

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

Zhongang Cai\*, Daxuan Ren\*, Ailing Zeng\*, Zhengyu Lin\*, Tao Yu\*, Wenjia Wang\*,  
Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang,  
Chen Change Loy, Lei Yang, Ziwei Liu

Shanghai AI Laboratory, S-Lab, Nanyang Technological University,  
SenseTime Research, The Chinese University of Hong Kong,  
Tsinghua University

ECCV'22 Oral

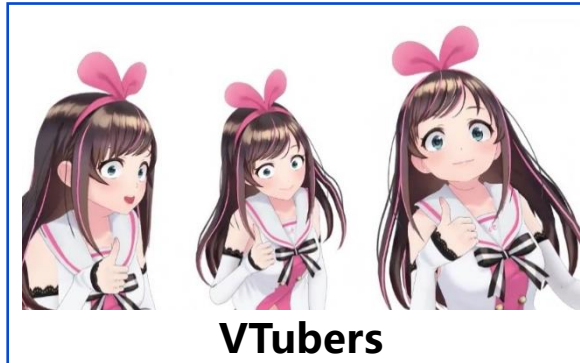
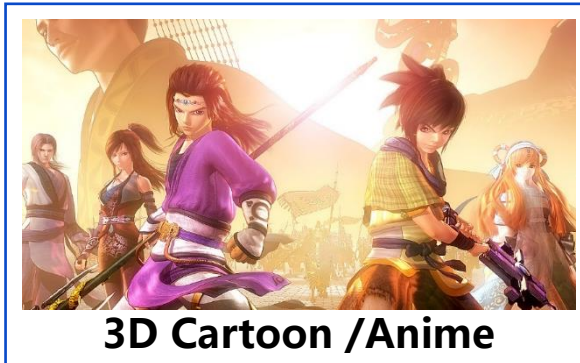




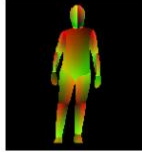


**S-LAB**  
FOR ADVANCED  
INTELLIGENCE

# Overview



# Background



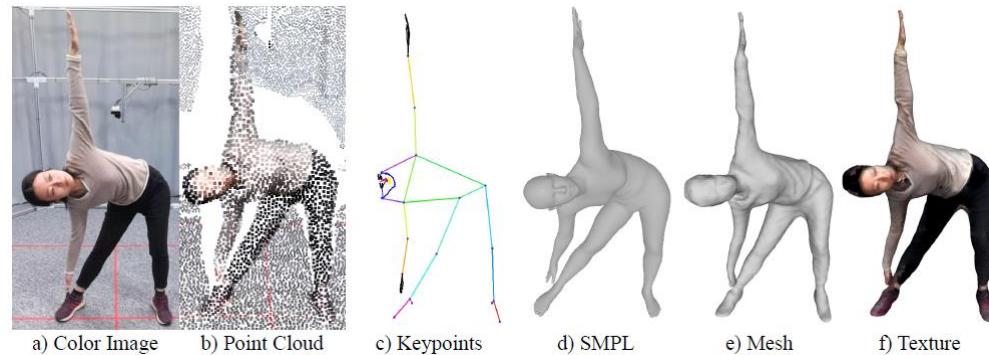
Annotation	Sparse 2D	Dense Labeling	Dense Correspondence	Constrained 3D	In-the-wild 3D
Examples					
Annotation Cost	\$	\$\$	\$\$\$	\$\$\$\$	\$\$\$\$\$

