HuMMan: Multi-Modal 4D Human Dataset for Versatile Sensing and Modeling

Zhongang Cai[⋆,1,2,3], Daxuan Ren[⋆,2], Ailing Zeng[⋆,4], Zhengyu Lin[⋆,3], Tao Yu[⋆,5], Wenjia Wang[⋆,3], Xiangyu Fan[3], Yang Gao[3], Yifan Yu[3], Liang Pan[2], Fangzhou Hong[2], Mingyuan Zhang[2], Chen Change Loy[2], Lei Yang[†,1,3], Ziwei Liu[†,2]

[1]Shanghai AI Laboratory, [2]S-Lab, Nanyang Technological University, [3]SenseTime Research, [4]The Chinese University of Hong Kong, [5]Tsinghua University
yanglei@sensetime.com, ziwei.liu@ntu.edu.sg

Abstract. 4D human sensing and modeling are fundamental tasks in vision and graphics with numerous applications. With the advances of new sensors and algorithms, there is an increasing demand for more versatile datasets. In this work, we contribute **HuMMan**, a large-scale multi-modal 4D human dataset with 1000 human subjects, 400k sequences and 60M frames. HuMMan has several appealing properties: **1)** multi-modal data and annotations including color images, point clouds, keypoints, SMPL parameters, and textured meshes; **2)** popular mobile device is included in the sensor suite; **3)** a set of 500 actions, designed to cover fundamental movements; **4)** multiple tasks such as action recognition, pose estimation, parametric human recovery, and textured mesh reconstruction are supported and evaluated. Extensive experiments on HuMMan voice the need for further study on challenges such as fine-grained action recognition, dynamic human mesh reconstruction, point cloud-based parametric human recovery, and cross-device domain gaps.[1]

1 Introduction

Sensing and modeling humans are longstanding problems for both computer vision and computer graphics research communities, which serve as the fundamental technology for a myriad of applications such as animation, gaming, augmented, and virtual reality. With the advent of deep learning, significant progress has been made alongside the introduction of large-scale datasets in human-centric sensing and modeling [29, 53, 60, 62, 103, 113]. In this work, we present **HuMMan**, a comprehensive human dataset consisting of 1000 human subjects, captured in total 400k sequences and 60M frames. More importantly, HuMMan features four main properties listed below.

- Multiple Modalities. HuMMan provides a basket of data formats and annotations in the hope to assist exploration in their potential complementary

⋆ co-first authors; † co-corresponding authors
[1] Homepage: https://caizhongang.github.io/projects/HuMMan/

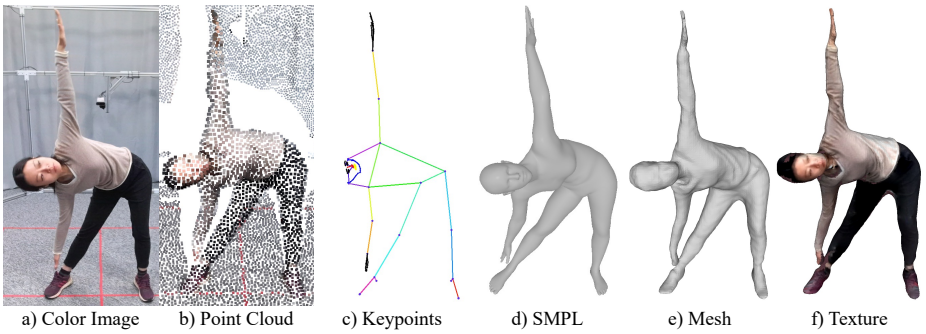


Fig. 1: HuMMan features multiple modalities of data format and annotations. We demonstrate a) color image, b) point cloud, c) keypoints, d) SMPL parameters and e) mesh geometry with f) texture. Each sequence is also annotated with an action label from 500 actions. Each subject has two additional high-resolution scans of naturally and minimally clothed body.

nature. We build HuMMan with a set of 10 synchronized RGB-D cameras to capture both video and depth sequences. Our toolchain then post-process the raw data into sequences of colored point clouds, 2D/3D keypoints, statistical model (SMPL) parameters, and model-free textured mesh. Note that all data and annotations are temporally synchronized, while 3D data and annotations are spatially aligned. In addition, we provide a high-resolution scan for each of the subjects in a canonical pose.

- Mobile Device. With the development of 3D sensors, it is common to find depth cameras or low-power LiDARs on a mobile device in recent years. In view of the surprising gap between emerging real-life applications and the insufficiency of data collected with mobile devices, we add a mobile phone with built-in LiDAR in the data collection to facilitate the relevant research.

- Action Set. We design HuMMan to empower comprehensive studies on human actions. Instead of empirically selecting daily activities, we propose to take an anatomical point of view and systematically divide body movements by their driving muscles. Specifically, we design 500 movements by categorizing major muscle groups to achieve a more complete and fundamental representation of human actions.

- Multiple Tasks. To facilitate research on HuMMan, we provide a whole suite of baselines and benchmarks for action recognition, 2D and 3D pose estimation, 3D parametric human recovery, and textured mesh reconstruction. Popular methods are implemented and evaluated using standard metrics. Our experiments demonstrate that HuMMan would be useful for multiple fields of study, such as fine-grained action recognition, point cloud-based parametric human recovery, dynamic mesh sequence reconstruction, and transferring knowledge across devices.

Table 1: Comparisons of HuMMan with published datasets. HuMMan has a competitive scale in terms of the number of subjects (#Subj), actions (#Act), sequences (#Seq) and frames (#Frame). Moreover, HuMMan features multiple modalities and supports multiple tasks. Video: sequential data, not limited to RGB sequences; Mobile: mobile device in the sensor suite; D/PC: depth image or point cloud, only genuine point cloud collected from depth sensors are considered; Act: action label; K2D: 2D keypoints; K3D: 3D keypoints; Param: statistical model (*e.g.* SMPL) parameters; Txtr: texture. -: not applicable or not reported.

