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

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

A Overview

We provide additional details of data collection (Section B), hardware (Section C), toolchain (Section D), action set (Section E), subjects (Section F), experiments (Section G), and a more complete dataset comparison (Section H).

B Additional Details of Data Collection

The data collection has two stages for each subject. **1)** each subject receives two high-resolution scans, one with natural clothes on and the other with a tight-fitting suit on, both captured by the Artex Eva 3D Scanner. To ensure the high quality of the scans, the subjects are instructed to stand in a special pose (the *canonical pose*) on a turntable, that allows for a 360-degree full-body scanning with minimal self-occlusion. Each high-resolution scan includes an MTL information file, an OBJ mesh file, and a BMP texture file. **2)** After that static body scanning, the subject enters the framework and follows instructions to perform 40-60 actions, randomly sampled from the action set that contains 500 actions. Each action that a subject performs is a *sequence*, that consists of ten Kinect RGB-D sequences and an iPhone RGB-D sequence. We show sample frames collected with our hardware setup in Fig. 1. Each sequence takes 5-15 seconds and 150-450 frames at 30 FPS per view. We compress all sequential data in a custom data format *SMC* that is developed based on HDF5 format. The SMC file also contains additional information such as camera parameters, subject ID, and action ID.

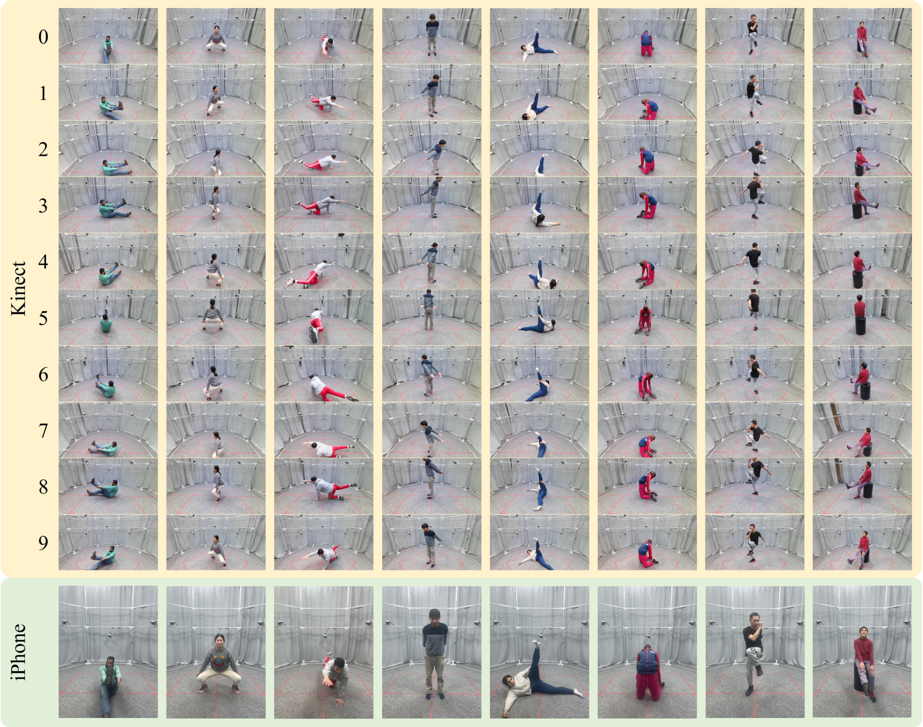

Fig. 1: HuMMan deploys ten Azure Kinects and an iPhone 12 Pro Max for multi-view sequential data collection. We show several synchronized RGB frames captured with our hardware setup. The numbers are device IDs

C Additional Details of Hardware

C.1 Sensors

We provide more details on the RGB-D sensor (Azure Kinect). We set operating mode to *NFOV unbinned* for the depth cameras, which results in the largest view overlap with the color camera and the densest point clouds. The depth camera in this mode has an FOV of $90° \times 59°$. The operating range of the depth sensor in this mode is between 0.5 m to 3.86 m. The typical systematic error of the depth sensor is less than 11 mm + 0.1% of distance with a standard deviation of less than 17 mm. In view of the limited FOV and depth error-distance relation, we design our aluminum framework such that the subject is around 2 m away from the Kinects: at that distance, the FOV can accommodate the subject's whole body, without incurring any extra depth error.

