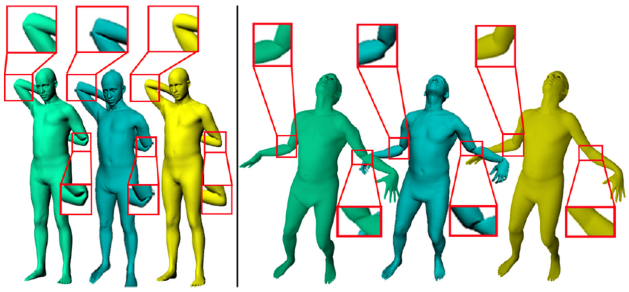

Fig. 13. Two examples with large twisting rotation around both elbow joints: Both SMPL (green) [Loper et al. 2015] and SMPL-X (gray) [Pavlakos et al. 2019] exhibit serious joint collapse, while PanoMan (yellow) works well.

we also use SMPL-X [Pavlakos et al. 2019] to generate the corresponding poses and depict them in the figure. The results show that both methods are visually comparable for these conventional poses.

SMPL+H dataset. This dataset includes scans across a variety of body poses and hand gestures. Besides point cloud data, it also labels 24 joint positions, 42 hand joint positions, and 10 facial feature points [Romero et al. 2017]. Similarly, we still fit PanoMan and SMPL+H to the sparse feature points as constraints, respectively. Figure 12 demonstrates that PanoMan is competent of recovering examples with large deformation near the waist region and comparable to SMPL+H in recovering these hand gestures with moderate degree of deformation.

Poses with large twisting. We mentioned in Section 1 that skeletal skinning–based models like SMPL usually suffer from joint collapse for data with large twisting rotation near joints and presented two examples to show this in Figure 1. Since they are built upon SMPL [Loper et al. 2015], SMPL+H [Romero et al. 2017], Frankenstein [Joo et al. 2018], and SMPL-X [Pavlakos et al. 2019] also suffer from this shortcoming. We augment two examples by using the motion data from the CMU database [CMU 2000] to illustrate this, as shown in Figure 13, in which results by SMPL, SMPL-X, and PanoMan are shown in succession. Two more examples are presented in Figure 14, where the original data (left of the figure) is scanned by ourselves and the same set of manually selected feature points are used for all parametric models. All these examples except for the last one demonstrate that results generated by SMPL and SMPL-X exhibit artifacts in elbow and arm joint regions while our results undergo a more natural deformation.

Unusual hand gesture evaluation. Examples in Figure 11 have shown that PanoMan is capable of expressing diverse hand gestures. Figure 15 illustrates two more examples with unusual hand poses that fall outside the low-dimensional pose space of SMPL+H and SMPL-X according to Romero et al. [2017], while our PanoMan still produces good gestures (yellow). It should be mentioned that extra 39 and 44 scan points are specified as vertex constraints for our first (yellow in top row) and second example (yellow in bottom row), respectively, besides 24 joint positions (only including 1 hand joint) and 10 facial feature points given in the SMPL+H dataset [Romero et al. 2017].

Motion composition. We apply PanoMan to composite body motion, facial expression, and hand gesture from different data sources into one full motion pose, as shown in Figure 16, which manifests that PanoMan is capable of integrating different motion data together naturally. In the figure, the body poses, respectively, come from Bogo et al. [2016] (top row) and from SMPL+H [Romero et al. 2017] (bottom row). The facial expression and hand gesture of all examples are taken from Totalcapture [Joo et al. 2018].

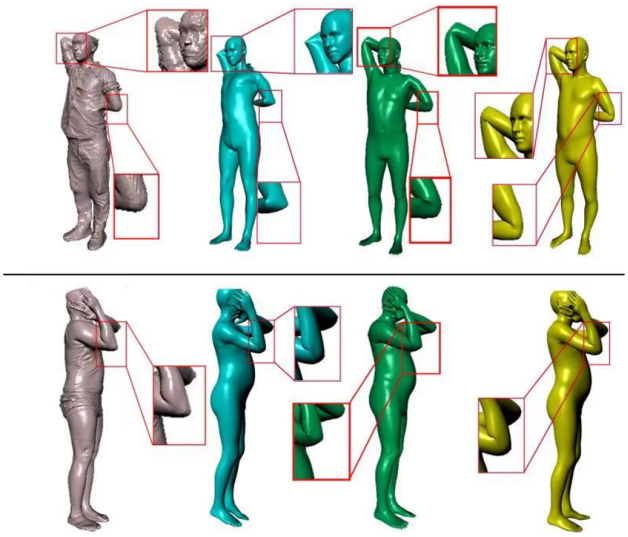

Fig. 14. Performing SMPL, SMPL-X, and PanoMan on our scanned data (left): SMPL (middle-left) [Loper et al. 2015] leads to serious distortion near elbow joints; SMPL-X (middle-right) [Pavlakos et al. 2019] is relatively better but with arm pose under-fitted; results by PanoMan (right) are the most natural.