Dataset	#Subj	#Act	#Seq	#Frame	Video	Mobile	Modalities							
							RGB	D/PC	Act	K2D	K3D	Param	Mesh	Txtr
UCF101 [87]	-	101	13k	-	✓	-	✓	-	✓	-	-	-	-	-
AVA [21]	-	80	437	-	✓	-	✓	-	✓	-	-	-	-	-
FineGym [84]	-	530	32k	-	✓	-	✓	-	✓	-	-	-	-	-
HAA500 [15]	-	500	10k	591k	✓	-	✓	-	✓	-	-	-	-	-
SYSU 3DHOI [27]	40	12	480	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D [83]	40	60	56k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D 120 [55]	106	120	114k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D X [93]	106	120	113k	-	✓	-	✓	✓	✓	-	✓	✓	-	-
MPII [4]	-	410	-	24k	-	-	✓	-	✓	✓	-	-	-	-
COCO [53]	-	-	-	104k	-	-	✓	-	-	✓	-	-	-	-
PoseTrack [3]	-	-	>1.35k	>46k	✓	-	✓	-	-	✓	-	-	-	-
Human3.6M [29]	11	17	839	3.6M	✓	-	✓	✓	✓	✓	✓	-	-	-
CMU Panoptic [35]	8	5	65	154M	✓	-	✓	✓	-	✓	✓	-	-	-
MPI-INF-3DHP [64]	8	8	16	1.3M	✓	-	✓	-	-	✓	✓	-	-	-
3DPW [62]	7	-	60	51k	✓	✓	✓	-	-	-	-	✓	-	-
AMASS [61]	344	-	>11k	>16.88M	✓	-	-	-	-	-	✓	✓	-	-
AIST++ [49]	30	-	1.40k	10.1M	✓	-	✓	-	-	✓	✓	✓	-	-
CAPE [60]	15	-	>600	>140k	✓	-	-	-	✓	-	✓	✓	✓	-
BUFF [107]	6	3	>30	>13.6k	✓	-	✓	✓	✓	-	✓	✓	✓	✓
DFAUST [7]	10	>10	>100	>40k	✓	-	✓	✓	✓	✓	✓	✓	✓	✓
HUMBI [103]	772	-	-	∼26M	✓	-	✓	-	-	✓	✓	✓	✓	✓
ZJU LightStage [78]	6	6	9	>1k	✓	-	✓	-	✓	✓	✓	✓	✓	✓
THuman2.0 [101]	200	-	-	>500	-	-	-	-	-	-	-	✓	✓	✓
HuMMan (ours)	1000	500	400k	60M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

In summary, HuMMan is a large-scale multi-modal dataset for 4D (spatio-temporal) human sensing and modeling, with four main features: **1)** multi-modal data and annotations; **2)** mobile device included in the sensor suite; **3)** action set with atomic motions; **4)** standard benchmarks for multiple vision tasks. We hope HuMMan would pave the way towards more comprehensive sensing and modeling of humans.

2 Related Works

Action Recognition. As an important step towards understanding human activities, action recognition is the task to categorize human motions into predefined classes. RGB videos [16, 17, 91, 92] with additional information such as optical flow and estimated poses and 3D skeletons typically obtained from RGB-D sequences [85, 86, 100, 105] are the common input to existing methods. Datasets for RGB video-based action recognition are often collected from the Internet. Some have a human-centric action design [15, 21, 39, 46, 84, 87] whereas others introduce interaction and diversity in the setup [11, 67, 111]. Recently, fine-grained action understanding [15, 21, 84] is drawing more research attention. However,

these 2D datasets lack 3D annotations. As for RGB-D datasets, earlier works are small in scale [27, 50, 96]. As a remedy, the latest NTU RGB-D series [55, 83, 93] features 60-120 actions. However, the majority of the actions are focused on the upper body. We develop a larger and more complete action set in HuMMan.

2D and 3D Keypoint Detection. Estimation of a human pose is a vital task in computer vision, and a popular pose representation is human skeletal keypoints. The field is categorized by output format: 2D [12, 47, 71, 88] and 3D [63, 77, 104–106, 112] keypoint detection, or by the number of views: single-view [12, 63, 71, 77, 88, 105, 112] and multi-view pose estimation [28, 30, 80]. For 2D keypoint detection, single-frame datasets such as MPII [4] and COCO [53] provide diverse images with 2D keypoints annotations, whereas video datasets such as J-HMDB [32], Penn Action [108] and PoseTrack [3] provide sequences of 2D keypoints. However, they lack 3D ground truths. In contrast, 3D keypoint datasets are typically built indoor data to accommodate sophisticated equipment, such as Human3.6M [29], CMU Panoptic [35], MPI-INF-3DHP [64], TotalCapture [94], and AIST++ [49]. Compared to these datasets, HuMMan not only supports 2D and 3D keypoint detection but also textured mesh reconstruction assist in more holistic modeling of humans.

3D Parametric Human Recovery. Also known as human pose and shape estimation, 3D parametric human recovery leverages human parametric model representation (such as SMPL [58], SMPL-X [75], STAR [73] and GHUM [99]) that achieves sophisticated mesh reconstruction with a small amount of parameters. Existing methods take keypoints [6, 75, 109], images [20, 22, 44, 45, 48, 72, 76], videos [13, 37, 59, 65, 68, 89], and point clouds [5, 33, 54, 97] as the input to obtain the parameters. Joint limits [1] and contact [69] are also important research topics. Apart from those that provide keypoints, various datasets also provide ground-truth SMPL parameters. MoSh [57] is applied on Human3.6M [29] to generate SMPL annotations. CMU Panoptic [35] and HUMBI [103] leverages keypoints from multiple camera views. 3DPW [62] combines a mobile phone and inertial measurement units (IMUs). Synthetic dataset such as AGORA [74] renders high-quality human scans in virtual environments and fits SMPL to the original mesh. Video games have also become an alternative source of data [9, 10]. In addition to SMPL parameters that do not model clothes or texture, HuMMan also provides textured meshes of clothed subjects.

Textured Mesh Reconstruction. To reconstruct the 3D surface, common methods include multi-view stereo [18], volumetric fusion [31, 70, 102], Poisson surface reconstruction [40, 41], and neural surface reconstruction [79, 82]. To reconstruct texture for the human body, popular approaches include texture mapping or montage [19], deep neural rendering [56], deferred neural rendering [90], and NeRF-like methods [66]. Unfortunately, existing datasets for textured human mesh reconstruction typically provide no sequential data [101, 113], which is valuable to the reconstruction of animatable avatars [81, 98]. Moreover, many have only a limited number of subjects [2, 7, 23–25, 60, 78, 107]. In contrast, HuMMan includes diverse subjects with high-resolution body scans and a large amount of dynamic 3D sequences.