* co-first authors; † co-corresponding authors

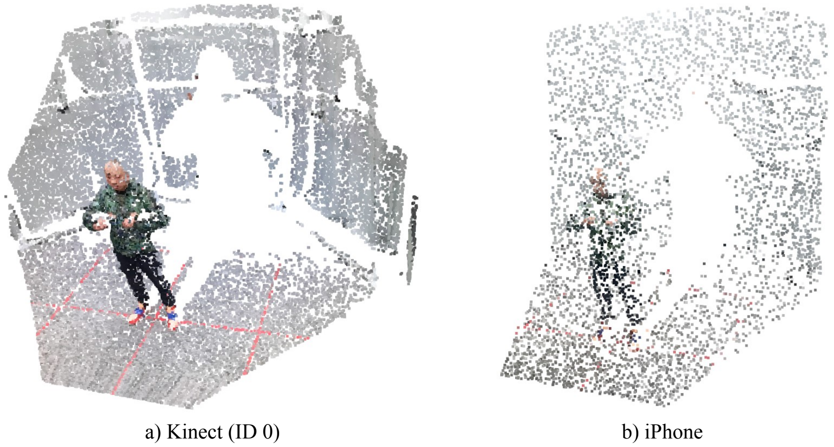

Fig. 2: The point clouds produced by the Kinect and the iPhone are different: the latter is significantly sparser. Note that the point clouds shown here are raw (not filtered or denoised). For visual comparison purpose, both point clouds are downsampled by the same factor of 10

C.2 Synchronization

Our data sampling program runs on a workstation, and it 1) integrates the Kinect SDK, and 2) communicates with the iPhone app developed based on ARKit through TCP. Since there is no existing hardware approach to Kinect-iPhone synchronization, we develop a method to compute the difference between Kinect clock and iPhone ARKit clock $t_{K\to A}$. Hence, we first obtain the offset from the workstation to the Kinects $t_{K\to W}$ as

$$t_{K\to W} = t_W - t_K$$

where t_K is the Kinect clock time and t_W is the workstation's system time, obtained at the same moment. We also send a message to the iPhone app, which records down the iPhone system clock t_I upon receiving the message and sends back a message to the workstation to complete a round trip. We compute the offset from the iPhone system clock to the workstation system clock $t_{W\to I}$ as

$$t_{W\to I} = t_I - t_W - \frac{t_{round}}{2}$$

where t_{round} is the round trip time taken. Note that there is an additional offset between the ARKit clock and the iPhone system clock $t_{I\to A}$, computed as

$$t_{I\to A} = t_A - t_I$$

where t_A is the ARKit clock. Finally, the required clock difference $t_{K\to A}$ is

$$t_{K\to A} = t_{K\to W} + t_{W\to I} + t_{I\to A}$$

C.3 Point Clouds

Both Kinect and iPhone produce depth maps that can be converted to point clouds. However, iPhone's point cloud is much sparser than Kinect's. We show unprocessed raw point clouds produced by the two types of sensors in Fig. 2. In addition, iPhone does not report the LiDAR accuracy; we empirically find that iPhone point clouds are noisier, especially at the object boundaries, than Kinect point clouds.

D Additional Details of Toolchain

D.1 Keypoint Annotation

The overall pipeline for keypoint annotation is summarized in Algorithm 1.

Algorithm 1 Keypoint Annotation

Input: Detected 2D Keypoints $\hat{\mathcal{P}}_{2D}$, camera parameters set $\mathcal{C}$, keypoint threshold τ_k, reprojection minimal threshold τ_{min}, reprojection maximum threshold τ_{max}, camera threshold step Δ_c, best camera number N_c.

