Multiview Deformation for Dynamic Human Modeling

ANONYMOUS AUTHOR(S)*

We present a novel multi-view dynamic 3D human reconstruction technique based on model-based shape deformation. Our approach specifically targets at handling challenging cases such as textureless appearance, heavy occlusions, and depth order ambiguity that are problematic to stereo-based techniques. We propose to pose match and shape deform a human template model to avoid meshing the point cloud. To robustly match the template pose with image observations, we present a novel Graph Convolutional Networks (GCN) to gradually filter out erroneous views and impose appropriate weights on the optimal subset for recovering the 3D skeleton and warping the template shape. Next, We use the warped human template to guide the cross-view consistent semantic segmentation. We set out to deform the warped 3D model so that the silhouette of the deformed model best matches the target in respective views while maintaining semantic consistency. Comprehensive experiments on publicly available and our newly generated complex motion datasets show our approach significantly outperforms the state-of-the-art on sparse cameras, textureless regions (e.g., under black clothing), complex motions, etc.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; **Image segmentation**; **Shape representations**.

Additional Key Words and Phrases: 3D human modeling, pose estimation, semantic driven shape deformation, semantic labeling

ACM Reference Format:
Anonymous Author(s). 2018. Multiview Deformation for Dynamic Human Modeling. *ACM Trans. Graph.* 37, 4, Article 111 (August 2018), 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The three-dimensional human model has a wide range of applications and has received more and more attention. Typical examples include digital entertainment, distance education, visual special effects, virtual clothes fitting and biomedical. However, the human body has an extremely high degree of freedom (DoF) of movement, and each body part can exhibit distinct non-rigid motion, hence the human reconstruction quality has long been unsatisfactory.

Among the existing methods, the vast majority of the dynamic human reconstruction systems adopt multi-view images based approaches [Cheung et al. 2003; Dou et al. 2016; Joo et al. 2018; Mikhnevich and Hébert 2011]. A common practice is to build a suitable "Dome", i.e. multi-camera array system [Collet et al. 2015a; Joo et al. 2018; Yang and Liu 2009]. Intuitively more cameras and better controlled lighting environment can produce better human geometry [Kanade et al. 1997], but such systems are expensive and difficult to generalize. Also, a small calibration error can lead to large inaccuracies in fine objects, such as the fingers and toes. Furthermore, this technique relies on feature matching results across multiple viewing positions, which cannot handle the textureless regions (even on bare skin). No need to mention that the complex and varied movements of the human body generate enormous self occlusions, which cannot be reconstructed at all. These errors are conducive to subsequential surface reconstructions, creating adhesions or noise, making the final results difficult to meet the needs.

The MVS-based human reconstruction methods did not take the prior topological knowledge of the human body into count, e.g. hands are connected to arms and head is above the shoulder. One way to solve the above problems is to deform a human body template model w.r.t. the image observations. These methods effectively avoid the surface reconstruction step and preserve the original topology information of the human body and hence avoid sticking. The work of [Jiang et al. 2019; Joo et al. 2018] *et al.* deformed a statistical parameter model (SMPL [Loper et al. 2015]) to match with the target point cloud. But it still cannot cope with the sparse cameras and textureless scenarios. [Gall et al. 2009; Vlasic et al. 2008] avoided point cloud reconstruction by deforming a pre-scanned subject-specific template model according to the silhouettes. It requires the rigged template model for different people, and careful weight adjustment of the rigging. Recently, deep learning based approach [Huang et al. 2018] achieved end-to-end human reconstruction through encoding the human prior and feature into a high-dimensional parameter space. However, this method requires a large amount of training data and is limited to a specific camera configuration.

To address the above problems, we present a method for robust high quality human body reconstruction with sparse cameras and textureless regions. Reconstructing the human body from sparse camera views is very challenging as many body parts are invisible to all cameras. With textureless regions, the problem gets more severe. In order to address this challenge, we propose a novel method of human body reconstruction based on template deformation.

We observed that the key step of deformation is to extract high-fidelity skeleton from the multi-view images. For complex movements, the extraction of the accurate skeleton is extremely challenging. We observe that: 1) not all views are equally beneficial for skeleton estimation, and the skeleton detection in some views is more robust [Kanade et al. 1997]. 2) The human skeleton points are coherent and constrained. For example, when the left arm is confidently estimated, the left shoulder estimation should be more confident. Therefore, we propose a graph convolutional network (GCN) to gradually filter out error-prone views and assign weights to inlier views through view selection and attention propagation. Experiments show that the process greatly improves the quality of 3D skeleton extraction and template model deformation. Through the above steps, we can take the extracted skeleton as the target, and the warp template model to match with the target pose. In order to better match the real surface of the human body, we further develop a deformation scheme based on the semantic masks. We perform the deformation on the warped template so that the outline of the

0730-0301/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

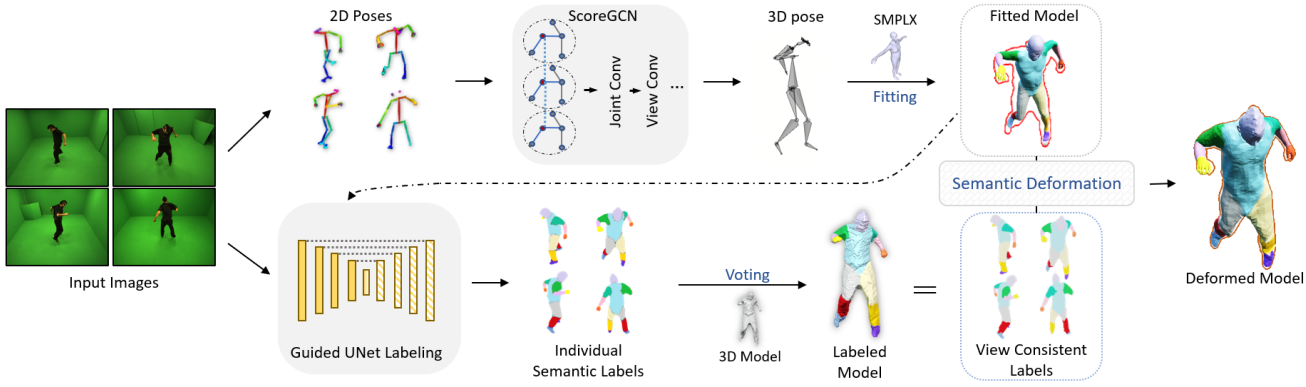

Fig. 1. Our human reconstruction pipeline. We take the multi-view images as input, and then generate a 3D human pose through the ScoreGCN. Then we fit SMPLX model with the pose and use it as our prototype. However, the deformation is solely based on 3D skeleton and not well match the image observations. To produce a more accurate geometric model, we utilize a semantic driven deformation approach to match the model with the consistent labels of each view. Our method can recover a relatively accurate human model even in the featureless and sparse camera views scenario.

final deformed model is perfectly aligned with the silhouettes and semantic segmentations of each view.

We conduct comprehensive experiments on evaluating our approach vs. the state-of-the-art reconstruction-based and deformation-based techniques on publicly available datasets and our newly generated complex motion datasets. We show our technique significantly outperforms the state-of-the-art on sparse cameras, texture-less regions (e.g., wearing black clothes), complex motions, etc.

2 RELATED WORK

Freedom Surface Reconstruction: In the past few decades, people have paid increasing attention to the reconstruction of real human appearance and movement. Most existing methods can be classified as passive and active reconstruction [Zhang et al. 2019].