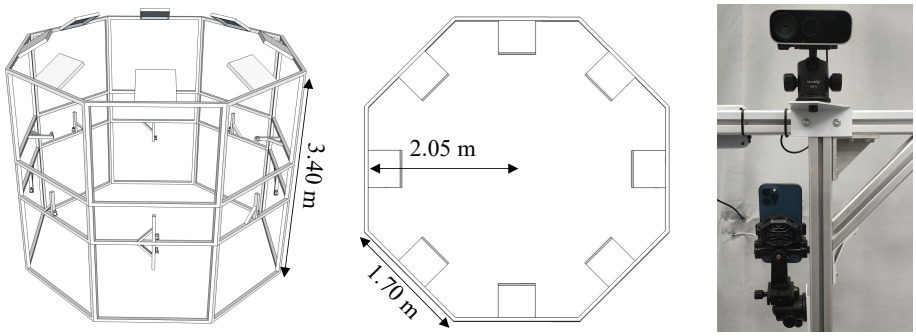


Fig. 2: Hardware setup. a) and b) we build a octagonal prism-shaped framework to accommodate the data collection system. c) sensors used to collect sequential data include ten Azure Kinects and an iPhone 12 Pro Max. Besides, an Artec Eva is used to produce high-resolution static scans of the subjects.

3 Hardware Setup

We customize an octagonal prism-shaped multi-layer framework to accommodate calibrated and synchronized sensors. The system is 1.7 m in height and 3.4 m in side length of its octagonal cross-section as illustrated in Fig. 2.

3.1 Sensors

RGB-D Sensors. Azure Kinect is popular with both academia and the industry with a color resolution of 1920×1080, and a depth resolution of 640×576. We deploy ten Kinects to capture multi-view RGB-D sequences. The Kinects are strategically placed to ensure a uniform spacing, and a wide coverage such that any body part of the subject, even in most expressive poses, is visible to at least two sensors. We develop a program that interfaces with Kinect's SDK to obtain a data throughput of 74.4 MB per frame and 2.2 GB per second at 30 FPS before data compression.

Mobile Device. An iPhone 12 Pro Max is included in the sensor suite to allow for the study on a mobile device. Besides the regular color images of resolution 1920×1440, the built-in LiDAR produces depth maps of resolution 256×192. We develop an iOS app upon ARKit to retrieve the data.

High-Resolution Scanner. To supplement our sequential data with high-quality body shape information, a professional handheld 3D scanner, Artec Eva, is used to produce a body scan of resolution up to 0.2 mm and accuracy up to 0.1 mm. A typical scan consists of $300k$ to $500k$ faces and $100k$ to $300k$ vertices, with a 4K (4096×4096) resolution texture map.

3.2 Two-Stage Calibration

Image-based Calibration. To obtain a coarse calibration, we first perform image-based calibration following the general steps in Zhang's method [110]. However, we highlight that Kinect's active IR depth cameras encounter over-exposure with regular chessboards. Hence, we customize a light absorbent material to cover the black squares of the chessboard pattern. In this way, we acquire reasonably accurate extrinsic calibration for Kinects and iPhones.
Geometry-based Calibration. Image-based calibration is unfortunately not accurate enough to reconstruct good-quality mesh. Hence, we propose to take advantage of the depth information in a geometry-based calibration stage. We empirically verify that image-based calibration serves as a good initialization for geometry-based calibration. Hence, we randomly place stacked cubes inside the framework. After that, we convert captured depth maps to point clouds and apply multi-way ICP registration [14] to refine the calibration.

3.3 Synchronization

Kinects. As the Azure Kinect implements the Time-of-Flight principle, it actively illuminates the scene multiple times (nine exposures in our system) for depth computation. To avoid interference between individual sensors, we use the synchronization cables to propagate a unified clock in a daisy chain fashion, and reject any image that is 33 ms or above out of synchronization. We highlight that there is only a 1450-us interval between exposures of 160 us; our system of ten Kinects reaches the theoretical maximum number.
Kinect-iPhone. Due to hardware limitations, we cannot apply the synchronization cable to the iPhone. We circumvent this challenge by implementing a TCP-based communication protocol that computes an offset between the Kinect clock and the iPhone ARKit clock. As iPhone is recording at 60 FPS, we then use the offset to map the closest iPhone frames to Kinect frames. Our test shows the synchronization error is constrained below 33 ms.

4 Toolchain

To handle the large volume of data, we develop an automatic toolchain to provide annotations such as keypoints and SMPL parameters. Moreover, dynamic sequences of textured mesh are also reconstructed. The pipeline is illustrated in Fig. 3. Note that there is a human inspection stage to reject low-quality data with erroneous annotations.

4.1 Keypoint Annotation

There are two stages of keypoint annotation (I and II) in the toolchain. For stage I, virtual cameras are placed around the minimally clothed body scan to render multi-view images. For stage II, the color images from multi-view RGB-D are

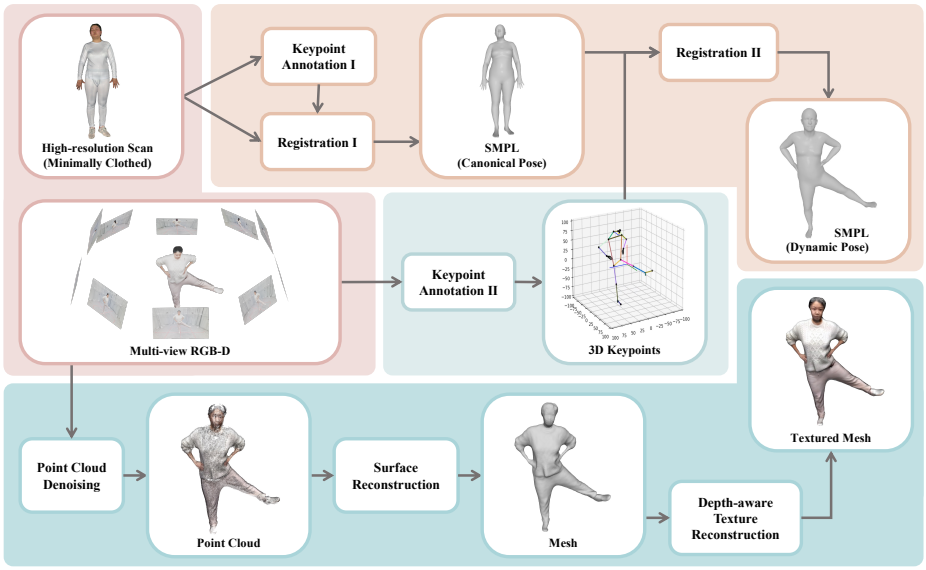


Fig. 3: Our toolchain produces multiple annotation formats such as 3D keypoint sequences, SMPL parameter sequences, and textured mesh sequences

used. The core ideas of the keypoint annotation are demonstrated below, with the detailed algorithm in the Supplementary Material.