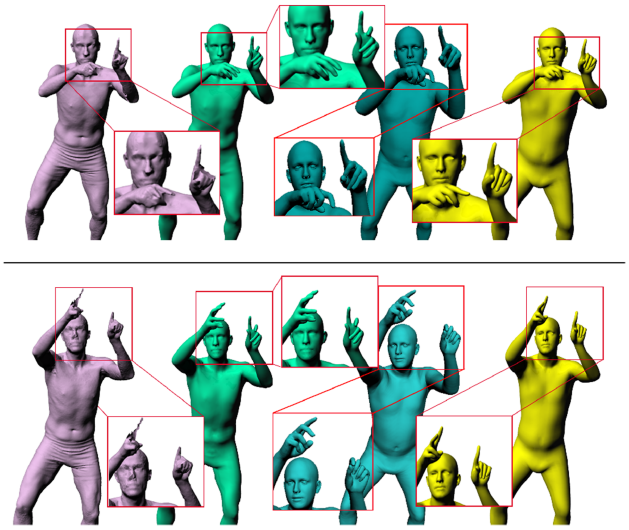

Fig. 15. Two examples with unusual hand poses (see scans in purple). From left to right are, respectively, scan, and results of SMPL+H, SMPL-X, PanoMan. SMPL+H and SMPL-X fail to generate correct gestures, while PanoMan performs well. Note that results by SMPL+H were generated by the authors of SMPL+H and 20 shape bases are used, while 10 shape bases are employed to approximate the shape of SMPL-X and PanoMan.

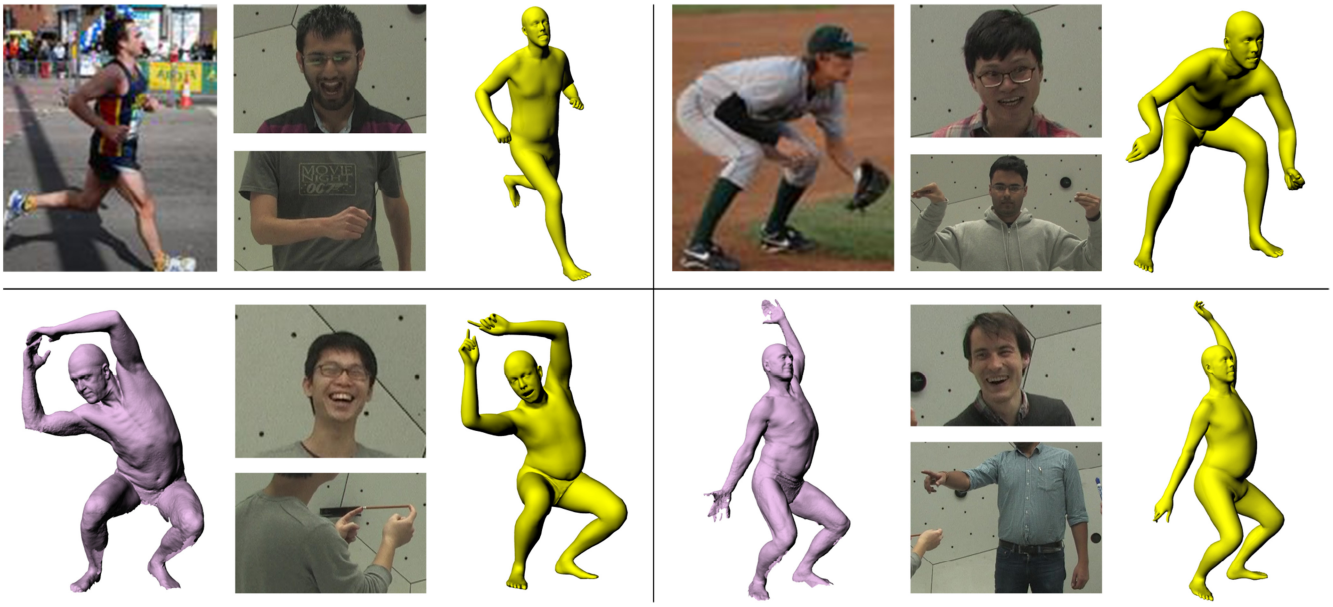

Fig. 16. Composite body pose, facial expression, and hand gesture into one and four examples are given: In each example, the left image is associated with body motion data, the middle two show the scenes embody the facial expression and hand gesture parameters, respectively, and the right one is our composite result.

Motion sequence reconstruction. To show the robustness of PanoMan in pose capture, we apply it to reconstruct four clips of animation from sequences "body-Hands A08," "body-Hands B08," "bodyHands_B12," and "bodyHands_B24" of the SMPL+H dataset [Romero et al. 2017]. Here, we fit the clips in a frame-by-frame manner. Fifteen body joints, 70 facial feature points, and 42 finger joints are used for fitting each frame of the Totalcapture data, while 24 body joints, 10 upper body points, 10 facial feature points, 97 hand surface points, and 8 foot feature points are used for fitting each frame of the SMPL+H sequences. Figure 17 illustrates the fitting result of SMPL+H sequence "bodyHands_B24."