Output: 3D Keypoints $\mathcal{P}_{3D}$, 2D Keypoints $\mathcal{P}_{2D}$

```
$\tau_c = \tau_{min}$, $\hat{\mathcal{C}} = \emptyset$
$\bar{\mathcal{P}}_{2D} = \textsc{FilterKeypoints}(\hat{\mathcal{P}}_{2D}, \tau_k)$
while $\tau_c \leq \tau_{max}$ do
    $\mathcal{P}_{3D} = \textsc{Triangulate}(\bar{\mathcal{P}}_{2D}, \mathcal{C})$
    $\mathcal{P}_{2D} = \textsc{Reprojection}(\mathcal{P}_{3D})$
    while $\tau_c \leq \tau_{max}$ and $|\hat{\mathcal{C}}| < 3$ do
        $\hat{\mathcal{C}} = \textsc{SelectCam}(\mathcal{P}_{2D}, \bar{\mathcal{P}}_{2D}, \tau_c, N_c)$
        $\tau_c = \tau_c + \Delta_c$
    end while
    if $\mathcal{C} == \hat{\mathcal{C}}$ then
        return $\mathcal{P}_{3D}$, $\mathcal{P}_{2D}$
    else
        $\mathcal{C} = \hat{\mathcal{C}}$
    end if
end while
return Fail
```

D.2 Full-body Angle Prior

It is surprisingly difficult to find literature that provides a complete analysis of joint movement ranges, especially rotations in three degrees of freedom (DOF). Hence, we take references from artists' guidelines on human anatomy[1] and 3D

[1] https://design.tutsplus.com/articles/human-anatomy-fundamentals-flexibility-and-joint-limitations--vector-25401

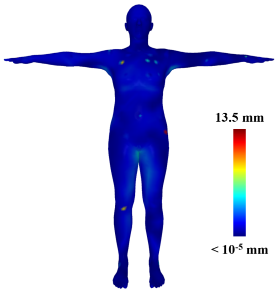

Fig. 3: The registration accuracy on high-resolution mesh (minimally clothed). The metric is mean uni-directional Chamfer distance (from SMPL vertices to high-resolution mesh vertices). Our registration (and subsequently the body shape obtained) is mostly accurate

modelers' suggested practices[2], to simplify the constraint such that the three DOF movement range is bounded by the maximum ranges in each of the DOF. Despite that this formulation is not perfect, it provides constraints that are otherwise completely absent. To easily apply the per-axis ranges, we convert the axis-angle representation into Euler angles and define the Z-axis to be aligned with the child bone of the joint in the kinematic tree (for example, *forearm* is the child bone of the joint *elbow*). To circumvent *gimbal lock* as much as possible, we define the joint frame coordinate such that the second rotation axis (Y-axis) always falls on the less flexible axis (for which the rotation is unlikely to reach 90°). Hence, we define the X-axis as the axis around which the largest rotation is achieved. Y-axis is finally defined with X- and Z-axis fixed. All values undergo manual inspection and are adjusted empirically. Note that the Euler angle rotation is used to generate a loss only; the joint rotation is still in axis-angle representation.

D.3 Annotation Quality of SMPL Parameters.

To evaluate the body shape, we compute the per-vertex error on the high-resolution scan that is the uni-directional Chamfer distance from registered SMPL mesh vertices to the high-resolution scan vertices. Note that high-resolution scans have been scaled to the real height of scanned persons. The mean per-vertex error is 0.16 mm. We also visualize the registration quality in Fig. 3. To evaluate the body pose, we compute the per-joint error as the L2 Euclidean distance between 3D keypoints and 3D joints of registered SMPL on the dynamic sequences. The mean per-joint error is 38.18 mm. Note that the error is largely attributed to the difference in the joint definition of the keypoint detector and the parametric model. As a reference, registration with an accurate optical marker system [14,23] yields a per-joint error of 29.34 mm.