**3D Human Data Is Expensive to Acquire**

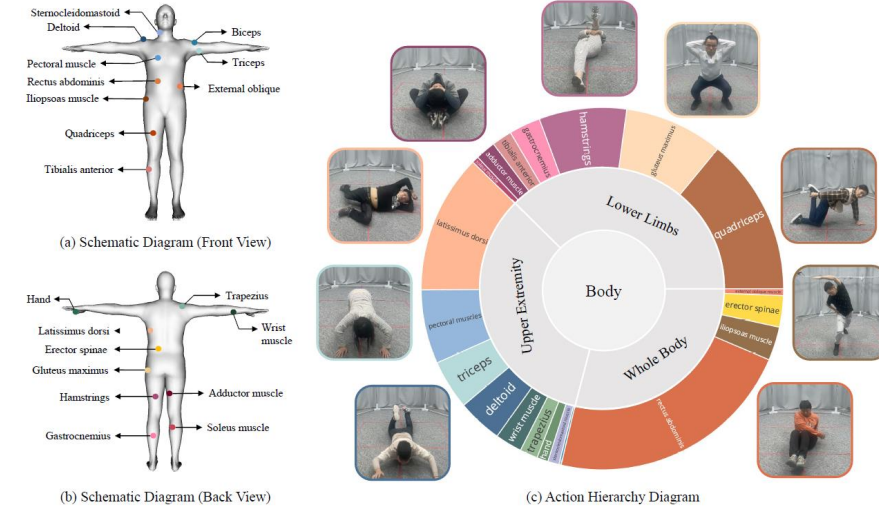


# HuMMan

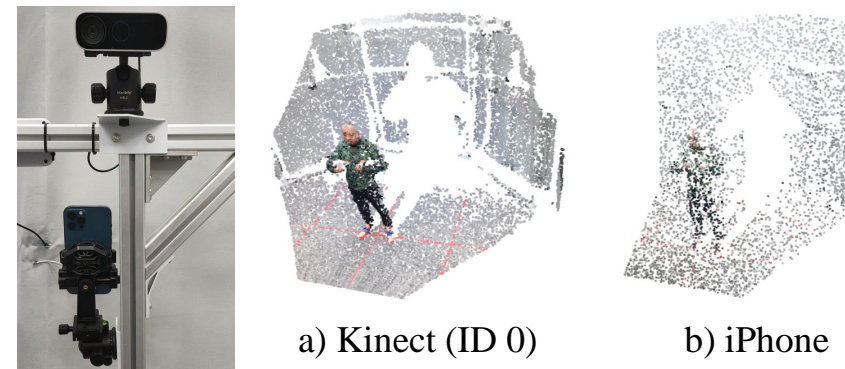
Dataset	#Subj	#Act	#Seq	#Frame	Video	Mobile	Modalities							
							RGB	D/PC	Act	K2D	K3D	Param	Mesh	Txtr
UCF101 [85]	-	101	13k	-	✓	-	✓	-	✓	-	-	-	-	-
AVA [20]	-	80	437	-	✓	-	✓	-	✓	-	-	-	-	-
FineGym [82]	-	530	32k	-	✓	-	✓	-	✓	-	-	-	-	-
HAA500 [14]	-	500	10k	591k	✓	-	✓	-	✓	-	-	-	-	-
SYSU 3DHOI [26]	40	12	480	-	✓	-	✓	-	✓	-	✓	-	-	-
NTU RGB+D [81]	40	60	56k	-	✓	-	✓	-	✓	-	✓	-	-	-
NTU RGB+D 120 [54]	106	120	114k	-	✓	-	✓	-	✓	-	✓	-	-	-
NTU RGB+D X [91]	106	120	113k	-	✓	-	✓	-	✓	-	✓	-	-	-
MPII [3]	-	410	-	24k	-	-	✓	-	✓	-	-	-	-	-
COCO [52]	-	-	-	104k	-	-	✓	-	✓	-	-	-	-	-
PoseTrack [2]	-	-	>1.35k	>46k	✓	-	✓	-	✓	-	-	-	-	-
Human3.6M [28]	11	17	839	3.6M	✓	-	✓	-	✓	-	✓	-	-	-
CMU Panoptic [34]	8	5	65	154M	✓	-	✓	-	✓	-	✓	-	-	-
MPI-INF-3DHP [63]	8	8	16	1.3M	✓	-	✓	-	✓	-	✓	-	-	-
3DPW [61]	7	-	60	51k	✓	✓	✓	-	✓	-	✓	-	-	-
AMASS [60]	344	-	>11k	>16.88M	✓	-	✓	-	✓	-	✓	-	-	-
AIST++ [48]	30	-	1.40k	10.1M	✓	-	✓	-	✓	-	✓	-	-	-
CAPE [59]	15	-	>600	>140k	✓	-	✓	-	✓	-	✓	-	-	-
BUFF [105]	6	3	>30	>13.6k	✓	-	✓	-	✓	-	✓	-	-	-
DEAUST [6]	10	>10	>100	>40k	✓	-	✓	-	✓	-	✓	-	-	-
HUMBI [101]	772	-	-	~26M	✓	-	✓	-	✓	-	✓	-	-	-
ZJU LightStage [76]	6	6	9	>1k	✓	-	✓	-	✓	-	✓	-	-	-
THuman2.0 [99]	200	-	-	>500	✓	-	✓	-	✓	-	✓	-	-	-
<b>HuMMan (ours)</b>	<b>1000</b>	<b>500</b>	<b>400k</b>	<b>60M</b>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



The Largest-scale Multi-modal Dataset for Human Sensing and Modeling



Complete and Unambiguous Action Set (500)

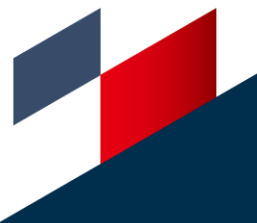


The First Large-scale Multi-modal Dataset Captured with a Mobile Device

# Scale

- ✓ Large-scale
  - ✓ Subjects
  - ✓ Actions
  - ✓ Sequences
  - ✓ Frames
- ✓ Multi-modal
- ✓ Mobile device
- ✓ Multi-task