As comparison, we also depict the results by SMPL-X for the Totalcapture clip and by SMPL+H for the SMPL+H clip. It can be seen that our results for these conventional poses are visually comparable to those by the state-of-the-art models. We refer the reader to the accompanied videos for full fitting sequences of the four clips.

Ablation on the MLP correction & qualitative comparison. First, we make a brief ablation on our model with and without the MLP corrected term by making use of FAUST and DFAUST datasets to evaluate their reconstruction accuracy. The latter one is called the baseline. As comparison, the reconstruction accuracy of SMPL and SMPL+H are also evaluated. FAUST contains total 100 poses with 10 poses for each of the 5 males and 5 females. DFAUST contains 5 (males) + 5 (females) = 10 motion sequences. Each contains 4,000 frames. We randomly select in total 50 frames of males in FAUST and 450 frames of males in DFAUST as our test set. Among 450 frames of the DFAUST dataset, 90 frames are randomly chosen for each single male. Figure 18 demonstrates that PANOMAN improves the fitting accuracy compared to the baseline by about 1 millimeter or 30%. Even so, the fitting accuracy of the baseline is still slightly higher than that of SMPL and SMPL+H.

Next, we compare the reconstruction error of SMPL-X [Pavlakos et al. 2019], SMPL+H [Romero et al. 2017], and PanoMan on Totalcapture sequences and SMPL+H sequences. Considering that only 3D feature points are known for the Totalcapture animations and that body, face, and hands are human parts with completely different scale, we analyze the average fitting error of feature points on each part independently. Full 3D meshes are given in the SMPL+H sequences, therefore, we estimate the maximal and average error of all vertices of the reconstructed mesh to the surface of the ground truth by using the Hausdorff distance function of Meshlab. All evaluations are conducted frame-by-frame.

Figure 19 illustrates the reconstruction errors of feature points for the two clips of the Totalcapture sequence by SMPL-X and PanoMan, respectively. In the figure, errors of SMPL-X and PanoMan are separately drawn with dotted and solid curves, while green, blue, and yellow indicate errors for body, hand, and face, respectively. Similarly, we also fit the parametric models to two sequences from the SMPL+H dataset. Figure 20 respectively depicts the error curves of SMPL (green), SMPL+H (pink), and XMPL-X (yellow) and PanoMan (blue). Both figures demonstrate that our approach achieves higher accuracy than SMPL-X and SMPL+H on the whole.

To observe the spatial error distribution on a frame, we select the 50th frame in sequence "bodyHands_B24" and visualize the Hausdorff distance between each vertex of its reconstructed pose and its ground truth mesh in Figure 21. It demonstrates that our result exhibits smaller fitting error in joint regions than the SMPL+H result.

Timings. To assess the time performance of reconstruction based on PanoMan, we record the time of generating each result in Figure 11 on a PC computer with I5 8300H CPU+16 G memory and GTX 1060 GPU+6 G video memory. We divide the algo-