[2] https://wiki.secondlife.com/wiki/Suggested_BVH_Joint_Rotation_Limits

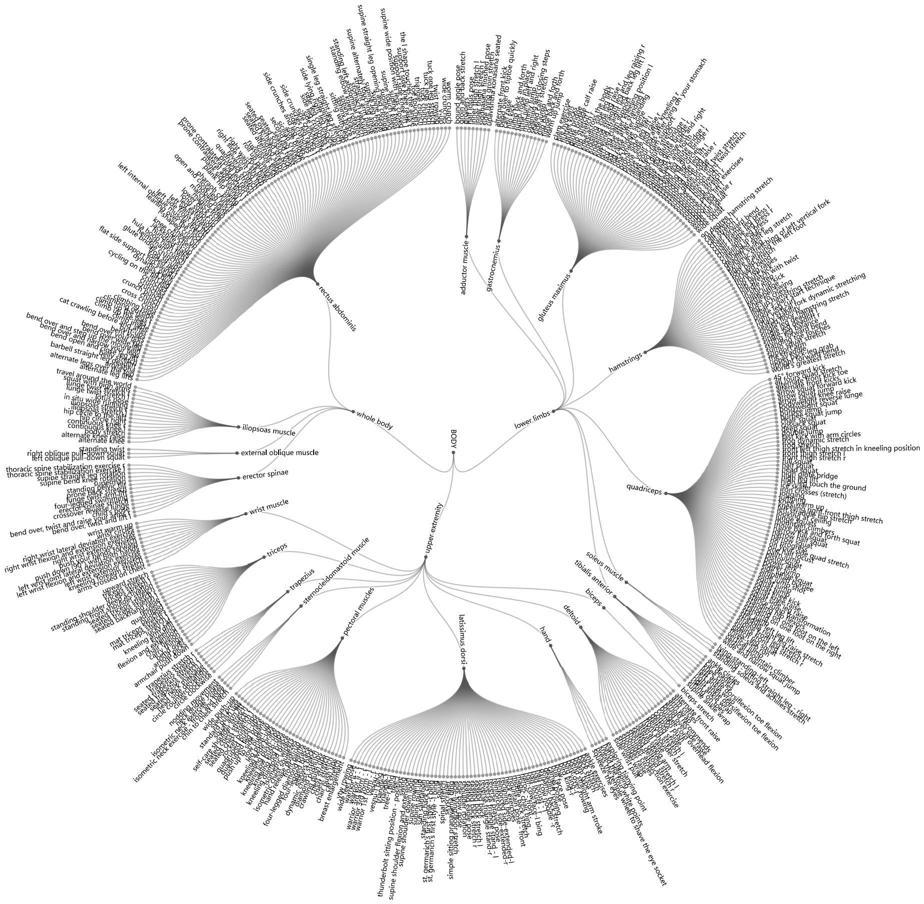

Fig. 4: The complete set of 500 actions

E Additional Details of Action Set

Design Process. In HuMMan, we design a hierarchical structure for a systematic coverage of different body parts to collate a *complete* and *unambiguous* action set. Specifically, we have *body* at the center as the first order. The second order consists of *whole body*, *upper extremity* and *lower limbs* that categorize actions by major body parts. After that, we propose a *muscle-driven* strategy to further split each major body part into main muscle groups according to human anatomy as the third order. Finally, we involve domain experts to design a series of action variants associated with each muscle in the fourth order. The full action hierarchy is demonstrated in Fig. 4.

Motion Diversity. As HuMMan contains a large amount of data, we further conduct a preliminary study on the motion diversity for further research on the

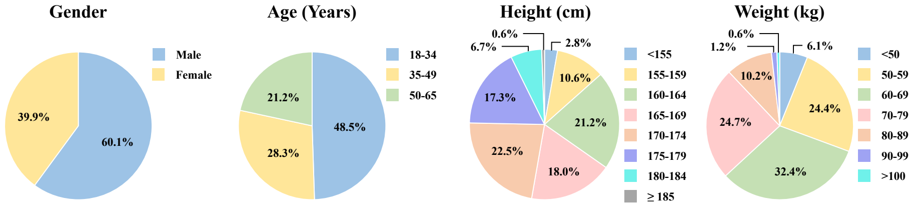

Fig. 5: Statistics of HuMMan subjects

motion prior learning. Specifically, We compute the mean standard deviation of joint angles of three datasets: 3DPW (0.159), AMASS (0.208), and HuMMan (**0.269**). The higher mean standard deviation indicate higher diversity in motions. Although the AMASS dataset with a large-scale MoCap data is wildly used in many recent works to pretrain models, HuMMan has more diversity in joint angles, showing its potential for human motion-related tasks.

F Additional Details of Subjects

Statistics. HuMMan consists of 1000 subjects. To evaluate the diversity, we include key statistics (gender, age, height and weight) of the subjects in Fig. 5.