Dataset	#Subj	#Act	#Seq	#Frame	Video	Mobile	Modalities							
							RGB	D/PC	Act	K2D	K3D	Param	Mesh	Txtr
UCF101 [85]	-	101	13k	-	✓	-	✓	-	✓	-	-	-	-	-
AVA [20]	-	80	437	-	✓	-	✓	-	✓	-	-	-	-	-
FineGym [82]	-	530	32k	-	✓	-	✓	-	✓	-	-	-	-	-
HAA500 [14]	-	500	10k	591k	✓	-	✓	-	✓	-	-	-	-	-
SYSU 3DHOI [26]	40	12	480	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D [81]	40	60	56k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D 120 [54]	106	120	114k	-	✓	-	✓	✓	✓	-	✓	-	-	-
NTU RGB+D X [91]	106	120	113k	-	✓	-	✓	✓	✓	-	✓	✓	-	-
MPII [3]	-	410	-	24k	-	-	✓	-	✓	✓	-	-	-	-
COCO [52]	-	-	-	104k	-	-	✓	-	✓	✓	-	-	-	-
PoseTrack [2]	-	-	>1.35k	>46k	✓	-	✓	-	✓	✓	-	-	-	-
Human3.6M [28]	11	17	839	3.6M	✓	-	✓	✓	✓	✓	✓	-	-	-
CMU Panoptic [34]	8	5	65	154M	✓	-	✓	✓	-	✓	✓	-	-	-
MPI-INF-3DHP [63]	8	8	16	1.3M	✓	-	✓	-	-	✓	✓	-	-	-
3DPW [61]	7	-	60	51k	✓	✓	✓	-	-	-	-	✓	-	-
AMASS [60]	344	-	>11k	>16.88M	✓	-	-	-	-	-	✓	✓	-	-
AIST++ [48]	30	-	1.40k	10.1M	✓	-	✓	-	-	✓	✓	✓	-	-
CAPE [59]	15	-	>600	>140k	✓	-	-	-	✓	-	✓	✓	✓	-
BUFF [105]	6	3	>30	>13.6k	✓	-	✓	✓	✓	-	✓	✓	✓	✓
DFAUST [6]	10	>10	>100	>40k	✓	-	✓	✓	✓	✓	✓	✓	✓	✓
HUMBI [101]	772	-	-	~26M	✓	-	✓	-	-	✓	✓	✓	✓	✓
ZJU LightStage [76]	6	6	9	>1k	✓	-	✓	-	✓	✓	✓	✓	✓	✓
THuman2.0 [99]	200	-	-	>500	-	-	-	-	-	-	✓	✓	✓	✓
<b>HuMMan (ours)</b>	<b>1000</b>	<b>500</b>	<b>400k</b>	<b>60M</b>	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



# Modalities

- ✓ Large-scale
- ✓ Multi-modal
  - ✓ RGB
  - ✓ Depth/Point Cloud
  - ✓ Action Label
  - ✓ 2D Keypoints
  - ✓ 3D Keypoints
  - ✓ SMPL
  - ✓ Mesh
  - ✓ Texture
- ✓ Mobile device
- ✓ Multi-task



a) Color Image

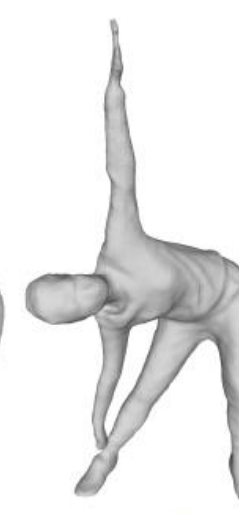
b) Point Cloud



c) Keypoints



d) SMPL



e) Mesh

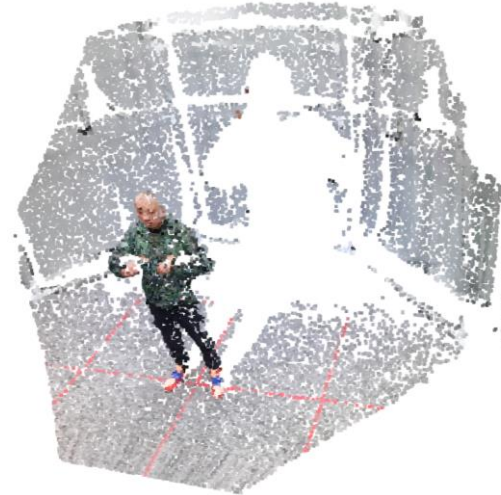


f) Texture



# Mobile Device

- ✓ Multiple Modalities
- ✓ Mobile Device
  - ✓ With Build-in LiDAR
- ✓ Action Set
- ✓ Multiple Tasks



a) Kinect (ID 0)