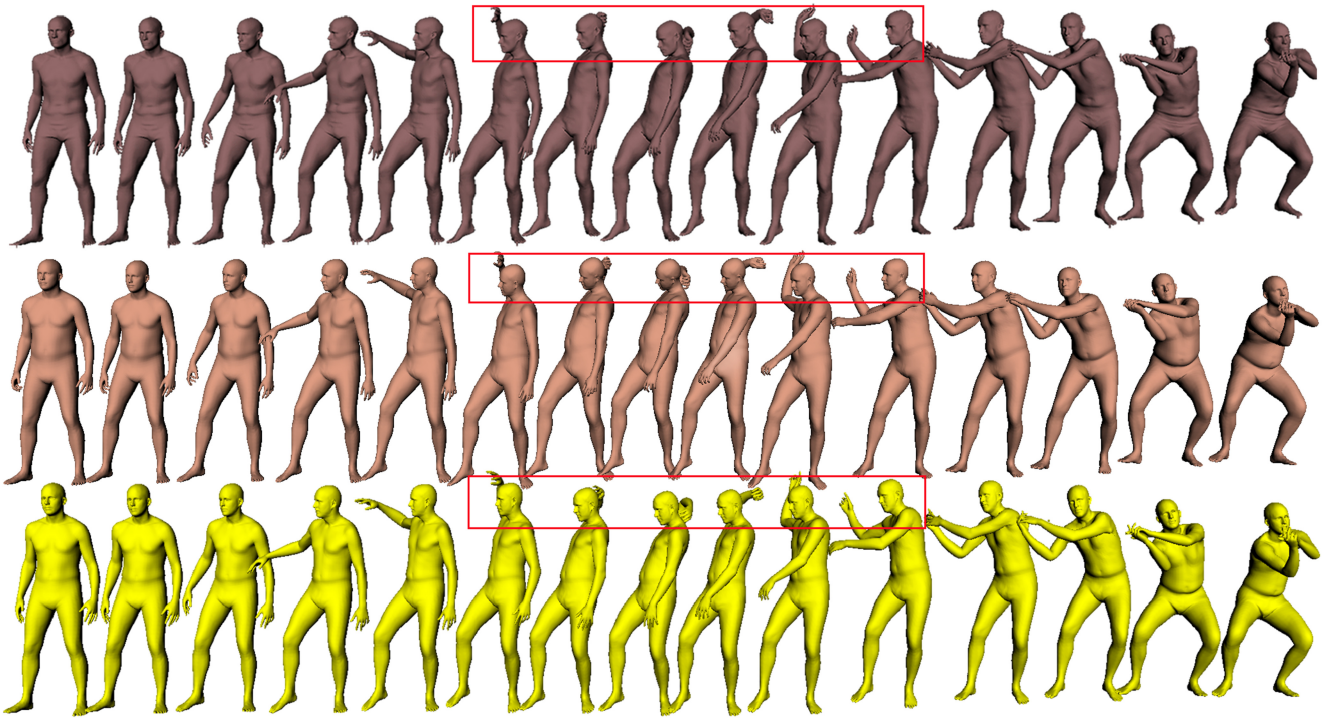

Fig. 17. Reconstruction of a SMPL+H motion sequence. The top row shows the ground truth, and the middle and bottom rows are, respectively, frames reconstructed by SMPL+H and PanoMan. Our results in the rectangular regions look better than those by SMPL+H.

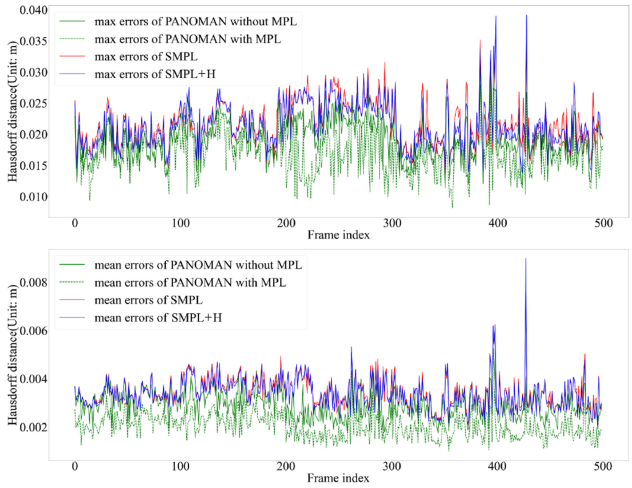

Fig. 18. Fitting errors (Top: max error; Bottom: average error) of the baseline, PANOMAN, SMPL, and SMPL+H: PANOMAN shows higher accuracy than other three models.

rithm into five stages. They are estimating motion data θ from joint positions, calculating β using MLP $\beta(\alpha, \theta)$, reconstructing initial shape with g, and fitting feature points of three parts. We summarize the average, maximal, and minimal time-consuming of all these stages in Table 1. The numbers listed in parentheses indicate the corresponding iterations, which are determined automatically by the algorithm. For comparison, the total runtimes of SMPL+H and SMPL-X are, respectively, 14.1 (ave), 19.0 (max), 11.42 (min), and 99.54 (ave), 114.0 (max), 85.47 (min). As PanoMan includes hundreds of parameters, it consumes much more time for reconstruction than SMPL+H and SMPL-X.

Table 1. Timings for the Reconstruction Algorithm (Unit: s)

Stage	IK	g	face	body	hand	Total
Ave.	47.2	1.9	32.2(13)	166.0(14)	27.2(25)	274.5
max	57.3	2.3	39.6(16)	295.2(23)	30.7(26)	425.1
min	32.7	1.7	26.3(11)	16.1(2)	20.2(21)	97.0

7 CONCLUSIONS

PanoMan is proposed to parameterize body poses, facial expressions, and hand gestures in a unified framework. It describes a specific full shape by blending the shape PCA bases and simulates full motion variations from body, face, and hand parts using a set of sparse localized components extracted from RLA vectors of the training data. We show that it is feasible to synthesize our training data from existing single-task datasets and not necessary to include data with full body, face, and hand motions in the training data, thanks to the locality of our motion bases. A robust algorithm is proposed to fit PanoMan to 3D human dynamic geometries with a variety of constraints. Multilayer perceptrons are employed to regress the residual of our parametric model and estimate initial values of blending motion coefficients when applying the model to recover human dynamic geometries. Experiments show that PanoMan is capable of both representing full human motions and producing more pleasing poses than state-of-the-art methods, which are mainly based on SMPL, in dealing with body