Ethics. HuMMan involves a large number of human subjects so that we pay special attention to address ethic concerns. The recruitment process is conducted on an entirely voluntary basis. Actors and actresses who participate in HuMMan are well-informed, with legal agreements signed to acknowledge that the data will be made public for research purposes.

G Additional Details of Experiments

G.1 Splits and Protocols

HuMMan contains a massive scale of subjects (1000), actions (500), sequences (400k) and frames (60M). To constrain training and testing within a reasonable computation budget, we sample only 10% of the data. We then develop three protocols to split iPhone and Kinect data into training and test sets. **Protocol 1 (P1)**: split by subjects, the training and test set are mutually exclusive and contain 70% and 30% of the subjects respectively. P1 is used for all experiments in the main paper. **Protocol 2 (P2)**: split by actions. We split actions into three categories according to major body parts involved: *upper extremity*, *lower limbs*, and *whole body*. Training is conducted on one category whereas the test is conducted on the other two. **Protocol 3 (P3)**: split by views. Model is trained on only one view (the *front* view, or the view of the iPhone and the Kinect with ID 0) and tested on all views.

Table 2: 3D keypoint detection under Protocol 2 on Kinect splits. FCN is used as the base model.

Training	Testing	MPJPE ↓	PA-MPJPE ↓
Lower Limbs	Upper Extremity	70.3	55.7
Lower Limbs	Whole Body	97.5	72.3
Upper Extremity	Lower Limbs	75.8	55.1
Upper Extremity	Whole Body	99.6	72.5
Whole Body	Lower Limbs	77.4	56.2
Whole Body	Upper Extremity	86.2	66.4
Mean Error		84.4	63.0

Table 3: 3D keypoint detection under Protocol 3 on Kinect splits. FCN is used as the base model. The model is trained on View 0 and tested on all views.

View	0	1	2	3	4	5	6	7	8	9	Mean
MPJPE ↓	66.4	97.2	167.1	172.0	247.2	268.4	245.1	175.3	165.4	95.9	170.0
PA-MPJPE ↓	41.2	67.5	100.9	103.5	112.3	118.7	111.8	103.9	100.2	67.1	92.7

G.2 2D Keypoint Detection

Table 1: 2D Keypoint Detection under Protocol 1. Input image is resized to 384×288

Method	AP^{50} ↑	AP^{75} ↑
CPN [6]	0.86	0.93
HRNet [35]	0.91	0.97
Lite-HRNet [39]	0.87	0.93

We study 2D keypoint detection baselines on HuMMan primarily for 2D-to-3D keypoint lifting. CPN [6] is a cascaded pyramid network to improve hard keypoints detection. HRNet [35] is a novel high-resolution network that obtains high performance on COCO dataset [21], and LiteHRNet is an efficient version of HRNet. The comparison results are listed in Table 1. Because 2D keypoints are often used as an intermediate representation of 3D keypoints in a two-stage manner [27, 30], the good performance in this task can be helpful to the estimation of subsequent 3D.

G.3 3D Keypoint Detection

3D keypoint detection benchmarks under P1 setting are presented in the main paper and additional benchmarks under P2 and P3 are provided here. In Table 2, we show results on the cross-action (P2) performance of the FCN method [27]. Compared with Protocol 1, we observe that training with fewer actions and testing on unseen actions degrade the precision significantly, especially for cross-evaluation on the *whole body* category which seems to have a large action distribution misalignment with the other two categories. Furthermore, we report

Table 4: 3D parametric human recovery under Protocol 2 on Kinect splits. HMR is used as the base model.

Training	Testing	MPJPE ↓	PA-MPJPE ↓
Lower Limbs	Upper Extremity	77.2	57.0
Lower Limbs	Whole Body	109.8	77.9
Upper Extremity	Lower Limbs	80.6	56.5
Upper Extremity	Whole Body	114.2	73.3
Whole Body	Lower Limbs	85.4	61.9
Whole Body	Upper Extremity	98.3	72.6
Mean Error		94.2	66.5

Table 5: 3D parametric human recovery under Protocol 3 on Kinect splits. HMR is used as the base model. The model is trained on View 0 and tested on all views.