The passive methods use multi-view camera systems [Mikhnevich and Hébert 2011] and strongly rely on stereo matching, triangulation, and meshing. A series of work by the CMU group create the most notable ones in passive approaches. Kanade [Kanade et al. 1997] and Narayanan [Narayanan et al. 1998] invent one of the earliest dome systems (with 51 cameras) and utilize multi-view stereo solution to reconstruct dynamic human models. Shape-from-silhouette based methods [Franco et al. 2006; Kutulakos and Seitz 2000; Matusik et al. 2000; Moezzi et al. 1997] are widely studied because they are fast and reliable. But the quality remains unsatisfactory as they cannot represent concavities and could lead to visible artifacts.

The active approaches use active sensors [Dou et al. 2016; Newcombe et al. 2015; Orts-Escolano et al. 2016; Xu et al. 2019; Yu et al. 2018] to get the additional depth information and employ fusion based methods to recover human motion sequence. These methods can achieve real-time performance, but they can not generate geometry with surface details as multi-view systems do.

Currently, [Collet et al. 2015b] and [Guo et al. 2019] combine active and passive sensors to build large controlled studios, they are able to produce high-quality results on human motions. But these require expensive infrastructure and are computationally intensive.

Learning-based Reconstruction: With the fast development of deep learning, many methods are proposed to learn deep features to rebuild human shapes.

Most of them define the problem using statistical models. They boil down the problem to estimate the parameters of the model. Dibra *et al.*[Dibra et al. 2016, 2017] encode statistical models into the network and regressed the model parameters from 2D silhouettes. But their methods can only get those naked shapes in a rest pose. Some pay more attention to static pose to get more realistic captures. [Bogo et al. 2016; Pavlakos et al. 2019a] propose SMPLify series to recover human pose and shape from a single image. They first estimate the 2D joints by a network. Then they regress the 3D model to fit those 2D estimations by minimizing several terms: data term, pose term, and shape term. Their methods can recover 3D humans from arbitrary poses. However, their recovered models are also naked models without detailed surface geometry.

Instead of using statistical models, some incorporate class-specific knowledge into the reconstruction. For example, Huang *et al.* [Huang et al. 2018] regard reconstruction as a classification problem. They map sparse images to a 3D volumetric field that encodes the probabilistic distribution of surface points. Then they reconstruct the surface geometry from the 3D field.

Deformable Models: There is also a trend in substituting the reconstruction pipeline with shape deformation. Loper *et al.* [Loper et al. 2015] introduced a learned model of human body pose and shape called Skinned Multi Person Linear (SMPL) that is capable of representing a wide variety of body shapes under natural human poses. A similar approach is used to produce a generative 3D hand model [Romero et al. 2017] for represent a fully articulated human body. It may even be possible to infer SMPL from a single image or video [Alldieck et al. 2018; Bogo et al. 2016]. Total capture [Joo et al. 2018] extends [Loper et al. 2015] with deformable hand and face geometry along with hair and clothing to form the template *Adam*

model and has shown great success on motion capture and shape reconstruction. However, the amount of allowable deformations is still limited and large errors can occur when the target motion varies greatly from the template, either in movements or clothing. Xu *et al.* [Xu et al. 2018] replaced the generic parametric model with a person-specific template mesh captured under T-pose. We instead use semantic silhouette based deformation to refine the template.

The key step in successful shape warping is pose alignment. Qiu *et al.*[Qiu et al. 2019] use a cross-view fusion scheme to jointly estimate the 2D poses in multi-view images and then develop a Pictorial Structure Model to recover the 3D pose. Recent work [Iskakov et al. 2019] conducts differentiable triangulation to generate multi-view 3D poses and achieves 20.5 mm MPJPE (mean per joint position error) on the Human3.6M dataset [Ionescu et al. 2014]. The most challenging part of the human body is hands. Simon *et al.*[Simon et al. 2017] use convolutional pose machine and multiview bootstrapping to detect hand keypoints, provided the image covers the complex body including hands. They further develop a robust triangulation method to generate the 3D skeleton. We observe not all views are equally beneficial for skeleton estimation and the human skeleton is anatomically related, so we encode this constraint into the GCN network to score each view. And with the better view scores, we can get the more accurate 3D skeleton.

3 OVERVIEW

We aim to reconstruct the 3D human shape from a set of sparse images captured at different viewing positions. Such sparsity of views imposes challenges for traditional MVS and SfM approaches as it leads to noise or missing geometry. We turn to the recent progress in parametric human models, such as SMPL [Loper et al. 2015] and SMPL-X [Pavlakos et al. 2019b], to fully exploit the human shape prior knowledge. Instead of recovering each pixel's 3D position, our method first roughly fits a parametric human model with the observed pose and then fine-tune the fitted model through semantic-driven deformation. The pipeline of our human reconstruction approach is shown in Fig. 1. Specifically, we first extract the 2D human pose in each image using the backbone network in [Iskakov et al. 2019] and feed all the 2D poses to a ScoreGCN for weighting the joints in each view. A more accurate 3D skeleton position is obtained via weighted triangulation (Sec. 1) and then we fit the SMPL-X model to the 3D skeleton (Sec. 4.2). However, the fitted model is usually over smoothed and contains artifacts due to incorrect deformations. Hence we introduce a semantic based shape deformation method to match the parametric model with the image observations (Sec. 5). In this step, we estimate consistent semantic labels of the human body in each image with the guidance of the skinned parametric model (Sec. 5.1) and the visual hull geometry. We then deform each semantic part of the parametric model to match with the consistent semantic labels as our final human shape.

4 POSE MATCHING

To fully exploit the human shape prior, it is an essential step to match the poses of the parametric model with the image observations. However, the human poses presented in the images are 2D dimensional while we need the 3D skeleton for model fitting.

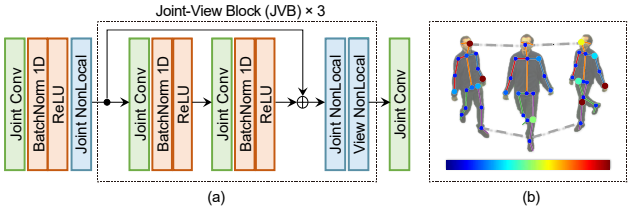

Fig. 2. We perform score refinement on input 2D pose estimation results through ScoreGCN. ScoreGCN consists of multiple layers of joint and view graph convolution, shown in (a). (b) is the visualization of the joint view graph. Dots with color denote the body joints, blue means lower weight while red means higher weight. The solid lines denote intra-body edges, they are defined based on biological connections. Dotted lines denote the connection between different views. We only show part of the view graph for clarity.

To estimate the 3D pose, we first extract the 2D pose and the corresponding confidence of each joint from each image via the 2D human pose detector provided in [Iskakov et al. 2019]. Yet recovering the 3D pose from 2D observations is non-trivial as not all joints are accurately perceived in an image. One image may only be able to recover the 2D positions of a subset of joints with high fidelity due to errors or occlusions. The confidences reported by [Iskakov et al. 2019] consider each joint and view individually, and hence are not accurate indicators for weighted triangulation. In this section, we present a ScoreGCN to estimate the accuracy scores of the 2D joints. Specifically, we take the 2D poses and joint confidences of all images as input and use the graph convolution network for generating the accuracy scores. The final 3D pose is obtained through weighted triangulation from the 2D estimations of joint positions and accuracy scores.