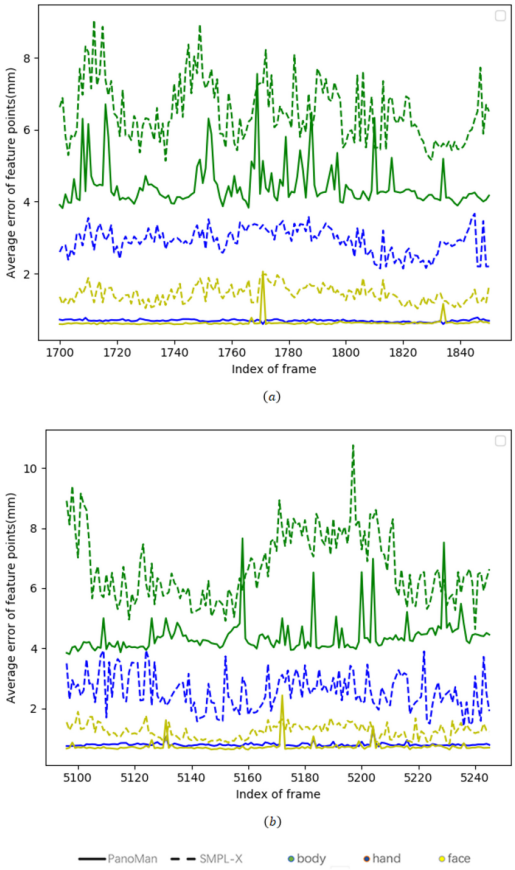

Fig. 19. Reconstruction accuracy of SMPL-X (dotted curves) and PanoMan (solid curves) for two clips extracted from "sequence 171204_pose3" of Totalcapture [Joo et al. 2018]. Y-axis is the fitting error (unit: millimeter) of feature points, and X-axis indicates the serial number of frames in the original sequence. Green, blue, and yellow are related to body, hand, and face parts separately.

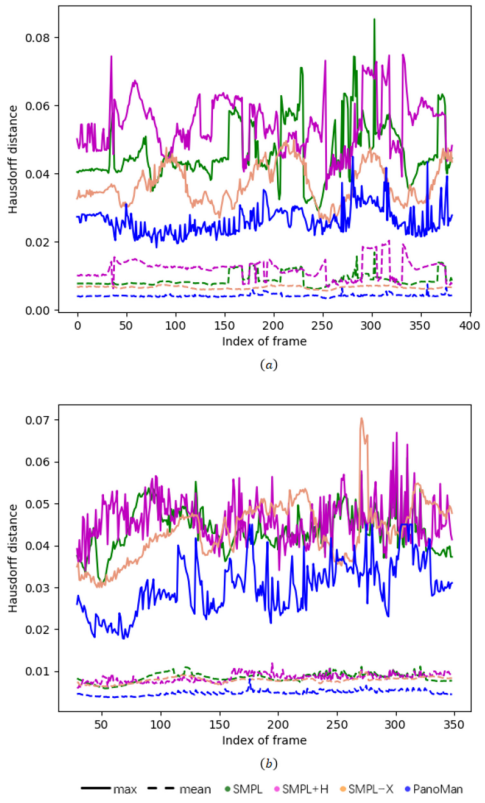

Fig. 20. Reconstruction accuracy of SMPL, SMPL+H, SMPL-X, and PanoMan for two clips of motion, respectively, extracted from sequences "bodyHands_B12" (top) and "bodyHands_B24" (bottom) of the SMPL+H dataset [Romero et al. 2017]. Y-axis records the Hausdorff distance between reconstructed frames and the corresponding ground truth, which has been divided by the diagonal length of the bounding box of the object, and X-axis indicates the frame indices. Solid and dotted lines stand for maximal and mean errors separately. Note that we sample new vertices on SMPL+H reconstructions to keep their topology the same as ours.

motions with large twisting rotation and unusual hand gestures. In addition, numerical analysis manifests that PanoMan is able to achieve higher reconstruction accuracy in general than the state-of-the-art approaches such as SMPL+H and SMPL-X.

As future work, accelerating the reconstruction efficiency is in an important position for our model. There are two ways to achieve this goal. As we have shown that exploring deep network regression among different kinds of parameters is probably promising due to the data-driven essence of our method. A point-cut is to establish a deep neural network to approximate the initial shape for given shape and pose parameters in this avenue. As the iteration algorithm is easily parallelized, GPU acceleration is a feasible option.

Thoroughly assessing the model is also significant. For example, we may also perform localized PCA analysis on the shape subset to increase the flexibility of fitting shapes. After all, dependency between face, hand, and body shapes is relatively low. In our current implementation, only 10 shape bases are available. Enhancing the training dataset to get more bases is necessary for practical purpose. In addition, totally 700 sparse localized components are generated, and it is useful to examine how many bases are adequate for capturing full motions from real data.

In our experiments, we mainly apply the model to fit sparse feature points or motion data of 3D scans. It is possible to develop robust algorithms to register PanoMan with dense point clouds automatically. Recovering 3D poses from a single image or a video is also an important application for a successful parametric model. We also plan to integrate vibration modes of Brandt and Hildebrandt [2017] to our model to simulate physical properties of human body and to include more human elements such as eyes, feet, hair, and clothes into the framework.

APPENDIX A g: SHAPE RECONSTRUCTION FROM RLA VECTORS

Given reference mesh $M_0 = (V_0, E, F)$ and the RLA vector of $M = (V, E, F)$ with respect to M_0, reconstructing the shape of M is straightforward. Obviously, Equation (1) directly leads to estimate of the edge lengths and dihedral angles of M:

$$a(e) = a_0(e) + \hat{a}(e), l(e) = l_0(e)(1 + \hat{l}(e)).$$

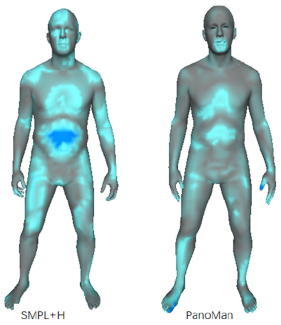

Fig. 21. Spatial error distribution of the reconstructed pose of frame 50 in "bodyHands_B24" by SMPL+H (left) and PanoMan (right). The blue color indicates large error.