b) iPhone

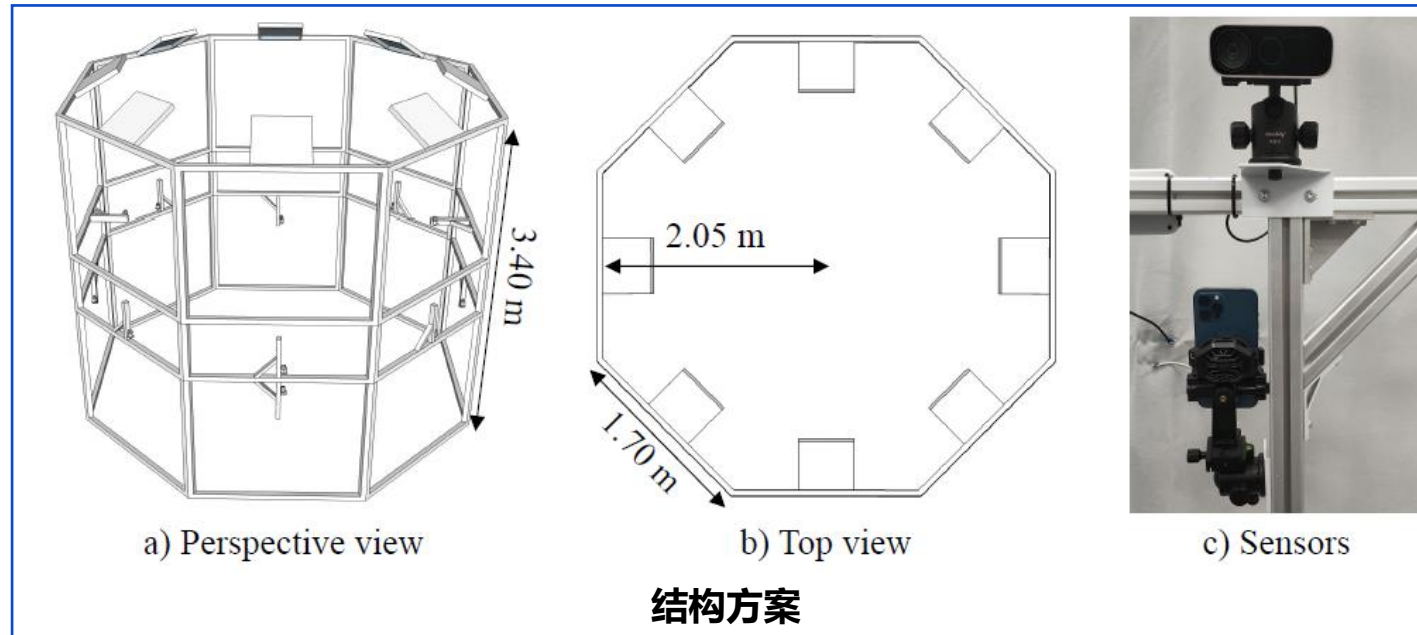


# Hardware





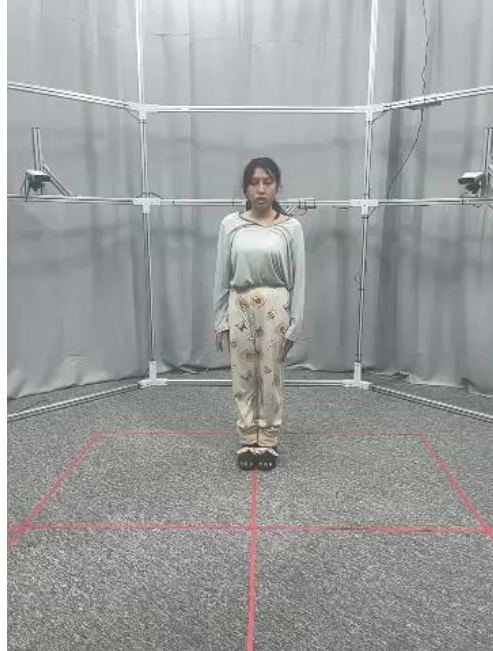
# Hardware



# Data Collection



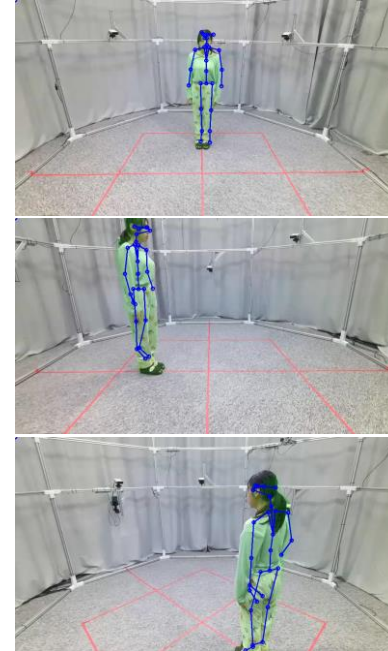
Artec Eva



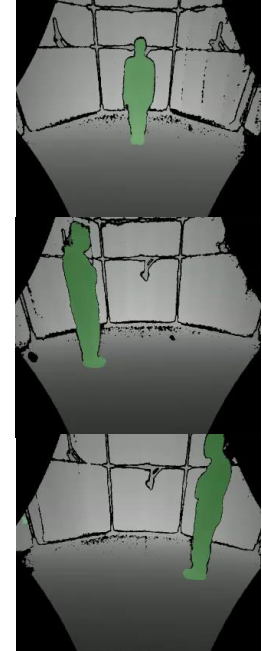
iPhone RGB



iPhone Depth



Kinect RGB