4.1 GCN-based 3D joints estimation

Our 2D joint accuracy score estimation network ScoreGCN is inspired by ST-GCN[Yan et al. 2018], which treats the joints of the human skeleton as graph nodes, bones as edges and the skeletons are further temporally connected. Similar to ST-GCN's structure, we construct two graphs, the joint graph and the view graph, and which are generally represented as $G=(V, E)$. The node set $V = \{v_{ci} | c = 1, ..., M, i = 1, ...N\}$ includes the 2D joints from all views. M is the number of views, and N is the number of joints. The input feature vector $F(v_{ci})$ consists of the 2D position and confidence of the ith joint in view c. We construct the joint view graph as follows: first connect the joints within one view according to the kinetic structure of the human body as illustrated in Fig. 2(b), then connect the same joint in different views as the dotted lines in Fig. 2(b).

In this setup, there are two types of edges in the edge set E: the one that describes the biological connection of joints, denoted as $E_S = \{v_{ci}, v_{cj} | (i, j) \in H\}$, where H is the set joints connected to each other biologically, and the one that connects the same joint in different camera views as $E_F = \{v_{ci}, v_{di} | (c, d) \in C\}$, where C is the set of camera views. We use a fully connected scheme for the view

graph, that is we connect each two joints which refer to the same joint and belong to different views.

We use the *Semantic Graph Convolution (SemGConv)*, proposed by Zhao *et al.* [Zhao et al. 2019], for graph convolution. The SemGConv introduces a set of learnable edge weighting matrices for each channel of the node features. Through learning this *input-independent* weights for edges, we can model the priors implied in the graph structures appropriately, e.g. how one joint influences other body parts in a single view. As in a certain view, an accurate estimation of the shoulder usually implies we can recover the arm with relatively high confidence. We call the SemGConv along the joints in the same view the Joint Convolution Layer (Joint Conv).

We combine the Joint Conv and non-local layers, similar to [Zhao et al. 2019], to capture the local and global relations of nodes. The structure of our ScoreGCN is shown in Fig. 2(a). We concatenate 3 Joint-view Blocks, and each Joint-view Block consists of one residual block with two Joint Convs followed by one joint non-local layer and one view non-local layer. Each Joint Conv is followed by a batch normalization layer and a ReLU activation layer except for the last one. The Joint Conv in the entry is for mapping the inputs into the latent space, while the one in the end projects the encoded features to the output space. Notice that we add the view non-local layer as we focus on multi-view configurations and our Joint-view Block downgrades to SemGCN[Zhao et al. 2019] without view non-local layer. The node features are updated locally first and then refined by SemGConv and non-local layers inside the Joint-view Block.

We then use a weighted triangulation approach to infer the 3D joints from multi-view 2D estimations x_i and its corresponding confidence score w_i, inspired by [Iskakov et al. 2019]. The method solves for the homogeneous 3D joint position vector X from the following overdetermined equation system:

$$(w_i \circ A_i)X = 0, \tag{1}$$

where i denotes the ith view in C, $A_i \in \mathbb{R}^{(2C,4)}$ is the matrix generated from camera projection matrices and x_i, see [Hartley and Zisserman 2004] for details. The scores w_i are produced by ScoreGCN. $\circ$ denotes the Hadamard product (element-wise product). During the training process of ScoreGCN, we use the L2 distance between weighted triangulation results and the ground truth as the loss. And we solve Eqn.1 via differentiable Singular Value Decomposition (SVD), which allows the gradients to flow back to w_i from 3D joint X.

4.2 Parametric Model Fitting

We use the SMPL-X [Pavlakos et al. 2019a], the current state-of-the-art 3D human model with minimized skinning artifacts, as our template and deform it to fit with the 3D skeleton. Specifically, the fitting process is according to:

$$E(\beta, \theta) = \lambda_j E_{joint} + \lambda_b E_{bone} + \lambda_\beta E_\beta + \lambda_\theta E_\theta + \lambda_p E_p, \tag{2}$$

where β and θ are the shape and pose parameters of SMPL-X respectively. The data term E_{joint} is for joint alignment, E_{bone} is the bone length term, E_β, E_θ are the regularization terms, and E_p penalizes penetrations. For the data term, we minimize the L2-distance between estimated 3D joint X_i and the corresponding posed SMPL-X joint $\mathcal{J}(\beta, \theta)_i$ for each joint i. It is defined as:

$$E_{joint} = \sum_{i=1}^{K} ||\mathcal{J}(\beta, \theta)_i - X_i||_2^2, \tag{3}$$

K is the number of joints, $\mathcal{J}$ is a sparse linear regressor provided by SMPL-X that regresses 3D joint locations from mesh vertices. The data term constrains the joints of the SMPL-X to match with our 3D joint estimations as much as possible.

We also add bone length constrain through E_{bone}. We define a bone vector b_j as the 3D offset of the i-th joint X_i relative to its immediate parent joint $X_{parent(i)}$:

$$b_j = X_i - X_{parent(i)} \tag{4}$$

Then the bone length term is given by the following equation:

$$E_{bone} = \sum_{j=1}^{B} ||b_j^s - b_j||_2^2, \tag{5}$$

where B is the number of bones, b_j^s denotes SMPL-X bones and b_j denotes our 3D estimation. We use this term to match SMPL-X to the current model scale.

For the regularization terms, following [Pavlakos et al. 2019a], we define them as $E_\beta(\beta) = ||\beta||^2$ and $E_\theta(\theta) = ||\theta||^2$. They both describe the Mahalanobis distance between the parameters being optimized and the distribution in the training dataset of SMPL-X. They are able to prevent the deformation deviating too far from regular human poses and shapes.

E_p is a penalization term that prevents self-collision and penetrations of several body parts during fitting. We employ the same collision penalizer from [Ballan et al. 2012; Pavlakos et al. 2019a; Tzionas et al. 2016]. We first detect a list of colliding mesh triangles and compute local 3D distance fields Φ. We penalize the penetration by the depth of the intrusion, computed by the position in the distance field. For the details about Φ and handling collision, we redirect the reader to [Ballan et al. 2012; Tzionas et al. 2016].

In practice, we set $\lambda_j, \lambda_b, \lambda_\beta, \lambda_\theta, \lambda_p$ to 4, 1, 0.05, 0.05, 1 respectively. We use a Limited-memory BFGS optimizer (L-BFGS) [Noc 2006] to minimize Eq.2. Gradients are computed on GPU by automatic differentiation using Pytorch [Paszke et al. 2017].

5 SHAPE MATCHING

The fitted template model aligns well with the actual 3D pose, except for its actual shape. We conduct non-rigid shape deformation to ensure the appearance of the deformed model matches the captured images. One possible approach is to conduct 3D deformation by deforming the template to the recovered 3D point cloud. Since the cases we aim to handle have minimal texture and sparse camera setting, a reliable 3D point cloud is not accessible. We, therefore, present a new 2D semantic-driven deformation scheme.

The first step is to generate the semantic masks(smasks for short) for the human body parts. However, such task is challenging as we focus on the multiview configurations and textureless scenarios. Current popular segmentation methods, such as Deeplab[Chen et al. 2018], rely on image textures and only tackle a single image, and hence are not applicable to our case. Here we propose a pose guided

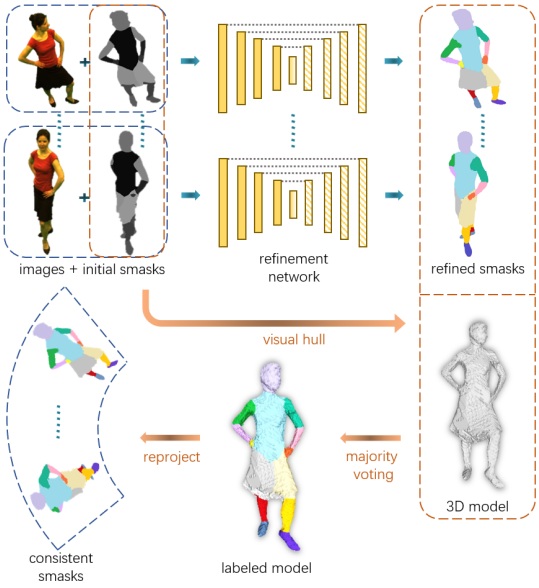

Fig. 3. Our view consistent semantic masks generation method. We utilize fitted model as priors to get the initial smasks and concatenate with the images as input, and fed them into our refinement network. To enforce view consistency, we run majority voting and reprojection using the network output(the refined smasks) and the visual hull geometry. Although the 'consistent smasks' may still not be perfect as the visual hull mesh is not the true geometry, we observe they could serve as an accurate guidance for our deformation.

semantic segmentation method to enforce the correctness and cross-view consistency of the smasks.