Multi-view 2D Keypoint Detection. We employ the whole-body pose model that includes body, hand and face 2D keypoints $\hat{\mathcal{P}}_{2D} \in \mathbb{R}^{P\times 2}$, where $P = 133$. A large deep learning model HRNet-w48 [88] is used which achieves AP 66.1 and AR 74.3 on COCO whole-body benchmark [34].

3D Keypoint Triangulation. As the camera intrinsic and extrinsic parameters are available, we triangulate 3D keypoints $\mathcal{P}_{3D} \in \mathbb{R}^{P\times 3}$ with the multi-view 2D estimated keypoints $\hat{\mathcal{P}}_{2D}$. However, 2D keypoints from any single view may not be always reliable. Hence, we use the following strategies to improve the quality of 3D keypoints. 1) *Keypoint selection.* To avoid the influence of poor-quality estimated 2D keypoints, we use a threshold τ_k to remove keypoints with a low confidence score. 2) *Camera selection.* As our system consists of ten Kinects, we exploit the redundancy to remove low-quality views. We only keep camera views with reprojection errors that are top-k smallest [38] and no larger than a threshold τ_c. 3) *Smoothness constraint.* Due to inevitable occlusion in the single view, the estimated 2D keypoints often have jitters. To alleviate the issue, we develop a smoothness loss to minimize the difference between consecutive triangulated 3D keypoints. Note that we design the loss weight to be inversely proportional to average speed, in order to remove jitters without compromising the ability to capture fast body motions. 4) *Bone length constraint.* As human bone length is constant, the per-frame bone length is constrained towards the median bone length $\mathcal{B}$ pre-computed from the initial triangulated 3D keypoints.

Fig. 4: HuMMan provides synchronized sequences of multiple data formats and annotations. Here we demonstrate textured mesh sequences and SMPL parameter sequences

The constraints are formulated as Eq. 1:

$$E_{tri} = \lambda_1 \sum_{t=0}^{T-1} \|\mathcal{P}_{3D}(t+1) - \mathcal{P}_{3D}(t)\| + \lambda_2 \sum_{(i,j)\in\mathcal{I}_\mathcal{B}} \|\mathcal{B}_{i,j} - f_\mathcal{B}(\mathcal{P}_{3D}(i,j))\| \quad (1)$$

where $\mathcal{I}_\mathcal{B}$ contains the indices of connected keypoints and $f_\mathcal{B}(\cdot)$ calculates the average bone length of a given 3D keypoint sequence. Note that 3) and 4) are jointly optimized.

2D Keypoint Projection. To obtain high-quality 2D keypoints $\mathcal{P}_{2D} \in \mathbb{R}^{P\times 2}$, we project the triangulated 3D keypoints to image space via calibrated camera parameters. Note that this step is only needed for stage II keypoint annotation.

Keypoint Quality. We use $\mathcal{P}_{2D}$ and $\mathcal{P}_{3D}$ as keypoint annotations for 2D Pose Estimation and 3D Pose Estimation, respectively. To gauge the accuracy of the automatic keypoint annotation pipeline, we manually annotate a subset of data. The average Euclidean distance between annotated 2D keypoints and reprojected 2D keypoints $\mathcal{P}_{2D}$ is 15.13 pixels on the resolution of 1920×1080.

4.2 Human Parametric Model Registration

We select SMPL [58] as the human parametric model for its popularity. There are two stages of registration (I and II). Stage I is used to obtain accurate shape parameters from the static high-resolution scan, whereas stage II is used to obtain pose parameters from the dynamic sequence, with shape parameters from stage I. The registration is formulated as an optimization task to obtain SMPL pose parameters $\theta \in \mathbb{R}^{n\times 72}$, shape parameters $\beta \in \mathbb{R}^{n\times 10}$ (stage I only) and translation parameters $t \in \mathbb{R}^{n\times 3}$ where n is the number of frames ($n = 1$ for stage I), with the following energy terms and constraints. We show a sample sequence of SMPL models with reconstructed textured mesh in Fig. 4.

Keypoint Energy. SMPLify [6] estimates camera parameters to leverage 2D keypoint supervision, which may be prone to depth and scale ambiguity. Hence,

Fig. 5: Examples of SMPL registered on high-resolution static body scans for accurate shape parameters. The subjects are instructed to wear tight clothes for this scan. Note that each subject has another naturally clothed scan

we develop the keypoint energy on 3D keypoints. For simplicity, we denote P_{3D} as P, the global rigid transformation derived from the SMPL kinematic tree as $\mathcal{T}$, the joint regressor as $\mathcal{J}$. We formulate the energy term:

$$E_{\mathcal{P}}(\theta, \beta, t) = \frac{1}{|\mathcal{P}|} \sum_{i}^{|\mathcal{P}|} \|\mathcal{T}(\mathcal{J}(\beta)_i, \theta), t) - \mathcal{P}_i\| \tag{2}$$

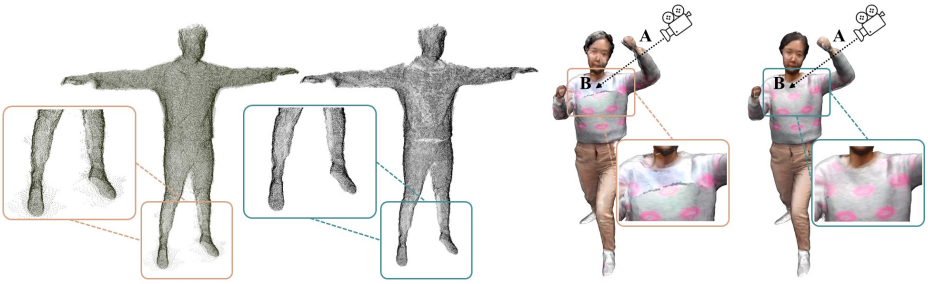
Surface Energy. To supplement 3D keypoints that do not provide sufficient constraint for shape parameters, we add an additional surface energy term for registration on the high-resolution minimally clothed scans in stage I only. We use bi-directional Chamfer distance to gauge the difference between two mesh surfaces:

$$E_S = \frac{1}{|\mathcal{V}_H|} \sum_{v_H \in \mathcal{V}_H} \min_{v_S \in \mathcal{V}_S} \|v_H - v_S\| + \frac{1}{|\mathcal{V}_S|} \sum_{v_S \in \mathcal{V}_S} \min_{v_H \in \mathcal{V}_H} \|v_H - v_S\| \tag{3}$$

where $\mathcal{V}_H$ and $\mathcal{V}_S$ are the mesh vertices of the high-resolution scan and SMPL.
Shape Consistency. Unlike existing work [74] that enforces an inter-beta energy term due to the lack of minimally clothed scan of each subject, we obtain accurate shape parameters from the high-resolution scan that allow us to apply constant beta parameters in the registration in stage II.
Full-body Joint Angle Prior. Joint rotation limitations serve as an important constraint to prevent unnaturally twisted poses. We extend existing work [6, 75] that only applies constraints on elbows and knees to all $J = 23$ joints in SMPL. The constraint is formulated as a strong penalty outside the plausible rotation range (with more details included in the Supplementary Material):

$$E_a = \frac{1}{J \times 3} \sum_{j}^{J \times 3} exp(\max(\theta_i - \theta_i^u, 0) + \max(\theta_i^l - \theta_i, 0)) - 1 \tag{4}$$

where θ_i^u and θ_i^l are the upper and lower limit of a rotation angle. Note that each joint rotation is converted to three Euler angles which can be interpreted as a series of individual rotations to decouple the original axis-angle representation.



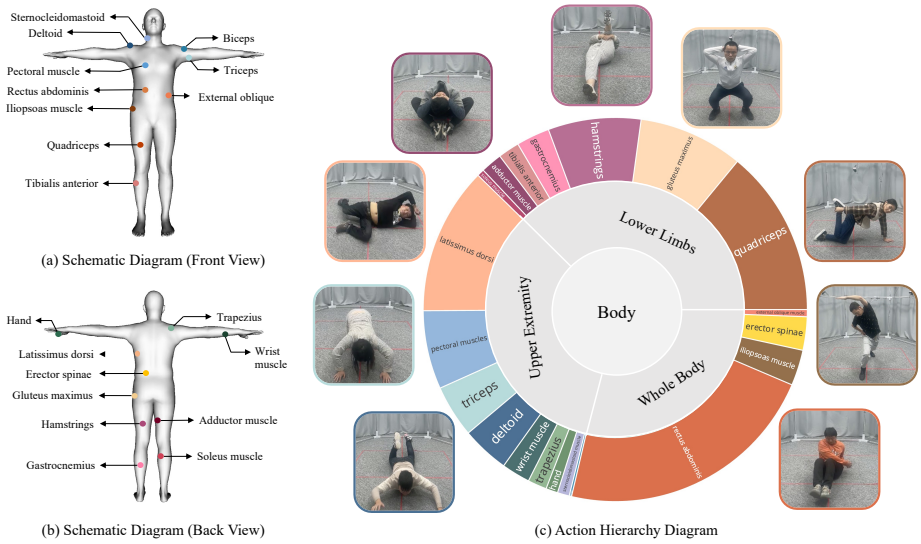
a) Point Cloud Denoising b) Depth-aware Texture Reconstruction

Fig. 6: Key steps to textured mesh reconstruction. a) Point cloud denoising removes noisy points. b) Depth-aware texture reconstruction prevents texture miss projection artifacts (such as projecting texture at point A to point B) due to misalignment between the actual subject and the reconstructed geometry

4.3 Textured Mesh Reconstruction

Point Cloud Reconstruction and Denoising. We convert depth maps to point clouds and transform them into a world coordinate system with camera extrinsic parameters. However, the depth images captured by Kinect contain noisy pixels, which are prominent at subject boundaries where the depth gradient is large. To solve this issue, we first generate a binary boundary mask through edge finding with Laplacian of Gaussian Filters. Since our cameras have highly overlapped views to supplement points for one another, we apply a more aggressive threshold to remove boundary pixels. After the point cloud is reconstructed from the denoised depth images, we apply Statistical Outlier Removal [26] to further remove sprinkle noises.

Geometry and Depth-aware Texture Reconstruction. With complete and dense point cloud reconstructed, we apply Poisson Surface Reconstruction with envelope constraints [42] to reconstruct the watertight mesh. However, due to inevitable self-occlusion in complicated poses, interpolation artifacts arise from missing depth information, which leads to a shrunk or a dilated geometry. These artifacts are negligible for geometry reconstruction. However, a prominent artifact appears when projecting a texture onto the mesh even if the inconsistency between the true surface and the reconstructed surface is small. Hence, we extend MVS-texturing [95] to be depth-aware in texture reconstruction. We render the reconstructed mesh back into the camera view and compare the rendered depth map with the original depth map to generate the difference mask. We then mask out all the misalignment regions where the depth difference exceeds a threshold τ_d. The masked regions do not contribute to texture projection. As shown in Fig. 6(b), the depth-aware texture reconstruction is more accurate and visually pleasing.



(a) Schematic Diagram (Front View)

(b) Schematic Diagram (Back View)

(c) Action Hierarchy Diagram

Fig. 7: Schematic diagram of muscles from a) front and b) back views. c) HuMMan categorizes 500 actions hierarchically, first by body parts to achieve *complete* body coverage, then by driving muscles for *unambiguous* action definition

5 Action Set

Understanding human actions is a long-standing computer vision task. In this section, we elaborate on the two principles, following which we design the action set of 500 actions: *completeness* and *unambiguity*. More details are included in the Supplementary Material.

Completeness. We build the action set to cover plausible human movements as much as possible. Compared to the popular 3D action recognition dataset NTU-RGBD-120 [55] whose actions are focused on upper body movements, we employ a hierarchical design to first divide possible actions into upper extremity, lower limbs, and whole-body movements. Such design allows us to achieve a balance between various body parts instead of over-emphasizing a specific group of movements. Note that we define whole body movements to be actions that require multiple body parts to collaborate, including different poses of the body trunk (*e.g.* lying down and sprawling). Fig. 7(c) demonstrates the action hierarchy and examples of interesting actions that are vastly diverse.