Kinect Depth



Scan Accuracy



Views



Data / second



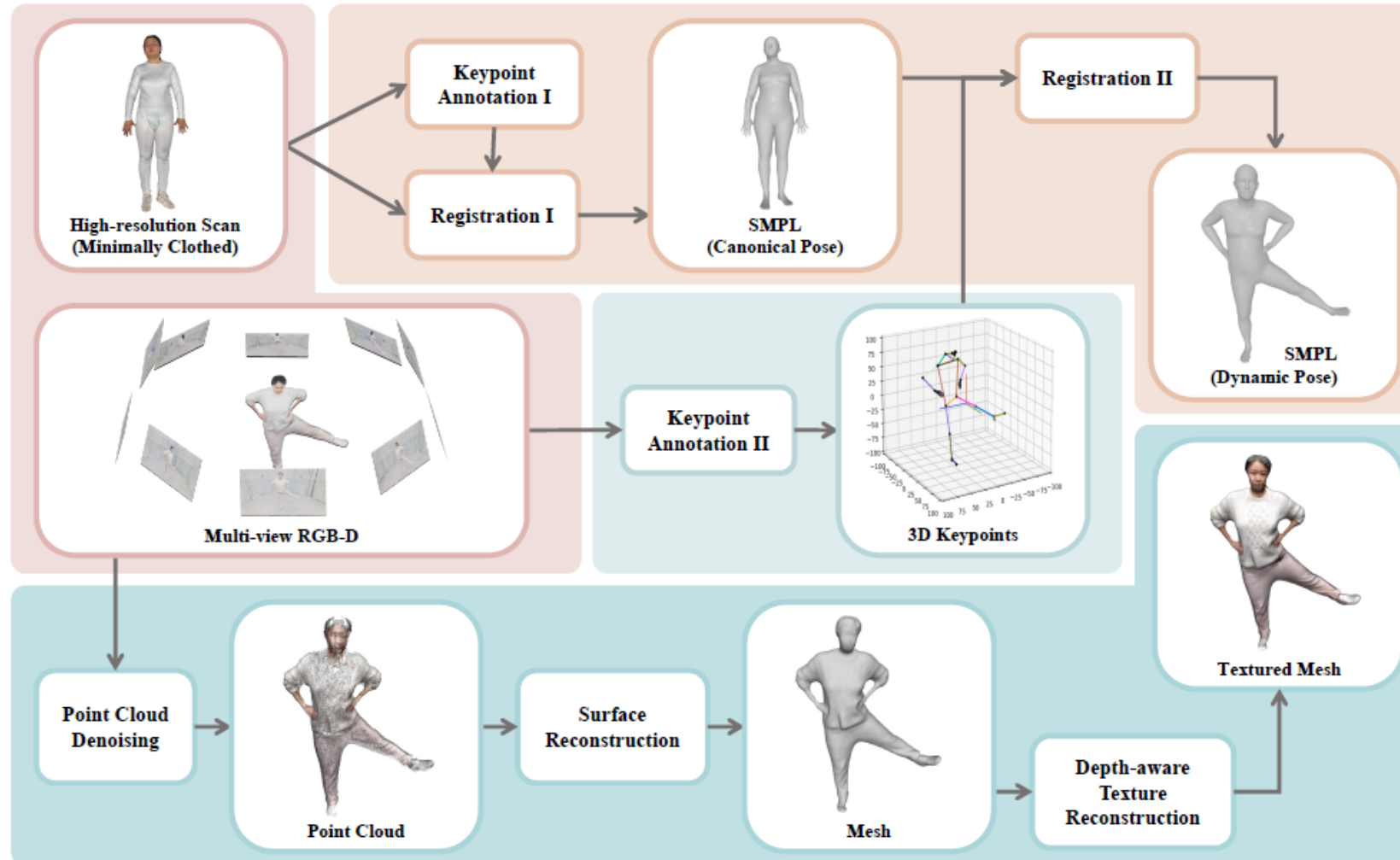
People / day



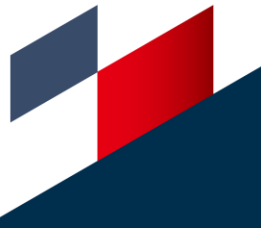
# Toolchain



# Toolchain



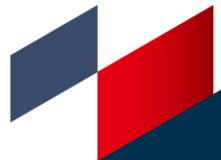
Toolchain



Video

# HuMMan

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