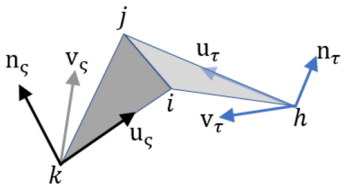

Fig. 22. Local frame reconstruction for computing the connection map of edge $e = (i, j)$ from five edge lengths and one dihedral angle.

The rest of the reconstruction process is then built upon the algorithm in Wang et al. [2017].

Define an orthogonal frame $R_t = (\mathbf{u}_t, \mathbf{v}_t, \mathbf{n}_t)$ for each triangle $t \in F$, where $\mathbf{u}_t$ and $\mathbf{n}_t$ are accordant with its first edge vector and outward normal, respectively. For each $e = (i, j) \in E$, assume $\zeta = (k, i, j)$ and $\tau = (h, j, i)$ are its two adjacent triangles (see Figure 22) in which their orthogonal frames R_ζ and R_τ are separately originated at vertices k and h. We further introduce the connection map of e as

$$Q_e = R_\zeta^{-1} R_\tau .$$

It is easy to show that Q_e is a 3×3 matrix irrelevant to coordinate-systems and can be estimated from edge lengths and dihedral angles.

For every $e = (i, j) \in E$, we establish a coordinate system $(\mathbf{u}_\zeta, \mathbf{v}_\zeta, \mathbf{n}_\zeta)$ that locates at vertex i and takes (i, j) as u axis and the outward normal of t_1 as n axis. Obviously, we have $\tilde{R}_\zeta = I$ (the orthogonal frame of ζ) and $\tilde{R}_\tau = (\tilde{\mathbf{u}}_\tau, \tilde{\mathbf{v}}_\tau, \tilde{\mathbf{u}}_\tau)$ (the orthogonal frame of τ) with respect to the above coordinate system. It follows $Q_e = (\tilde{R}_\zeta)^{-1} \tilde{R}_\tau = \tilde{R}_\tau$. Solving the following constrained optimization

$$arg \min_{\{R_t\}} \sum_{e \in E} \|R_\zeta - Q_e R_\tau\|^2, \text{s.t. } R_t^T R_t = I \text{ for all } t \in F.$$

with R_1 given yields all orthogonal frames $\{R_t, t \in F\}$ from which we easily obtain all edge vectors $\{\mathbf{e} : e \in E\}$ of M. Finally, minimizing

$$\sum_{e=(i,j) \in E} \|(v_i - v_j) - \mathbf{e}\|^2$$

via fixing v_1 yields vertex positions of the mesh.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their valuable comments. Thanks also go to Federica Bogo and Georgios Pavlakos for sharing source codes to reconstruct 3D human shapes with SMPL and SMPL+H/-X, respectively, and Loper et al. [2015], Romero et al. [2017], Gao et al. [2008], Cao et al. [2014], Joo et al. [2018], Varol et al. [2017], Sigal et al. [2010], and Bogo et al. [2017] for opening up their datasets.

REFERENCES

Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. 2019a. Learning to reconstruct people in clothing from a single RGB camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1175–1186.

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video based reconstruction of 3D people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8387–8397.

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019b. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*. 2293–2303.

Brett Allen, Brian Curless, and Zoran Popović. 2003. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Trans. Graph.* 22, 3 (July 2003), 587–594.

Brian Amberg, Reinhard Knothe, and Thomas Vetter. 2008. Expression invariant 3D face recognition with a morphable model. In *Proceedings of the 8th IEEE International Conference on Automatic Face & Gesture Recognition (FG'08)*. IEEE, 1–6.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. SCAPE: Shape completion and animation of people. *ACM Trans. Graph.* 24, 3 (July 2005), 408–416.

Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. 2012. Motion capture of hands in action using discriminative salient points. In *Proceedings of the European Conference on Computer Vision*. Springer, 640–653.

Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2020. Combining implicit function learning and parametric models for 3D human reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*. Springer, 311–329.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Conference on Computer Graphics and Interactive Techniques*. ACM Press/Addison-Wesley Publishing Co., 187–194.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision*. Springer, 561–578.

Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2017. Dynamic FAUST: Registering human bodies in motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5573–5582.

Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2013. Online modeling for realtime facial animation. *ACM Trans. Graph.* 32, 4 (2013), 76–86.

Christopher Brandt and Klaus Hildebrandt. 2017. Compressed vibration modes of elastic bodies. *Comput-aided Geom. Des.* 52 (2017), 297–312.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graph.* 20, 3 (2014), 413–425.