5.1 Veiw Consistent Semantic Masks Generation

To generate accurate and view consistent smasks, we rely on the fitted model to provide priors. Notice despite the high similarity between the skeletons of the fitted template and captured models, their actual 3D shapes can have large variations, e.g., exhibiting large gaps near the boundary, as shown in Fig. 4(c).

We first conduct semantic labeling on the segmented human images, i.e. assigning each pixel a semantic body part label and obtain the initial smasks. We adopt the 14 human body part segmentation scheme: head, torso, upper left and right arms, lower left and right arms, left and right hands, upper left and right legs, lower left and right legs, and two feet. Recent MRF-based techniques [Liang et al. 2016] and learning-based approaches [Liang et al. 2015, 2017] for 2D human parsing have shown great success. However, they all focus on labeling a single image and not suitable for the multi-view semantic labeling problem. We instead use the canonical model as prior to maintain the consistency.

Specifically, we project the fitted canonical model into each view. The canonical model itself already has semantic labels defined for each vertex and for each view we can obtain a semantically labeled image M_s. Next, we simply vote on each foreground pixel for its semantic label. For each foreground pixel p, we set its label as the most frequent label appears within a radius r circle around it. In cases where there are not sufficient labeled pixels (e.g, the gap area near the boundary), we simply increase r until it covers sufficient labeled pixels. We use the generated smasks as the initial labels for guiding the final smasks generation. Some example initial smasks are shown in Fig. 7(c).

Notice the initial smasks usually still exhibit errors (due to occlusions or inconsistent gaps). We resort to a learning based approach to identify and correct incorrect labels. We trained a refinement network based on a variant of the "U-Net" structure [Fang et al. 2018; Isola et al. 2017]. Our network takes the images and initial smasks as input and generates the refined masks. We adopt the L1 distance between the output and the ground truth label as our loss function. After refinement, the boundaries between different body parts are more accurate, see the boundaries on the legs and dresses in Fig. 7. To further enforce the across view consistency, we test the multiview labels on the visual hull geometry. For each face f on the visual hull geometry, we project it into each view and obtain the corresponding semantic label. We set the majority label as the final label for the corresponding face. Though the result may still not be perfect as the visual hull mesh is not the true geometry, we observe they can still serve as fair accurate guidance for our fine deformation.

5.2 Semantic-Driven Deformation

Finally, we conduct shape deformation based on the smasks to generate a more accurate geometry. Our approach resembles the method proposed in [Joo et al. 2018]. However, [Joo et al. 2018] uses a 3D point cloud generated by the MVS technique as the deformation target. In our case, the MVS is unable to produce satisfactory point cloud for the sparsity of the views and textureless regions. We, therefore, resort to deform the 3D template according to observed 2D images and labels. Our deformation approach is based on the graph node based non-rigid deformation technique which maintains a relative small set of control nodes. The transformation of the complete vertex set is according to a weighted combination of their neighboring control nodes. The average distance between two deformation nodes is ϵ and each node will affect vertices within 3ϵ geodesic distance from it. We construct a transformation graph from these nodes, where each node defines a local warp field. We follow the method in [Xu et al. 2018] to update each vertex according to node transformation.

To solve for optimal deformation of the template model, we set out to minimize the following function:

$$\mathbf{E} = \lambda_m E_s + \lambda_r E_r, \tag{6}$$

where E_s is the semantic label alignment term, E_r is the smoothness term and λ_m and λ_r are the weighting factors. The semantic label alignment term aims to enforce the semantic parts of the deformed 3D template closely match with the 2D labels.

$$E_s = \sum_{v \in \mathcal{V}} \sum_{j \in \mathcal{L}} \sum_{u_i \in \mathcal{B}_j} [n_j^v \cdot (\Pi_v(u_i) - s_j^v)]^2, \tag{7}$$

where $\mathcal{V}$ is the selected view set, $\mathcal{L}$ is the collection of labels, $\mathcal{B}_j$ is the set of boundary vertices on template model with the current label j. $\Pi_v(\cdot)$ is the 3D to 2D projection under view v. s_j^v is the closest

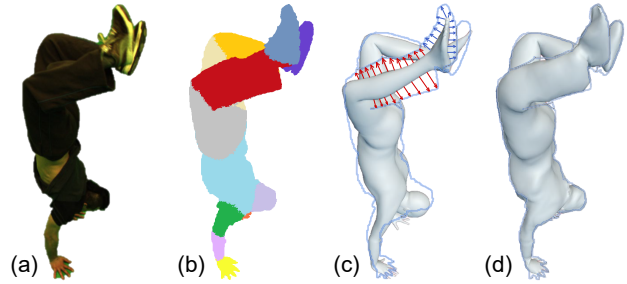

Fig. 4. An example result of our semantic-driven deformation. (a) Original image. (b) Semantic mask. (c) Before semantic deformation. Shape matching result overlays with outer silhouette. Red and blue arrows indicate surface deformation direction. (d) After semantic deformation.

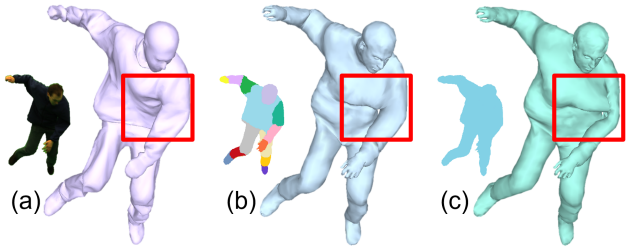

Fig. 5. With and without semantic information. (a) Ground truth mesh. (b) Non-rigid deformation with semantic masks. (c) Non-rigid deformation with binary masks. Note that the arm part is distorted in (c) due to wrong correspondence.

pixel of $\Pi_v(u_i)$ on the target semantic silhouette with normal n_j^v. We consider both the Euclidean distance and normal similarities while searching for a closest point. This term encourages each semantic part of the model to be tightly aligned with the target silhouette.

The smoothness term enforces the neighbor nodes to have similar transformations and hence ensures the deformation is smooth.

$$E_r = \sum_{p_i \in \mathcal{N}} \sum_{p_j \in \mathcal{M}_i} w_{ij}(\mathbf{T}_i p_j - \mathbf{T}_j p_j), \tag{8}$$

where $\mathcal{N}$ is the collection of nodes, and $\mathcal{M}_i$ is the neighbors of node i. $\mathbf{T}$ is the transformation parameters (represented as rotation matrix and translation vector). p_j is the 3D coordinates of node j. The weight between adjacent nodes p_i and p_j is defined as follows:

$$w_{ij} = e^{-\|p_i - p_j\|_2^2 / 2\sigma^2}, \tag{9}$$

larger weight corresponds to closer nodes.