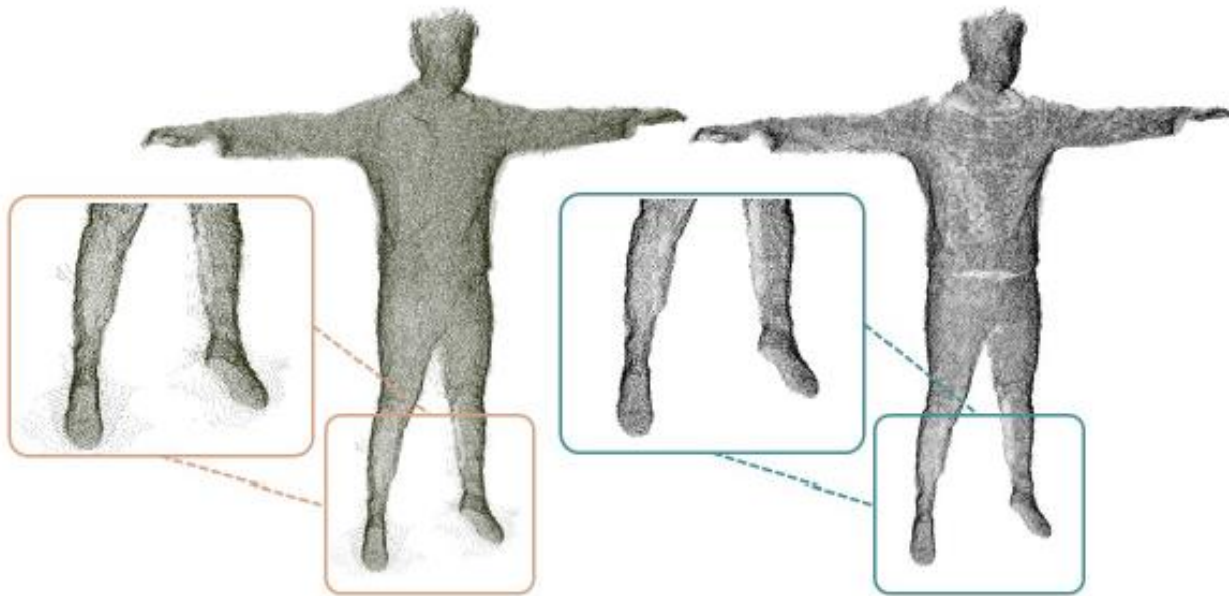
# Shape Registration



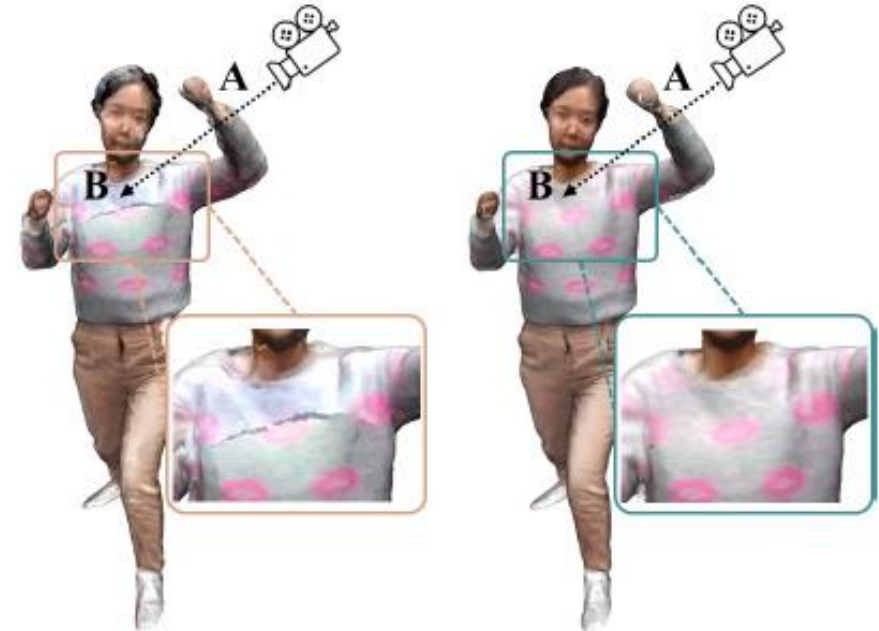
Registration on High-Resolution Scans



# Textured Mesh Reconstruction



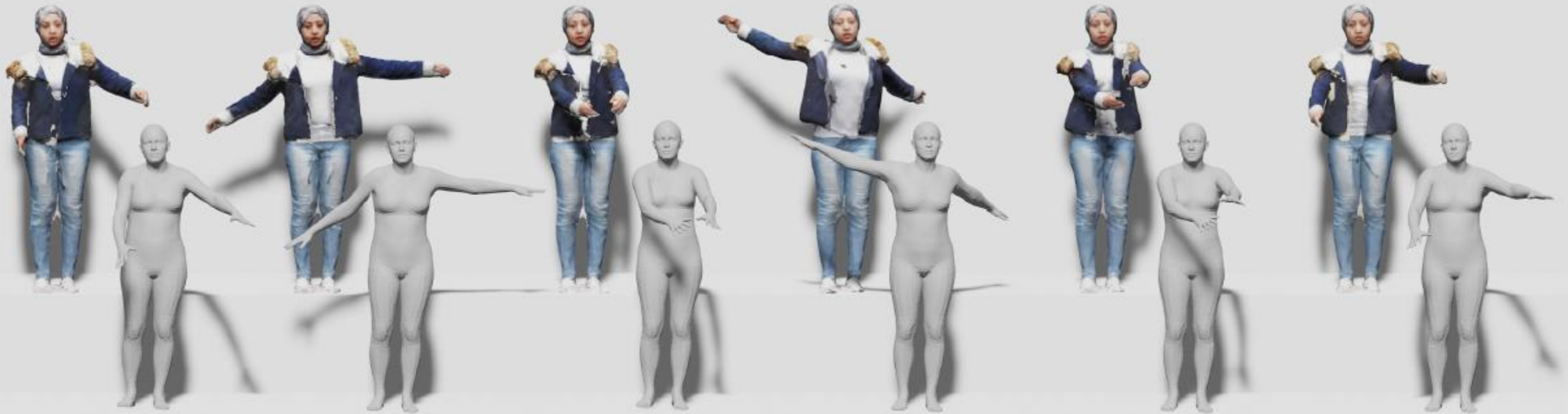
a) Point Cloud Denoising



b) Depth-aware Texture Reconstruction