View	0	1	2	3	4	5	6	7	8	9	Mean
MPJPE ↓	61.9	122.9	223.9	206.2	343.9	421.0	334.0	208.0	199.0	123.5	224.4
PA-MPJPE ↓	40.2	71.9	123.7	115.0	124.4	133.1	127.2	123.1	118.0	73.3	105.0

results of cross-view (P3) in Table 3. When the model is only trained on one view (*i.e.*, View 0), we observe a considerable domain gap across different views as the errors increase as the deviation from the test view from the training view increases. The experiment results indicate that cross-view 3D keypoint detection is challenging.

G.4 3D Parametric Human Recovery

In addition to P1 benchmarks for 3D parametric human recovery presented in the main paper, we also provide more benchmarks under P2 and P3. In Table 4, we evaluate the cross-action (P2) performance of the HMR baseline. We find that testing on unseen poses is challenging (compared to P1 benchmark results). Moreover, *whole body* actions seem to have a distribution that is further away from *lower limbs* and *upper extremity* actions. In Table 5, we study the cross-view setting (P3), which is even worse than the cross-action setting. The HMR baseline is trained on View 0, and gives a clear trend that the greater the viewing angle difference, the larger the errors. View 5 is directly opposite View 0 and yields the largest error.

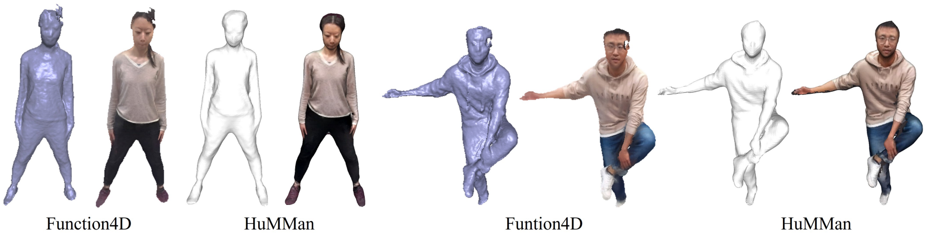

Fig. 6: We compare Function4D with HuMMan in textured mesh reconstruction

G.5 Textured Mesh Reconstruction

To fully demonstrate the capacity of HuMMan, we also provide the results of Function4D [40] as a baseline for textured mesh reconstruction since it combines both volumetric fusion and implicit surface reconstruction for volumetric capture in real-time. The results of Function4D, using 4 (ID: 0,3,6,9) views, are shown in Fig. 6.

H A More Complete Dataset Comparison

In Table 6, we provide a more thorough comparison of HuMMan with similar datasets for 1) action recognition, 2) 2D and 3D keypoint detection, 3) 3D parametric human recovery, and 4) mesh reconstruction.