Unambiguity. Instead of providing a general description of the motions [11, 29, 39, 62, 64, 67, 87], we argue that the action classes should be clearly defined and are easy to identify and reproduce. Inspired by the fact that all human actions are the result of muscular contractions, we propose a *muscle-driven* strategy to systematically design the action set from the perspective of human anatomy. As illustrated in Fig. 7(a)(b), 20 major muscles are identified by professionals in fitness and yoga training, who then put together a list of standard movements as-

Fig. 8: HuMMan contains 1000 subjects with diverse appearances. For each subject, a naturally clothed high-resolution scan is obtained

sociated with these muscles. Moreover, we cross-check with the action definitions from existing datasets [8, 11, 15, 21, 35, 39, 51, 55] to ensure a wide coverage.

6 Subjects

HuMMan consists of 1000 subjects with a wide coverage of genders, ages, body shapes (heights, weights), and ethnicity. The subjects are instructed to wear their personal daily clothes to achieve a large collection of natural appearances. We demonstrate examples of high-resolution scans of the subjects in Fig. 8. We include statistics in the Supplementary Material.

7 Experiments

In this section, we evaluate popular methods from various research fields on HuMMan. To constrain the training within a reasonable computation budget, we sample 10% of data and split them into training and testing sets for both Kinects and iPhone. The details are included in the Supplementary Material.

Action Recognition. HuMMan provides action labels and 3D skeletal positions, which can verify its usefulness on 3D action recognition. Specifically, we train popular graph-based methods (STGCN [100] and 2s-AGCN [85]) on HuMMan. Results are shown in Table 2. Compared to NTU RGB+D, a large-scale 3D action recognition dataset and a standard benchmark that contains 120 actions [55], HuMMan may be more challenging since 2s-AGCN [85] achieves Top-1 accuracy of 88.9% and 82.9% on NTU RGB+D 60 and 120 respectively, but 74.1% only on HuMMan. The difficulties come from

Table 2: **Action Recognition**

Method	Top-1 (%)↑	Top-5 (%)↑
ST-GCN	72.5	94.3
2s-AGCN	74.1	95.4

the whole-body coverage design in our action set, instead of over-emphasis on certain body parts (*e.g.* NTU RGB+D has a large proportion of upper body movements). Moreover, we observe a significant gap between Top-1 and Top-5 accuracy (∼30%). We attribute this phenomenon to the fact that there are plenty of *intra-actions* in HuMMan. For example, there are similar variants of push-ups such as quadruped push-ups, kneeling push-ups, and leg push-ups. This challenges the model to pay more attention to the fine-grained differences in these actions. Hence, we find HuMMan would serve as an indicative benchmark for fine-grained action understanding.

3D Keypoint Detection. With the well-annotated 3D keypoints, HuMMan supports 3D keypoint detection. We employ popular 2D-to-3D lifting backbones [63, 77] as single-frame and multi-frame baselines on HuMMan. We experiment with different training and test settings to obtain the baseline results in Table 3. First, in-domain training and testing on HuMMan are provided. The values are slightly higher than the same baselines on Human3.6M [29] (on which FCN obtains MPJPE of 53.4 mm). Second, methods trained on HuMMan tend to generalize better than on Human3.6M. This may be attributed to HuMMan's diverse collection of subjects and actions.

Table 3: **3D Keypoint Detection**. PA: PA-MPJPE. Row 1-3: FCN [63]; Row 4-6: Video3D [77]

Train	Test	MPJPE ↓	PA ↓
HuMMan	HuMMan	78.5	46.3
H36M	AIST++	133.9	73.1
HuMMan	AIST++	116.4	67.2
HuMMan	HuMMan	73.1	43.5
H36M	AIST++	128.5	72.0
HuMMan	AIST++	109.2	63.5

3D Parametric Human Recovery. HuMMan provides SMPL annotations, RGB and RGB-D sequences. Hence, we evaluate HMR [36], not only one of the first deep learning approaches towards 3D parametric human recovery but a fundamental component for follow-up works [43, 45], to represent image-based methods. In addition, we employ VoteHMR [54], a recent work that takes point clouds as the input. In Table 4, we find that HMR has achieved low MPJPE and PA-MPJPE, which may be attributed to the clearly defined action set and the training set already includes all action classes. However, VoteHMR is not performing well. We argue that existing point cloud-based methods [33,54,97] rely heavily on synthetic data for training and evaluation, whereas HuMMan provides genuine point clouds from commercial RGB-D sensors that remain challenging.

Table 4: **3D Parametric Human Recovery**. Image- and point cloud-based methods are evaluated

Method	MPJPE ↓	PA ↓
HMR	54.78	36.14
VoteHMR	144.99	106.32

Textured Mesh Reconstruction. We gauge mesh geometry reconstruction quality of PIFu, PIFuHD, and Function4D (F4D) in Table 5 with Chamfer distance (CD) as the metric. Note that benefiting from the multi-modality signals, HuMMan supports a wide range of surface reconstruction methods that leverage various input types like PIFu [82] (RGB-only), 3D Self-Portrait [52] (single-view RGBD video), and CON [79] (multi-view depth point cloud).

Table 5: **Geometry Reconstruction**

Method	PIFu	PIFuHD	F4D
CD (10^{-2} m)	7.92	7.73	1.80

Mobile Device. It is under-explored that if model trained with the regular device is readily transferable to the mobile device. In Table 6, we study the performance gaps across devices. For the image-based method, we find that there exists a considerable domain gap across devices, despite that they have similar resolutions. Moreover, for the point cloud-based method, the domain gap is much more significant as the mobile device tends to have much sparser point clouds as a result of lower depth map resolution. Hence, it remains a challenging problem to transfer knowledge across devices, especially for point cloud-based methods.

Table 6: **Mobile Device**. The models are trained with different training sets, and evaluated on HuMMan iPhone test set. Kin.: Kinect training set. iPh.: iPhone training set.

Method	Kin.	iPh.	MPJPE ↓	PA ↓
HMR	✓	-	97.81	52.74
HMR	-	✓	72.62	41.86
VoteHMR	✓	-	255.71	162.00
VoteHMR	-	✓	83.18	61.69

8 Discussion

We present HuMMan, a large-scale 4D human dataset that features multi-modal data and annotations, inclusion of mobile device, a comprehensive action set, and support for multiple tasks. Our experiments point out interesting directions that await future research, such as fine-grained action recognition, point cloud-based parametric human estimation, dynamic mesh sequence reconstruction, transferring knowledge across devices, and potentially, multi-task joint training. We hope HuMMan would facilitate the development of better algorithms for sensing and modeling humans.