# Dynamic Parametric / Mesh Sequences



Textured Mesh and SMPL Sequences



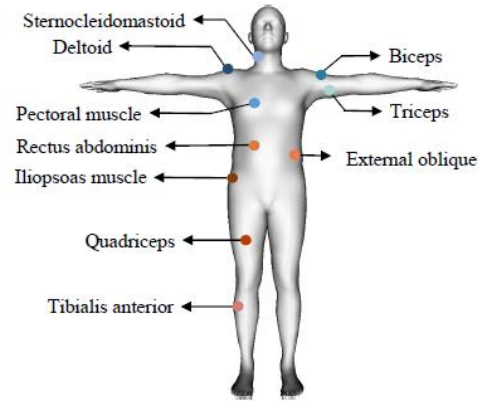


# Action Set

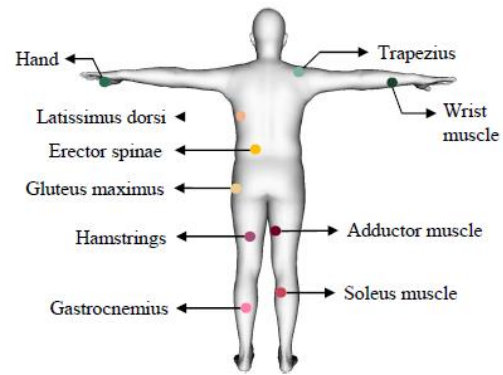


# Action Set

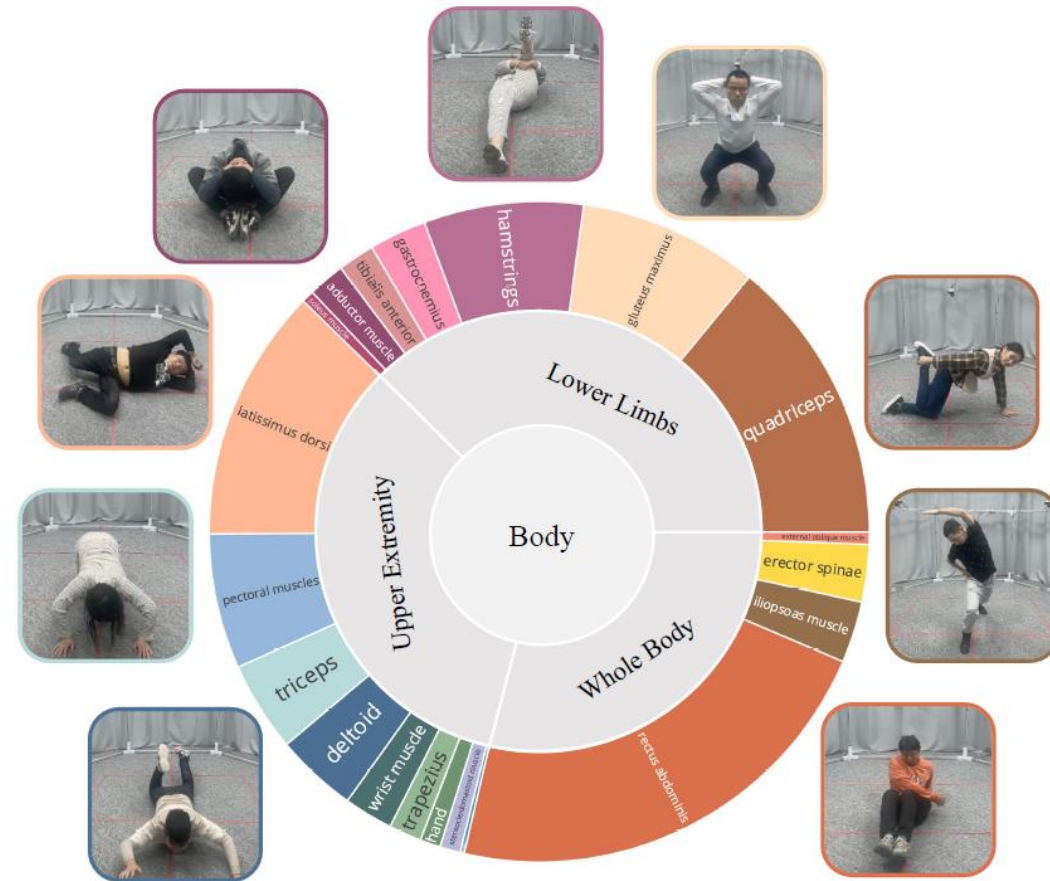
- Hierarchical
- Completeness
- Unambiguity