The vertices near the boundary between two semantic body parts are the most challenging to tackle due to the ambiguity, i.e. it's difficult to decide which target position a boundary vertex should deform into. We rely completely on the smoothness term E_r to ensure a natural transition between the labels. This significantly reduces the "detach" artifacts, e.g. the end of the palm disconnects from the forearm due to inconsistent transformations. Specifically, we separate the boundary vertices to the inner-set and outer-set. Inner-set vertices are the ones between each semantic part, while the outer-set are the rest, which encodes the actual shape of the body part. We detect the inner boundary vertices with the guidance of the warped model and set their λ_m to 0 in E_s.

Recall we have semantically labeled 2D images and the fitted model based on 3D human pose, and the human parts exhibit strong heterogeneity during deformation. We, therefore, set different weights for different semantic parts: e.g. for face and hand, we observe high coherence between the warped and ground truth models and hence we use $3\lambda_r$ for hand and $5\lambda_r$ for face to enforce stronger regularization.

In practice, we perform the alignment in a hierarchical coarse-to-fine manner. During iterations, we increase the number of nodes and λ_m, and decrease λ_r. We set the initial values of λ_m and λ_r to 20 and 10 respectively, and the final values to 80 and 0.8. Meanwhile, we gradually reduce ϵ from 80 mm to 20 mm to extract deformation nodes, which is about 200 to 2500 nodes in our experiment. The hierarchical approach allows the deformation to match with largely misaligned poses and to capture more geometric details at the same time. We apply a GPU-based Gauss-Newton solver similar to [Xu et al. 2018] to solve for the optimal transformations.

Fig. 5 shows a comparison between our semantic-driven deformation and pure silhouette based deformation. The deformed model generated by a pure silhouette approach exhibits large mismatches with the actual model, especially in the arm area, Fig. 5(c). Our semantic-driven method is able to fit the template model well with the boundary and keep the smoothness between semantic parts in the meanwhile.

Compared with [Xu et al. 2018], our method imposes consistency on the semantic parts instead of only foreground masks. We show a comparison of non-rigid deformation with and without semantic information in Fig. 5. Note that the upper arm is distorted in (c) because the torso and arm are not separated in the binary mask. Our semantic deformation method manages to compensate these differences and outperforms the pure silhouette based approaches [Xu et al. 2018].

6 EXPERIMENTAL RESULTS

To evaluate the performance of our method, we compare with existing methods on both publicly available MIT dataset [Vlasic et al. 2008] and our own dynamic motion datasets. All experiments have been conducted on a computer with Intel Core i7-8700K, 32 GB memory and one NVIDIA 1080 Ti GPU.

Our method consists of three major steps: the 3D pose estimation using ScoreGCN, the pose guided consistent labeling and the semantic driven deformation. In this section, we will show the importance and effectiveness of each step.

6.1 Comparison of 3D Pose Estimation from Multi-view 2D Poses

The first step of our framework is to recover the 3D human pose from a set of multi-view 2D poses. As there is no ground truth 2D and 3D skeletons provided in the MIT dataset, we train and test our ScoreGCN on the Human3.6M[Ionescu et al. 2014] dataset.

The Human3.6M is currently the largest dataset available for multi-view 3D human pose estimation. It contains 3.6 million images with ground truth 3D skeleton annotation captured by a 4-view MoCap System in an indoor environment, with 7 subjects and 15 daily activities for each subject. However, the images in Human3.6M differ greatly with that of the MIT dataset. As the subjects in Human3.6 wear clothes with rich texture while the MIT dataset focuses on textureless subjects. Also, the Human3.6M has 4 viewpoints, while the MIT dataset has 8. In order to bridge the gap, we generate a synthetic dataset with 10 artist rigged human model (as shown in Fig. 6) to mimic MIT setting. Each rigged model performs 2 types of activities through skeleton driven human body animation with data provided in Mixamo [Adobe 2019], and we sample around 60 frames from each motion sequence and project the animated mesh to 8 cameras views. The camera intrinsic and extrinsic parameters are designed to be similar to the MIT dataset. We split the synthetic dataset into training, validation and testing set, each contains 748, 321, 50 frames. As for Human3.6m, we follow the partition strategy proposed in [Iskakov et al. 2019].

For Human3.6m, we directly apply the 2D backbone network provided by [Iskakov et al. 2019] to generate the input data for ScoreGCN, then train ScoreGCN with the loss proposed in Equ.1. In order to bridge the gap between Human3.6M and MIT data, we fine-tune the backbone on the synthetic data for 10 epochs using the Adam optimizer with 10^{-4} learning rate with 2D skeleton supervision and per-joint Mean Square Error (MSE) as the loss.

For evaluation, we use the Mean Per Joint Position Error (MPJPE) as the metric, which computes the Euclidean distance between the ground truth joints and the triangulated 3D joints measured with respect to pelvis joint.

We evaluate our method on two configurations, with and without the view non-local layer (VNL) for ablation study. We also compare with LTN[Iskakov et al. 2019]. The baseline method LTN-algebraic is similar to our setup, they use fully connected layers for joint score prediction and weighted triangulation for 3D joint estimation. The current state-of-the-art method LTN-volumetric uses volumetric triangulation and 3D joint heatmaps for pose estimation. We show the comparison results in Table 1.

Our ScoreGCN model clearly outperforms the baseline LTN-algebraic model. And the network structure with view nonlocal layer performs better than that without view nonlocal layer. This indicates the effectiveness of joint and view graph. While on the Human3.6 dataset, our network exhibits little improvement compared to LTN-algebraic. We believe it is because the Human3.6 dataset only contains 4 views and hence the estimated score by ScoreGCN has relatively less effect on the final accuracy compared to 8 views, since ScoreGCN requires a subset of views has a good initial estimation of the 2D joints.

6.2 Comparison to Multi-view Labeling

To train our refinement network, we continue to utilize the synthetic dataset in Sec. 6.1. Our synthetic dataset contains both the multi-view images and animated human body sequence. We conduct semantic labeling on the animated body mesh and then project to each view as our ground truth semantic labels. We further split the images into two sets for training and validation purpose. After trained on our synthetic dataset, we fine-tune our refinement network on the MIT dataset (We only trained the network on the 'squat1', 'squat2', 'march1', 'march2' and 'swing' sequences, and left the others as testing data. All the shown cases are not in the training data). Since this mesh dataset also contains the multi-view images and animation meshes. We can generate ground truth labels in the same way. Because these two datasets are both captured in a controlled environment, we assume the human body as the foreground can be extracted easily. Therefore, we use the extracted human body combining with the initial smask to fed into the network. From our practice, this trick can improve the labeling accuracy and behave more robust.

In our method, there are multiple steps of the optimization in the labeling process, and each step refines the results. Fig. 7 shows a qualitative comparison between 2D and 3D. The 2D segmentation represents the result of each step, it can be seen that the result improves in each step. The 3D consistent error maps (the second row of each example) measure the labeling consistency among views. They represent the consistent error rate of each mesh face – We reversely project each pixel(with its label) to the ground truth mesh. Instead of directly comparing the back-projected pixel label to ground truth,

Fig. 6. The 10 artist rigged human models used for generating our synthetic data for the ScoreGCN training.

Method/MPJPE(mm)	H3.6M	Synthetic
LTN-algebraic	22.50	18.54
LTN-volumtric	**20.25**	17.72
ScoreGCN without VNL	22.23	17.34
ScoreGCN with VNL	22.16	**16.82**

Table 1. Compare with state of the art LTN[Iskakov et al. 2019] on H36M dataset[Ionescu et al. 2013] and our synthetic dataset using MPJPE (Mean Per Joint Position Error).