Acknowledgements. This work is supported by NTU NAP, MOE AcRF Tier 2 (T2EP20221-0033), NSFC No.62171255, and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1446–1455 (2015) 4
2. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8387–8397 (2018) 4
3. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5167–5176 (2018) 3, 4
4. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014) 3, 4
5. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Combining implicit function learning and parametric models for 3d human reconstruction. In: European Conference on Computer Vision. pp. 311–329. Springer (2020) 4
6. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European conference on computer vision. pp. 561–578. Springer (2016) 4, 8, 9
7. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic faust: Registering human bodies in motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6233–6242 (2017) 3, 4
8. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015) 12
9. Cai, Z., Zhang, M., Ren, J., Wei, C., Ren, D., Li, J., Lin, Z., Zhao, H., Yi, S., Yang, L., et al.: Playing for 3d human recovery. arXiv preprint arXiv:2110.07588 (2021) 4
10. Cao, Z., Gao, H., Mangalam, K., Cai, Q.Z., Vo, M., Malik, J.: Long-term human motion prediction with scene context. In: European Conference on Computer Vision. pp. 387–404. Springer (2020) 4
11. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019) 3, 11, 12
12. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G.g., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7103–7112 (2018) 4
13. Choi, H., Moon, G., Lee, K.M.: Beyond static features for temporally consistent 3d human pose and shape from a video. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
14. Choi, S., Zhou, Q.Y., Koltun, V.: Robust reconstruction of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5556–5565 (2015) 6
15. Chung, J., Wuu, C.h., Yang, H.r., Tai, Y.W., Tang, C.K.: Haa500: Human-centric atomic action dataset with curated videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13465–13474 (2021) 3, 12

16. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020) 3
17. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019) 3
18. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(8), 1362–1376 (2010). https://doi.org/10.1109/TPAMI.2009.161 4
19. Gal, R., Wexler, Y., Ofek, E., Hoppe, H., Cohen-Or, D.: Seamless montage for texturing models. Computer Graphics Forum **29**(2), 479–486 (2010). https://doi.org/https://doi.org/10.1111/j.1467-8659.2009.01617.x, https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2009.01617.x 4
20. Georgakis, G., Li, R., Karanam, S., Chen, T., Košecká, J., Wu, Z.: Hierarchical kinematic human mesh recovery. In: European Conference on Computer Vision. pp. 768–784. Springer (2020) 4
21. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6047–6056 (2018) 3, 12
22. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10884–10894 (2019) 4
23. Habermann, M., Liu, L., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Real-time deep dynamic characters. ACM Transactions on Graphics (TOG) **40**(4), 1–16 (2021) 4
24. Habermann, M., Xu, W., Zollhoefer, M., Pons-Moll, G., Theobalt, C.: Livecap: Real-time human performance capture from monocular video. ACM Transactions On Graphics (TOG) **38**(2), 1–17 (2019) 4
25. Habermann, M., Xu, W., Zollhofer, M., Pons-Moll, G., Theobalt, C.: Deepcap: Monocular human performance capture using weak supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5052–5063 (2020) 4
26. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial intelligence review **22**(2), 85–126 (2004) 10
27. Hu, J.F., Zheng, W.S., Lai, J., Zhang, J.: Jointly learning heterogeneous features for rgb-d activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5344–5352 (2015) 3, 4
28. Huang, F., Zeng, A., Liu, M., Lai, Q., Xu, Q.: Deepfuse: An imu-aware network for real-time 3d human pose estimation from multi-view image. arXiv preprint arXiv:1912.04071 (2019) 4
29. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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), 1325–1339 (2013) 1, 3, 4, 11, 13
30. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7718–7727 (2019) 4

31. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568 (2011) 4
32. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 3192–3199 (2013) 4
33. Jiang, H., Cai, J., Zheng, J.: Skeleton-aware 3d human shape reconstruction from point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5431–5441 (2019) 4, 13
34. Jin, S., Xu, L., Xu, J., Wang, C., Liu, W., Qian, C., Ouyang, W., Luo, P.: Whole-body human pose estimation in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 7
35. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3334–3342 (2015) 3, 4, 12
36. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7122–7131 (2018) 13
37. Kanazawa, A., Zhang, J.Y., Felsen, P., Malik, J.: Learning 3d human dynamics from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5614–5623 (2019) 4
38. Karashchuk, P., Rupp, K.L., Dickinson, E.S., Walling-Bell, S., Sanders, E., Azim, E., Brunton, B.W., Tuthill, J.C.: Anipose: a toolkit for robust markerless 3d pose estimation. Cell reports **36**(13), 109730 (2021) 7
39. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014) 3, 11, 12
40. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. ACM Trans. Graph. **32**(3) (jul 2013). https://doi.org/10.1145/2487228.2487237, https://doi.org/10.1145/2487228.2487237 4
41. Kazhdan, M., Chuang, M., Rusinkiewicz, S., Hoppe, H.: Poisson surface reconstruction with envelope constraints. Computer Graphics Forum (Proc. Symposium on Geometry Processing) **39**(5) (Jul 2020) 4
42. Kazhdan, M., Chuang, M., Rusinkiewicz, S., Hoppe, H.: Poisson surface reconstruction with envelope constraints. In: Computer graphics forum. vol. 39, pp. 173–182. Wiley Online Library (2020) 10
43. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5253–5263 (2020) 13
44. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. arXiv preprint arXiv:2104.08527 (2021) 4
45. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2252–2261 (2019) 4, 13

46. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011) 3
47. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: ICCV (2021) 4
48. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: CVPR. pp. 3383–3393. Computer Vision Foundation / IEEE (2021) 4
49. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021) 3, 4
50. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. pp. 9–14. IEEE (2010) 4
51. Li, Y.L., Liu, X., Wu, X., Li, Y., Qiu, Z., Xu, L., Xu, Y., Fang, H.S., Lu, C.: Hake: A knowledge engine foundation for human activity understanding (2022) 12
52. Li, Z., Yu, T., Zheng, Z., Liu, Y.: Robust and accurate 3d self-portraits in seconds. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–1 (2021). https://doi.org/10.1109/TPAMI.2021.3113164 14
53. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 1, 3, 4
54. Liu, G., Rong, Y., Sheng, L.: Votehmr: Occlusion-aware voting network for robust 3d human mesh recovery from partial point clouds. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 955–964 (2021) 4, 13
55. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019) 3, 4, 11, 12
56. Lombardi, S., Saragih, J., Simon, T., Sheikh, Y.: Deep appearance models for face rendering **37**(4) (jul 2018). https://doi.org/10.1145/3197517.3201401, https://doi.org/10.1145/3197517.3201401 4
57. Loper, M., Mahmood, N., Black, M.J.: Mosh: Motion and shape capture from sparse markers. ACM Transactions on Graphics (TOG) **33**(6), 1–13 (2014) 4
58. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34**(6), 1–16 (2015) 4, 8
59. Luo, Z., Golestaneh, S.A., Kitani, K.M.: 3d human motion estimation via motion compression and refinement. In: Proceedings of the Asian Conference on Computer Vision (2020) 4
60. Ma, Q., Yang, J., Ranjan, A., Pujades, S., Pons-Moll, G., Tang, S., Black, M.J.: Learning to dress 3d people in generative clothing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6469–6478 (2020) 1, 3, 4
61. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5442–5451 (2019) 3
62. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera.