Yin Chen, Zhi-Quan Cheng, Chao Lai, Ralph R. Martin, and Gang Dang. 2016. Realtime reconstruction of an animating human body from a single depth camera. *IEEE Trans. Vis. Comput. Graph.* 22, 8 (2016), 2000–2011.

Yinpeng Chen, Zicheng Liu, and Zhengyou Zhang. 2013. Tensor-based human body modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 105–112.

Yin Chen, Zhan Song, Weiwei Xu, Ralph R. Martin, and Zhi-Quan Cheng. 2019. Parametric 3D modeling of a symmetric human body. *Comput. Graph.* 81 (2019), 52–60.

Ke-Li Cheng, Ruo-Feng Tong, Min Tang, Jing-Ye Qian, and Michel Sarkis. 2016. Parametric human body reconstruction based on sparse key points. *IEEE Trans. Vis. Comput. Graph.* 22, 11 (2016), 2467–2479.

Zhi-Quan Cheng, Yin Chen, Ralph R. Martin, Tong Wu, and Zhan Song. 2018. Parametric modeling of 3D human body shape—A survey. *Comput. Graph.* 71 (2018), 88–100.

Baptiste Chu, Sami Romdhani, and Liming Chen. 2014. 3D-aided face recognition robust to expression and pose variations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1899–1906.

CMU. 2000. CMU graphics lab motion capture database. Retrieved from: http://mocap.cs.cmu.edu.

Stefan Fröhlich and Mario Botsch. 2011. Example-driven deformations based on discrete shells. *Comput. Graph. Forum* 30, 8 (2011), 2246–2257.

Lin Gao, Shu-Yu Chen, Yu-Kun Lai, and Shihong Xia. 2017. Data-driven shape interpolation and morphing editing. *Comput. Graph. Forum* 36, 8 (2017), 19–31.

Lin Gao, Yu-Kun Lai, Dun Liang, Shu-Yu Chen, and Shihong Xia. 2016. Efficient and flexible deformation representation for data-driven surface modeling. *ACM Trans. Graph.* 35, 5 (2016), 1–17.

Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. 2008. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Trans. Syst., Man, Cyber.-Part A: Syst. Hum.* 38, 1 (2008), 149–161.

Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Trans. Graph.* 35, 3 (2016), 28.

Behrend Heeren, Chao Zhang, Martin Rumpf, and William Smith. 2018. Principal geodesic analysis in the space of discrete shells. *Comput. Graph. Forum* 37, 5 (2018), 173–184.

David A. Hirshberg, Matthew Loper, Eric Rachlin, and Michael J. Black. 2012. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In *Proceedings of the European Conference on Computer Vision*. Springer, 242–255.

Zhichao Huang, Junfeng Yao, Zichun Zhong, Yang Liu, and Xiaohu Guo. 2014. Sparse localized decomposition of deformation gradients. *Comput. Graph. Forum* 33, 7 (2014), 239–248.

Alec Jacobson and Olga Sorkine. 2011. Stretchable and twistable bones for skeletal shape deformation. *ACM Trans. Graph.* 30, 6 (2011), 165.

Arjun Jain, Thorsten Thormählen, Hans-Peter Seidel, and Christian Theobalt. 2010. MovieReshape: Tracking and reshaping of humans in videos. *ACM Trans. Graph.* 29, 6 (2010), 148.

Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2019. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 1 (2019), 190–204.

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8320–8329.

Petr Kadleček, Alexandru-Eugen Ichim, Tiantian Liu, Jaroslav Křivánek, and Ladislav Kavan. 2016. Reconstructing personalized anatomical models for physics-based body animation. *ACM Trans. Graph.* 35, 6 (2016), 213.

Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Seungbae Bang, Jinwook Kim, Michael J. Black, and Sung-Hee Lee. 2017. Data-driven physics for human soft tissue animation. *ACM Trans. Graph.* 36, 4 (2017), 54.

YoungBeom Kim and JungHyun Han. 2014. Bulging-free dual quaternion skinning. *Comput. Anim. Virt. Worlds* 25, 3-4 (2014), 321–329.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. 2017. Unite the people: Closing the loop between 3D and 2D human representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*, Vol. 2. 3.

Binh Huy Le and Jessica K. Hodgins. 2016. Real-time skeletal skinning with optimized centers of rotation. *ACM Trans. Graph.* 35, 4 (2016), 37.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194.

Yuxin Liu, Guiqing Li, Yupan Wang, Yongwei Nie, and Aihua Mao. 2019. Discrete shell deformation driven by adaptive sparse localized components. *Comput. Graph.* 78 (2019), 76–86.

Matthew Loper, Naureen Mahmood, and Michael J. Black. 2014. MoSh: Motion and shape capture from sparse markers. *ACM Trans. Graph.* 33, 6 (2014), 220.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* 34, 6 (2015), 248.

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. GANerated hands for real-time 3D hand tracking from monocular RGB. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–59.

Thomas Neumann, Kiran Varanasi, Stephan Wenger, Markus Wacker, Marcus Magnor, and Christian Theobalt. 2013. Sparse localized deformation components. *ACM Trans. Graph.* 32, 6 (2013), 179.