Method	torso	arms	legs	hands	head	mean
U-Net	84.12	76.14	89.65	64.80	92.35	81.41
Ours-1	77.92	70.40	86.10	60.30	90.52	77.05
Ours-2	84.02	77.80	90.65	67.44	94.15	82.81
Ours-3	**87.25**	**81.03**	**94.29**	**75.67**	**95.66**	**86.78**

Table 2. The quantitative results of multi-view labeling in MIT testing data. Ours-1 (initial smask), Ours-2 (guided U-Net), Ours-3 (majority voting).

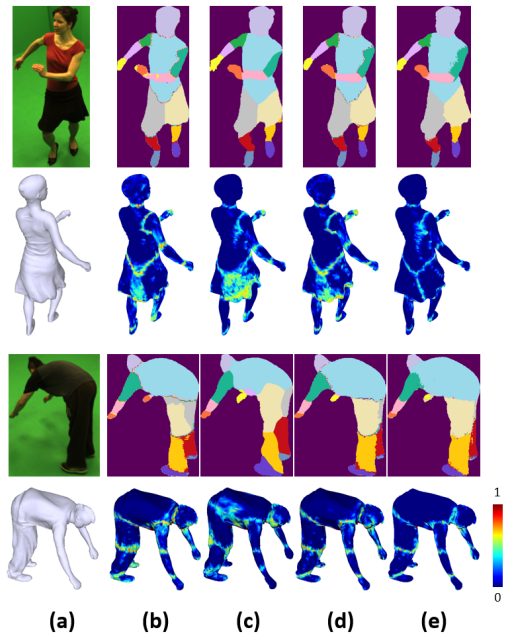

Fig. 7. Multi-view labeling qualitative comparisons of U-Net and our method step by step. (a) Reference images and ground truth models. From columns (b) to (e): U-Net, ours (initial smask), ours (guided U-Net) and ours (majority voting) 2D labeling results and 3D consistent error maps. Our labeling results and cross-view consistency improves step by step and outperform the U-Net apparently.

we set the maximum number of the reversely projected label as the face consistent label(to emphasize the consistency). And the consistent error rate is equal to all pixel numbers divided by inconsistent pixel numbers for each mesh face.

As shown in Fig. 7, our labeling results improve step by step and outperform the U-Net visibly. We also do the quantitative experiment on the MIT testing data. As shown in Table 2, our final labeling results get the highest IoU in every parsing part(especially in 'hands').

6.3 Comparison to Multi-view Reconstruction

As shown in Fig. 8, we provide a qualitative comparison between our approach and other conventional reconstruction methods. It can be seen that conventional stereo algorithm (PMVS) can reconstruct accurate point clouds, but due to their dependence on feature matching, it is difficult for them to reconstruct dense point clouds given sparse views and little features. Although SurfaceNet[Ji et al. 2017] is able to extract more features and obtains a denser point cloud, it is still difficult to recover accurate point clouds in the case of severely lack of features. [Gall et al. 2009] is quite similar to our method, they use a specific model template for each case, constrained by the template shape, they are far less effective than ours for flexible parts such as hands and face. Currently, many performance capture works are based on SMPL [Loper et al. 2015] series. SMPLX [Pavlakos et al. 2019a] is the latest work. SMPLX is a parametric human model, which is able to represent various human pose and shape with little parameter. However, it lacks surface detail such as cloth and hair especially in the movement of characters. We tackle this problem with an extra nonrigid deformation. Through pose and shape matching, we can reconstruct a more complete and accurate model.

6.4 Real Experiment

To verify our method's performance on real data, we construct a camera array dome system. All cameras are synchronized and capture at a resolution of 2048*1536 at 25 frames per second. We pre-calibrate the camera intrinsic and extrinsic parameters using the traditional structure from motion approach [Schonberger and Frahm 2016] with a calibration target (a textured mannequin). In order to extract human shape more accurately, we also put green curtains as the background, similar to the MIT data.

To get accurate 3D skeletons and labeling results, we re-render the synthetic dataset(Sec. 6.1) in our new setting and fine-tune the models in Sec. 6.1 and Sec. 6.2 to adapt them to our capturing setting. Fig. 9 shows our 3D skeleton, labeling, and textured model results in different poses. Our method is also robust even in challenge high-speed motion sequence cases (Fig. 10).

In Fig. 9, we select six most representative scenes: sign language, badminton, dance, basketball, salsa, and suits. For all black (textureless) area, our method can provide satisfactory geometry (sign language and badminton). For those occlusion cases, such as salsa and dance, their arm and leg movements produce severe self-occlusion. Since our approach learns the natural connections in human bodies, limbs and body are easily separated by our method. Therefore, even if there is serious occlusion (even under the occlusion of objects, basketball case) in the input images, our reconstruction result will eliminate adhesion. Hands are particularly difficult to recover in multi-view reconstruction, due to limited resolution and ultra-complex topology and toplogical variations. In the sign language and dance, our method manages to successfully recover the hand motion using our skeleton estimation and semantic deformation techniques.
Our method can be applied to dynamic motion sequence, we reconstruct the 3D pose of each frame and then reconstruct the 3D human body in each frame and utilize the method in [Li et al. 2018] to temporally smooth the result, as the models all share the same topology. We show our reconstructed 3D motion sequences of the jumping, squatting and dancing activity in Fig. 10.

7 CONCLUSIONS

We have presented a novel dynamic 3D human reconstruction technique based on multi-view pose matching and semantic shape deformation. We have tackled cases most challenging to traditional multi-view stereo matching including texturelessness (e.g., black clothing), heavy occlusions (arms occluding torsos) and depth order ambiguities (complex finger movements).

By far our technique relies solely on pose fitting and deformation, without considering multi-view stereo reconstruction. In reality, part of the human body may exhibit rich texture and hence is suitable for correspondence matching based solutions while the rest may be completely textureless. We plan to develop techniques that can