Table 6: A more complete comparison of HuMMan with published datasets. Subj: subjects; Act: actions; Seq: sequences; 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
Action Recognition														
HMDB51 [18]	-	51	7k	-	✓	-	✓	-	✓	-	-	-	-	-
UCF101 [34]	-	101	13k	-	✓	-	✓	-	✓	-	-	-	-	-
Sports1M [17]	-	487	1M	-	✓	-	✓	-	✓	-	-	-	-	-
AVA [9]	-	80	437	-	✓	-	✓	-	✓	-	-	-	-	-
Kinectics 700 [5]	-	700	650k	-	✓	-	✓	-	✓	-	-	-	-	-
HACS [44]	-	200	1.55M	-	✓	-	✓	-	✓	-	-	-	-	-
Moments-In-Time [29]	-	339	1M	-	✓	-	✓	-	✓	-	-	-	-	-
FineGym [33]	-	530	32k	-	✓	-	✓	-	✓	-	-	-	-	-
HAA500 [7]	-	500	10k	591k	✓	-	✓	-	✓	-	-	-	-	-
MSR-Action3D [20]	10	20	567	-	✓	-	-	✓	✓	-	✓	-	-	-
Northwestern-UCLA [38]	10	10	1.47k	>23k	✓	-	✓	✓	✓	-	✓	-	-	-
SYSU 3DHOI [13]	40	12	65	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D [32]	40	60	56k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D 120 [22]	106	120	114k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D X [36]	106	120	113k	-	✓	-	✓	✓	✓	-	✓	✓	-	-
2D/3D Keypoint Detection and 3D Parametric Human Recovery														
J-HMDB [15]	-	21	928	33.18k	✓	-	✓	-	✓	✓	-	-	-	-
Penn Action [43]	-	15	2.32k	-	✓	-	✓	-	✓	✓	-	-	-	-
MPII [3]	-	410	-	24k	-	-	✓	-	✓	✓	-	-	-	-
COCO [21]	-	-	-	104k	-	-	✓	-	-	✓	-	-	-	-
PoseTrack [2]	-	-	>1.35k	>46k	✓	-	✓	-	-	✓	-	-	-	-
Human3.6M [14]	11	17	839	3.6M	✓	-	✓	✓	✓	✓	✓	-	-	-
CMU Panoptic [16]	8	5	65	154M	✓	-	✓	✓	-	✓	✓	-	-	-
MPI-INF-3DHP [28]	8	8	16	1.3M	✓	-	✓	-	-	✓	✓	-	-	-
TotalCapture [37]	5	5	60	1.89M	✓	-	✓	-	-	✓	✓	-	-	-
3DPW [26]	7	-	60	51k	✓	✓	✓	-	-	-	-	✓	-	-
AMASS [25]	344	-	>11k	>16.88M	✓	-	-	-	-	-	✓	✓	-	-
Mirrored-Human [8]	-	56	56	>1.5M	✓	-	-	-	✓	✓	✓	✓	-	-
AIST++ [19]	30	-	1.40k	10.1M	✓	-	✓	-	-	✓	✓	✓	-	-
Mesh Reconstruction														
ZJU LightStage [31]	6	6	9	>1k	✓	-	✓	-	✓	✓	✓	✓	✓	✓
CAPE [24]	15	-	>600	>140k	✓	-	-	-	✓	-	✓	✓	✓	-
BUFF [42]	6	3	>30	>13.6k	✓	-	✓	✓	✓	-	✓	✓	✓	✓
DFAUST [4]	10	>10	>100	>40k	✓	-	✓	✓	✓	✓	✓	✓	✓	✓
People Snapshot [1]	9	-	24	15k	✓	-	✓	-	-	-	✓	✓	✓	✓
LiveCap [11]	7	11	11	36k	✓	-	✓	-	✓	✓	✓	✓	✓	✓
DynaCap [10]	4	5	5	35k	✓	-	✓	-	✓	✓	✓	✓	✓	✓
DeepCap [12]	4	17	17	26k	✓	✓	✓	-	✓	✓	✓	-	✓	✓
HUMBI [41]	772	-	-	~26M	✓	-	✓	-	-	✓	✓	✓	✓	✓
THuman [45]	200	-	-	>6k	-	-	✓	✓	-	-	-	✓	✓	✓
THuman2.0 [40]	200	-	-	>500	-	-	-	-	-	-	-	✓	✓	✓
Multi-task														
HuMMan (ours)	1000	500	400k	60M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

References

1. 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) 11
2. 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) 11
3. 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) 11
4. 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) 11
5. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019) 11
6. 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) 8
7. 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) 11
8. Fang, Q., Shuai, Q., Dong, J., Bao, H., Zhou, X.: Reconstructing 3d human pose by watching humans in the mirror. In: CVPR (2021) 11
9. 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) 11
10. 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) 11
11. 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) 11
12. 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) 11
13. 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) 11
14. 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) 5, 11
15. 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) 11

16. 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) 11
17. 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) 11
18. 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) 11
19. 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) 11
20. 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) 11
21. 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) 8, 11
22. 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) 11
23. 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) 5
24. 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) 11
25. 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) 11
26. 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) 11
27. 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) 8
28. 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) 11
29. 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) 11
30. 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) 8

31. 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) 11
32. 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) 11
33. 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) 11
34. 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) 11
35. 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) 8
36. 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) 11
37. 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) 11
38. 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) 11
39. Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J.: Lite-hrnet: A lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10440–10450 (2021) 8
40. 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) 10, 11
41. 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) 11
42. 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) 11
43. 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) 11
44. 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) 11
45. 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) 11