Verónica Orvalho, Pedro Bastos, Frederic I. Parke, Bruno Oliveira, and Xenxo Alvarez. 2012. A facial rigging survey. In *Eurographics (STARs)*. 183–204.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10975–10985.

Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 459–468.

Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. 2017. Building statistical shape spaces for 3D human modeling. *Pattern Recog.* 67 (2017), 276–286.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. 2017. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Trans. Graph.* 36, 4 (2017), 73.

Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J. Black. 2015. Dyna: A model of dynamic human shape in motion. *ACM Trans. Graph.* 34, 4 (2015), 120.

Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. 2014. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1106–1113.

Elad Richardson, Matan Sela, and Ron Kimmel. 2016. 3D face reconstruction by learning from synthetic data. In *Proceedings of the 4th IEEEE International Conference on 3D Vision (3DV'16)*. IEEE, 460–469.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.* 36, 6 (2017), 245.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2020. FrankMocap: Fast monocular 3D hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324* (2020).

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. 2019. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7763–7772.

Josua Sassenm, Klaus Hildebrandt, and Martin Rumpf. 2020. Nonlinear deformation synthesis via sparse principal geodesic analysis. *Comput. Graph. Forum* 39, 5 (2020), 119–132.

Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei et al. 2015. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd ACM Conference on Human Factors in Computing Systems*. ACM, 3633–3642.

Leonid Sigal, Alexandru O. Balan, and Michael J. Black. 2010. HUMANEVA: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Int. J. Comput. Vis.* 87 (2010), 4–27.

Tomas Simon, Hanbyul Joo, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 4645–4653.

Dan Song, Ruofeng Tong, Jian Chang, Xiaosong Yang, Min Tang, and Jian Jun Zhang. 2016. 3D body shapes estimation from dressed-human silhouettes. *Comput. Graph. Forum* 35, 7 (2016), 147–156.

Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. 2015. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3213–3221.

Robert W. Sumner and Jovan Popović. 2004. Deformation transfer for triangle meshes. *ACM Trans. Graph.* 23, 3 (2004), 399–405.

Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. 2015. Cascaded hand pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 824–832.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020. GRAB: A dataset of whole-body human grasping of objects. In *Proceedings of the European Conference on Computer Vision*, Vol. 12349. Springer, Cham, 581–600.

Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. 2018a. Variational autoencoders for deforming 3D mesh models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5841–5850.

Qingyang Tan, Lin Gao, Yu-Kun Lai, Jie Yang, and Shihong Xia. 2018b. Mesh-based autoencoders for localized deformation component analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2452–2459.

Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff et al. 2016. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Trans. Graph.* 35, 4 (2016), 143.

Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. 2016. Sphere-meshes for real-time hand modeling and tracking. *ACM Trans. Graph.* 35, 6 (2016), 222.

Anastasia Tkach, Andrea Tagliasacchi, Edoardo Remelli, Mark Pauly, and Andrew Fitzgibbon. 2017. Online generative model personalization for hand tracking. *ACM Trans. Graph.* 36, 6 (2017), 243.

Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2017. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. IEEE, 4627–4635.

Yupan Wang, Guiqing Li, Zhichao Zeng, and Huayun He. 2017. Articulated-motion-aware sparse localized decomposition. *Comput. Graph. Forum* 36, 8 (2017), 247–259.

Ofir Weber, Olga Sorkine, Yaron Lipman, and Craig Gotsman. 2007. Context-aware skeletal shape deformation. *Comput. Graph. Forum* 26, 3 (2007), 265–274.

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM Trans. Graph.* 30, 4 (2011), 77.

Lance Williams. 1990. Performance-driven facial animation. *ACM SIGGRAPH Comput. Graph.* 24, 4 (1990), 235–242.

Tim Winkler, Jens Drieseberg, Marc Alexa, and Kai Hormann. 2010. Multi-scale geometry interpolation. *Comput. Graph. Forum* 29, 2 (2010), 309–318.

Hongyi Xu and Jernej Barbič. 2016. Pose-space subspace dynamics. *ACM Trans. Graph.* 35, 4 (2016), 1–14.

Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. 2018. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7287–7296.

Chao Zhang, Sergi Pujades, Michael J. Black, and Gerard Pons-Moll. 2017. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. 3.

Xiong Zhang, Qiang Li, Wenbo Zhang, and Wen Zheng. 2019. End-to-end hand mesh recovery from a monocular RGB image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, 2354–2364.

Jiaxiang Zheng, Ming Zeng, Xuan Cheng, and Xinguo Liu. 2014. SCAPE-based human performance reconstruction. *Comput. Graph.* 38 (2014), 191–198.

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2020. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. arxiv:*cs.CV/2007.03858* (2020)

Denis Zorin, Peter Schröder, and Wim Sweldens. 1996. Interpolating subdivision for meshes with arbitrary topology. In *Proceedings of the 23rd Conference on Computer Graphics and Interactive Techniques*. ACM, 189–192.

Silvia Zuffi and Michael J. Black. 2015. The stitched puppet: A graphical model of 3D human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3537–3546.

Received November 2019; revised October 2020; accepted January 2021