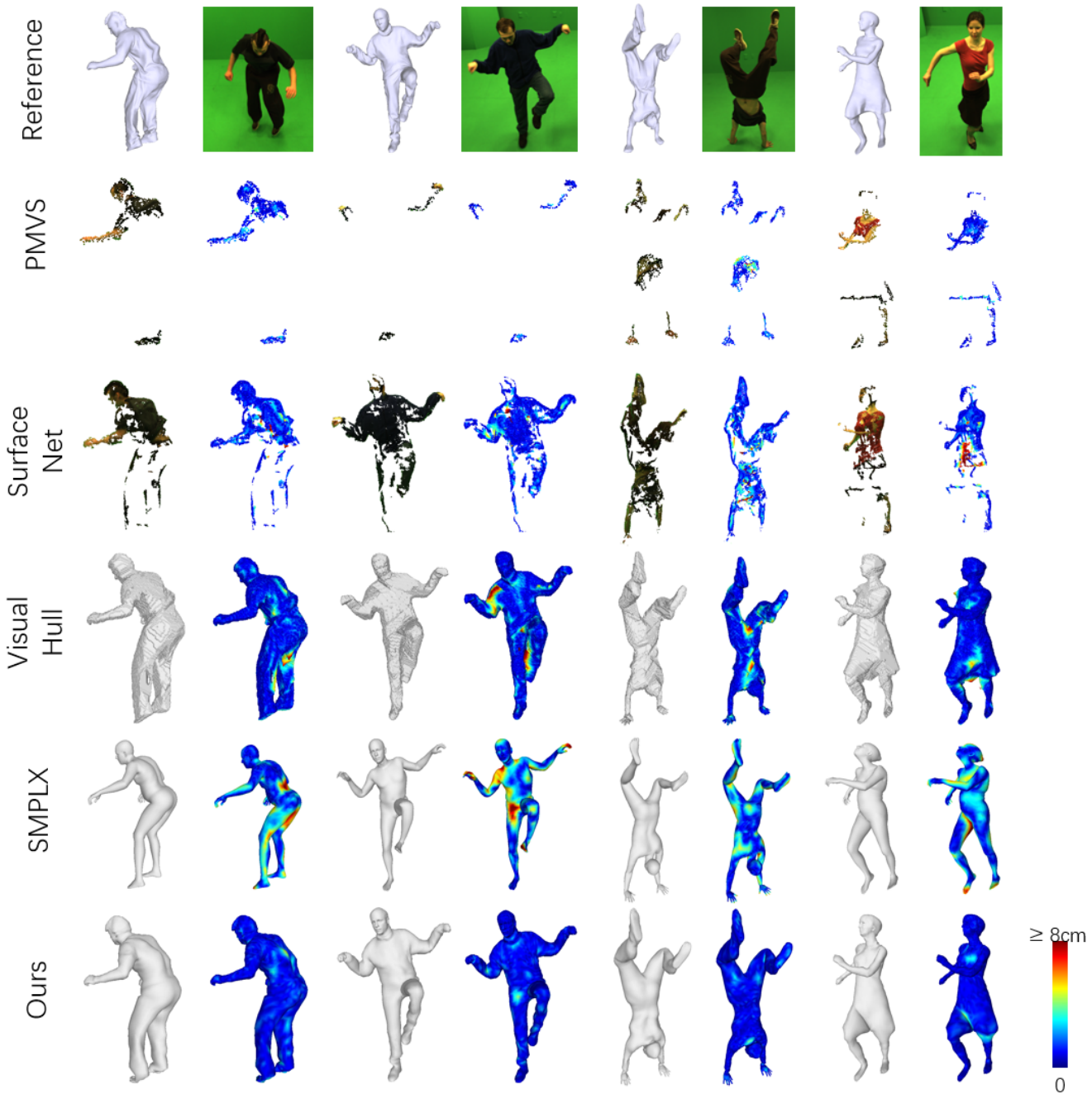

Fig. 8. Multi-view reconstruction comparisons. From top to bottom rows: reference models and images, PMVS, SurfaceNet, visual hull, our fitted smplx and our final deformed results. For our results, mean error distance: left to right: 1.19cm, 0.88cm, 1.01cm, 0.76cm.

adapt to both cases when appropriate. For example, one can use the 3D point cloud (even though sparse) from stereo as prior to 3D deformation while using 2D silhouettes in textureless cases for 2D deformation.

REFERENCES

2006. *Nonlinear Equations.* Springer New York, New York, NY, 270–302. https://doi.org/10.1007/978-0-387-40065-5_11

Adobe. 2019. *Mixamo.* https://www.mixamo.com/

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV).* IEEE, 98–109.

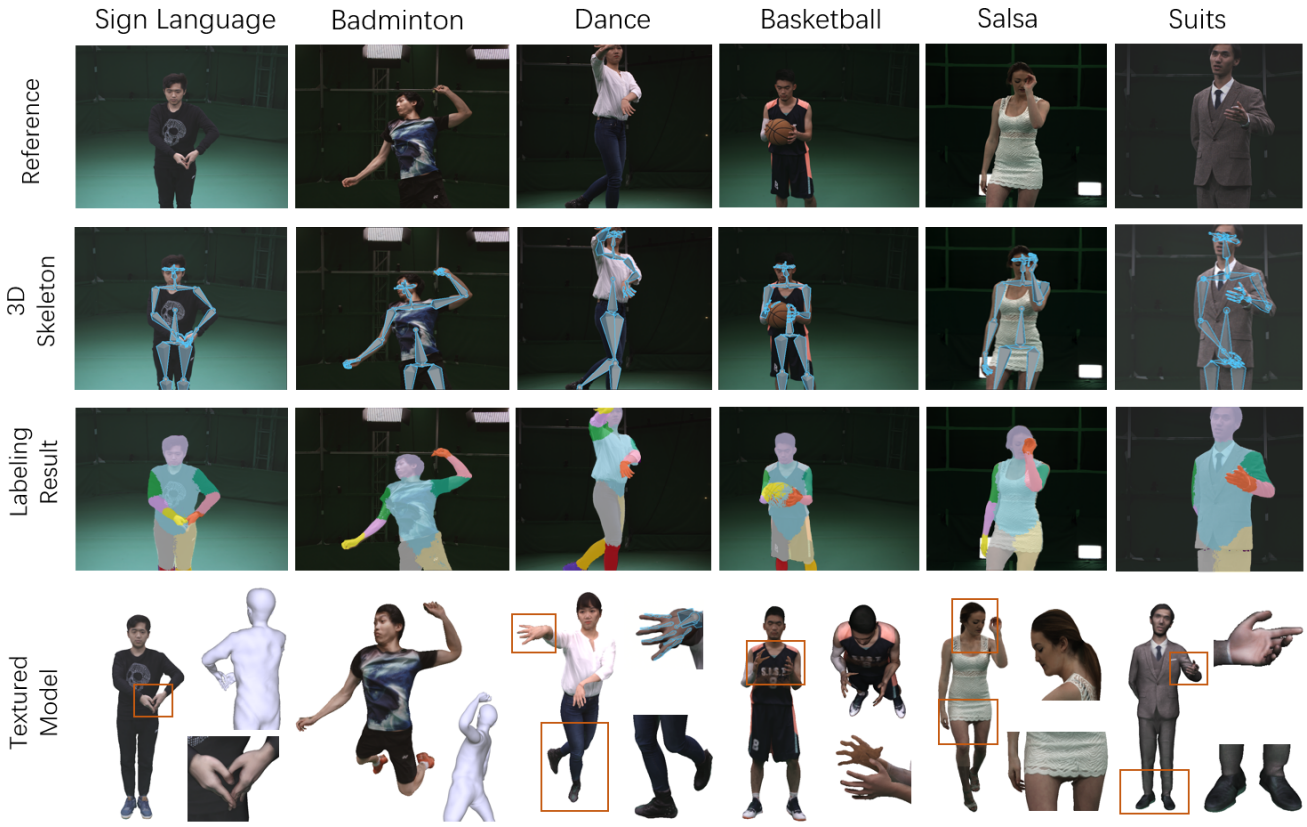

Fig. 9. Our new capturing static results. We select six most representative scenes: sign language, badminton, dance, basketball, salsa, and suits. They can stand for those challenge situations such as featureless, occlusions and hands problems. Note that our technique can recover the high-quality 3D skeleton, multi-view labeling and reconstruction results in these tough situations.

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

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 *European Conference on Computer Vision*. Springer, 561–578.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

KMG Cheung, Simon Baker, and Takeo Kanade. 2003. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, Vol. 1. IEEE, I–I.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015a. High-quality Streamable Free-viewpoint Video. *ACM Trans. Graph.* 34, 4, Article 69 (July 2015), 13 pages. https://doi.org/10.1145/2766945

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015b. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 69.

Endri Dibra, Himanshu Jain, Cengiz Öztireli, Remo Ziegler, and Markus Gross. 2016. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE, 108–117.

Endri Dibra, Himanshu Jain, Cengiz Oztireli, Remo Ziegler, and Markus Gross. 2017. Human shape from silhouettes using generative hks descriptors and cross-modal neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4826–4836.

Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, Pushmeet Kohli, Vladimir Tankovich, and Shahram Izadi. 2016. Fusion4D: Real-time Performance Capture of Challenging Scenes. *ACM Trans. Graph.* 35, 4, Article 114 (July 2016), 13 pages. https://doi.org/10.1145/2897824.2925969

Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. 2018. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310* (2018).