In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 601–617 (2018) 1, 3, 4, 11
63. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649 (2017) 4, 13
64. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017) 3, 4, 11
65. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: Real-time multi-person 3d motion capture with a single rgb camera. ACM Transactions on Graphics (TOG) **39**(4), 82–1 (2020) 4
66. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 4
67. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence **42**(2), 502–508 (2019) 3, 11
68. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single RGB image. In: ECCV (7). Lecture Notes in Computer Science, vol. 12352, pp. 752–768. Springer (2020) 4
69. Muller, L., Osman, A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9990–9999 (2021) 4
70. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015) 4
71. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016) 4
72. Omran, M., Lassner, C., Pons-Moll, G., Gehler, P., Schiele, B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). pp. 484–494. IEEE (2018) 4
73. Osman, A.A.A., Bolkart, T., Black, M.J.: STAR: sparse trained articulated human body regressor. In: ECCV (6). Lecture Notes in Computer Science, vol. 12351, pp. 598–613. Springer (2020) 4
74. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13468–13478 (2021) 4, 9
75. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10975–10985 (2019) 4, 9
76. Pavlakos, G., Zhu, L., Zhou, X., Daniilidis, K.: 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. pp. 459–468 (2018) 4

77. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019) 4, 13
78. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021) 3, 4
79. Peng, S., Niemeyer, M., Mescheder, L., Pollefeys, M., Geiger, A.: Convolutional occupancy networks. In: ECCV (2020) 4, 14
80. Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W.: Cross view fusion for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4342–4351 (2019) 4
81. Raj, A., Tanke, J., Hays, J., Vo, M., Stoll, C., Lassner, C.: Anr-articulated neural rendering for virtual avatars. In: arXiv:2012.12890 (2020) 4
82. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2304–2314 (2019) 4, 14
83. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016) 3, 4
84. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2616–2625 (2020) 3
85. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. arXiv preprint arXiv:1912.06971 (2019) 3, 12
86. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. IEEE Transactions on Image Processing **29**, 9532–9545 (2020) 3
87. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012) 3, 11
88. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703 (2019) 4, 7
89. Sun, Y., Ye, Y., Liu, W., Gao, W., Fu, Y., Mei, T.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5349–5358 (2019) 4
90. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Trans. Graph. **38**(4) (jul 2019). https://doi.org/10.1145/3306346.3323035, https://doi.org/10.1145/3306346.3323035 4
91. Tran, D., Wang, H., Torresani, L., Feiszli, M.: Video classification with channel-separated convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5552–5561 (2019) 3
92. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018) 3

93. Trivedi, N., Thatipelli, A., Sarvadevabhatla, R.K.: Ntu-x: An enhanced large-scale dataset for improving pose-based recognition of subtle human actions. arXiv preprint arXiv:2101.11529 (2021) 3, 4
94. Trumble, M., Gilbert, A., Malleson, C., Hilton, A., Collomosse, J.P.: Total capture: 3d human pose estimation fusing video and inertial sensors. In: BMVC. vol. 2, pp. 1–13 (2017) 4
95. Waechter, M., Moehrle, N., Goesele, M.: Let there be color! — Large-scale texturing of 3D reconstructions. In: Proceedings of the European Conference on Computer Vision. Springer (2014) 10
96. Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C.: Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2649–2656 (2014) 4
97. Wang, S., Geiger, A., Tang, S.: Locally aware piecewise transformation fields for 3d human mesh registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7639–7648 (2021) 4, 13
98. Xiang, D., Prada, F., Bagautdinov, T., Xu, W., Dong, Y., Wen, H., Hodgins, J., Wu, C.: Modeling clothing as a separate layer for an animatable human avatar. ACM Trans. Graph. **40**(6) (dec 2021). https://doi.org/10.1145/3478513.3480545, https://doi.org/10.1145/3478513.3480545 4
99. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Ghum & ghuml: Generative 3d human shape and articulated pose models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6184–6193 (2020) 4
100. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. arXiv preprint arXiv:1801.07455 (2018) 3, 12
101. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021) 3, 4
102. Yu, T., Zheng, Z., Guo, K., Zhao, J., Dai, Q., Li, H., Pons-Moll, G., Liu, Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In: IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 7287–7296. IEEE, Salt Lake City (June 2018) 4
103. Yu, Z., Yoon, J.S., Lee, I., Venkatesh, P., Park, J., Yu, J., Park, H.: Humbi: A large multiview dataset of human body expressions. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2987–2997 (2020) 1, 3, 4
104. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.C.F.: Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In: ECCV (2020) 4
105. Zeng, A., Sun, X., Yang, L., Zhao, N., Liu, M., Xu, Q.: Learning skeletal graph neural networks for hard 3d pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (2021) 3, 4
106. Zeng, A., Yang, L., Ju, X., Li, J., Wang, J., Xu, Q.: Smoothnet: A plug-and-play network for refining human poses in videos. arXiv preprint arXiv:2112.13715 (2021) 4
107. Zhang, C., Pujades, S., Black, M.J., Pons-Moll, G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4191–4200 (2017) 3, 4

108. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2248–2255 (2013) 4
109. Zhang, Y., Li, Z., An, L., Li, M., Yu, T., Liu, Y.: Lightweight multi-person total motion capture using sparse multi-view cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5560–5569 (October 2021) 4
110. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence **22**(11), 1330–1334 (2000) 6
111. Zhao, H., Torralba, A., Torresani, L., Yan, Z.: Hacs: Human action clips and segments dataset for recognition and temporal localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8668–8678 (2019) 3
112. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3425–3435 (2019) 4
113. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: Deephuman: 3d human reconstruction from a single image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 1, 4