(a) Schematic Diagram (Front View)



(b) Schematic Diagram (Back View)



(c) Action Hierarchy Diagram



# Subjects



# Subjects

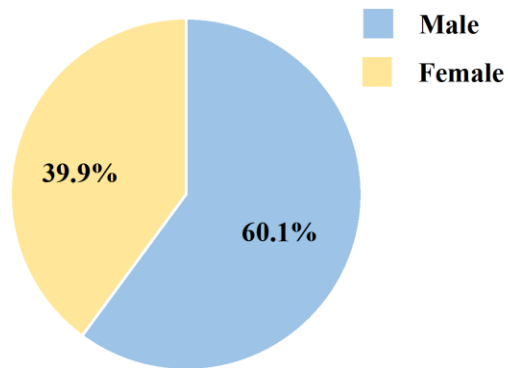


Varieties in Genders, Ages, Body Shapes (Heights, Weights), Ethnicity, and Clothing

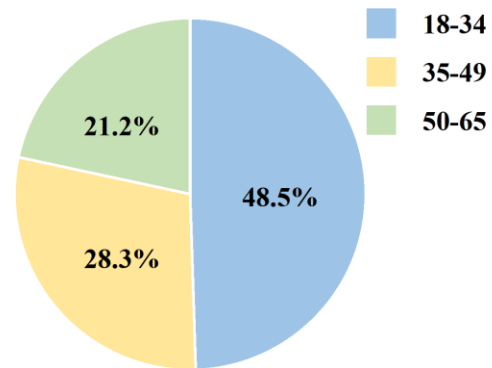


# Statistics

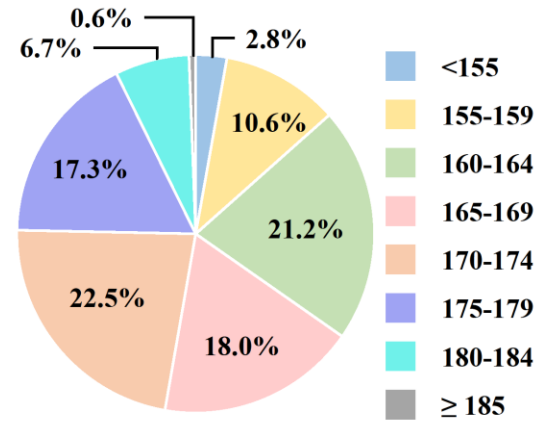
### Gender



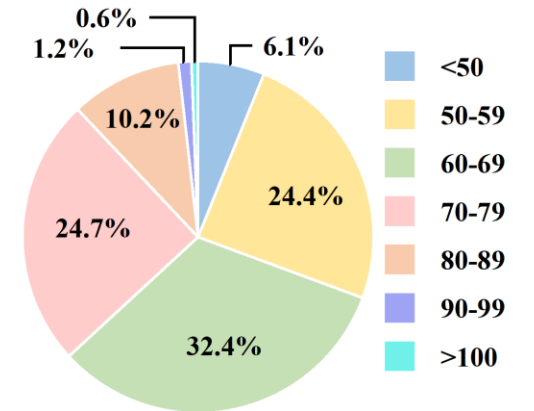
### Age (Years)



### Height (cm)



### Weight (kg)



# Experiments



# Action Recognition

- Challenging action set
  - 2s-AGCN obtains Top-1 accuracy of 88.9% and 82.9% on NTU RGB-D 60/120
- Fine-grained actions
  - Large Top-1 vs Top-5 gap

Table 2: **Action Recognition**

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



# 3D Keypoints

- 3D keypoint estimation is challenging in HuMMan
- Model trained on HuMMan exhibits better transferability

Table 3: **3D Keypoint Detection**. PA: PA-MPJPE

Train	Test	MPJPE ↓	PA ↓
FCN [62]			
HuMMan	HuMMan	78.5	46.3
H36M	AIST++	133.9	73.1
HuMMan	AIST++	116.4	67.2
Video3D [75]			
HuMMan	HuMMan	73.1	43.5
H36M	AIST++	128.5	72.0
HuMMan	AIST++	109.2	63.5





# 3D Parametric Human Recovery

- Point cloud-based parametric human recovery is challenging

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

Method	MPJPE ↓	PA-MPJPE ↓
HMR	54.78	36.14
VoteHMR	144.99	106.32



# Mobile Device

- Cross-device domain gap exists
- More severe in point cloud applications

Table 5: **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. PA: PA-MPJPE

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



Thank you!

Homepage:

<https://caizhongang.github.io/projects/HuMMan/>