Jean-Sébastien Franco, Marc Lapierre, and Edmond Boyer. 2006. Visual shapes of silhouette sets. In *Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*. IEEE, 397–404.

Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. 2009. Motion capture using joint skeleton tracking and surface estimation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1746–1753.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Ortsescolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics* 38, 6 (2019), 1–19.

R. I. Hartley and A. Zisserman. 2004. *Multiple View Geometry in Computer Vision* (second ed.). Cambridge University Press, ISBN: 0521540518.

Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. 2018. Deep volumetric video from very sparse multi-view performance capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 336–354.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural

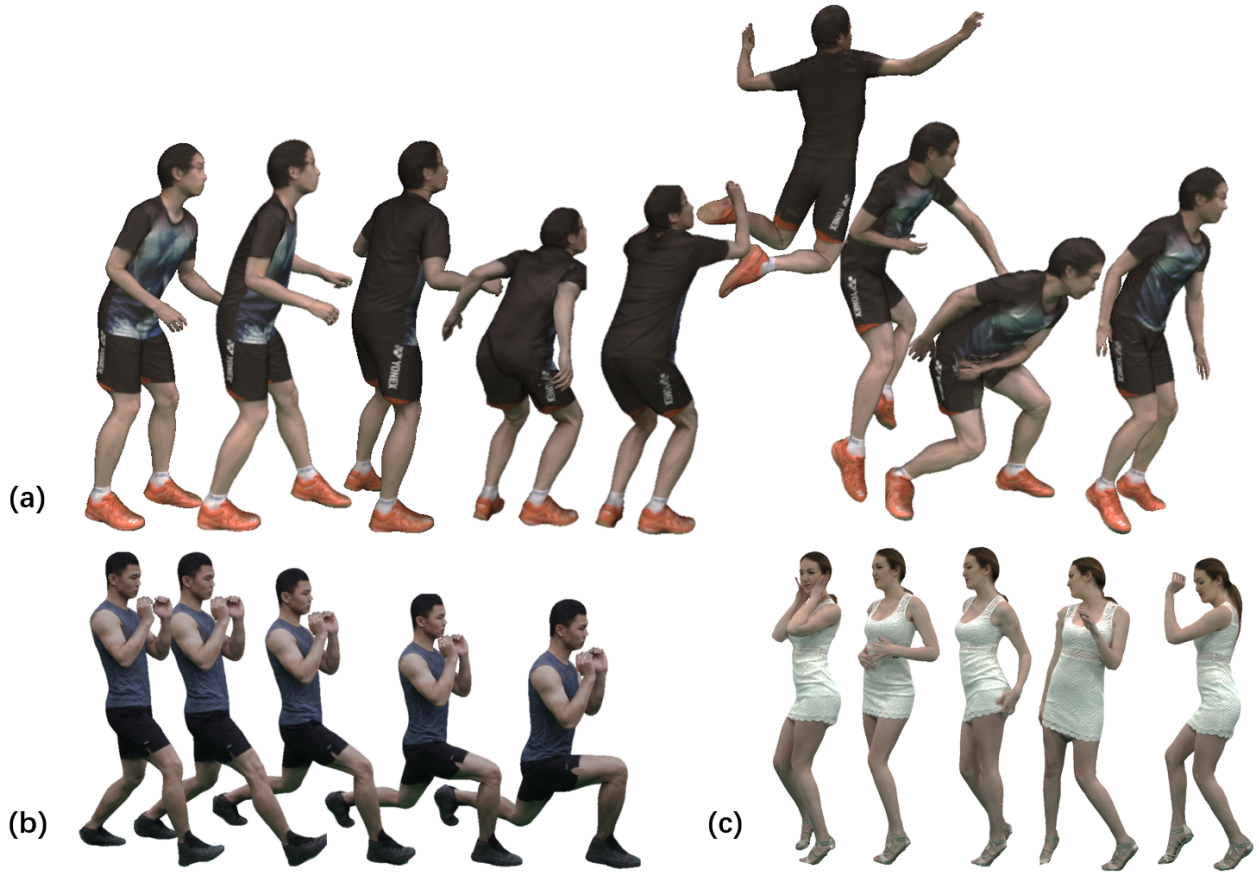

Fig. 10. Our approach can be applied to dynamic motion sequence. (a) jumping, (b) squatting, and (c) dancing.

environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. 2019. Learnable Triangulation of Human Pose. *arXiv preprint arXiv:1905.05754* (2019).

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. 2017. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*. 2307–2315.

Haiyong Jiang, Jianfei Cai, and Jianmin Zheng. 2019. Skeleton-Aware 3D Human Shape Reconstruction From Point Clouds. In *Proceedings of the IEEE International Conference on Computer Vision*. 5431–5441.

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Takeo Kanade, Peter Rander, and PJ Narayanan. 1997. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE multimedia* 4, 1 (1997), 34–47.

Kiriakos N Kutulakos and Steven M Seitz. 2000. A theory of shape by space carving. *International journal of computer vision* 38, 3 (2000), 199–218.

Zhong Li, Minye Wu, Wangyiteng Zhou, and Jingyi Yu. 2018. 4D Human Body Correspondences from Panoramic Depth Maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2877–2886.

Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. 2015. Deep Human Parsing with Active Template Regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 12 (2015), 2402–2414.

Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. 2016. Semantic Object Parsing with Graph LSTM. *european conference on computer vision* (2016), 125–143.

Xiaodan Liang, Chunyan Xu, Xiaohui Shen, Jianchao Yang, Jinhui Tang, Liang Lin, and Shuicheng Yan. 2017. Human Parsing with Contextualized Convolutional Neural Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 1 (2017), 115–127.

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

Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. 2000. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 369–374.

Maxim Mikhnevich and Patrick Hébert. 2011. Shape from Silhouette Under Varying Lighting and Multi-viewpoints. In *Canadian Conference on Computer and Robot Vision, CRV 2011, St Johns, Newfoundland, Canada, May 25-27, 2011*. IEEE Computer Society, 285–292. https://doi.org/10.1109/CRV.2011.45

Saied Moezzi, Li-Cheng Tai, and Philippe Gerard. 1997. Virtual view generation for 3D digital video. *IEEE multimedia* 4, 1 (1997), 18–26.

PJ Narayanan, Peter W Rander, and Takeo Kanade. 1998. Constructing virtual worlds using dense stereo. In *null*. IEEE, 3.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. 2015. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 343–352.

Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 741–754.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019a. 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, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019b. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. 2019. Cross View Fusion for 3D Human Pose Estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 4342–4351.

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

Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4104–4113.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2.

Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. 2016. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision* 118, 2 (2016), 172–193.

Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popovic. 2008. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.* 27, 3 (2008), 97:1–97:9. https://doi.org/10.1145/1360612.1360696

Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and FANG Lu. 2019. Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercialrgbd cameras. *IEEE transactions on pattern analysis and machine intelligence* (2019).

Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (TOG)* 37, 2 (2018), 27.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Guangwei Yang and Yebin Liu. 2009. 3D object relighting based on multi-view stereo and image based lighting techniques. In *2009 IEEE International Conference on Multimedia and Expo*. IEEE, 934–937.

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. *arXiv preprint arXiv:1804.06023* (2018).

Yingliang Zhang, Xi Luo, Wei Yang, and Jingyi Yu. 2019. Fragmentation guided human shape reconstruction. *IEEE Access* 7 (2019), 45651–45661.

Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. 2019. Semantic Graph Convolutional Networks for 3D Human Pose Regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3425–3435.