Data Infrastructure for Scaling up Human Pose and Shape Estimation to the Real World

Zhongang Cai 蔡中昂

Ph.D. Student
S-Lab, Nanyang Technological University

12 Jan 2023

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

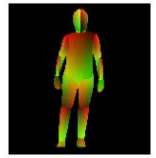
Background

Movies

Games

3D Cartoon / Anime

VTubers

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

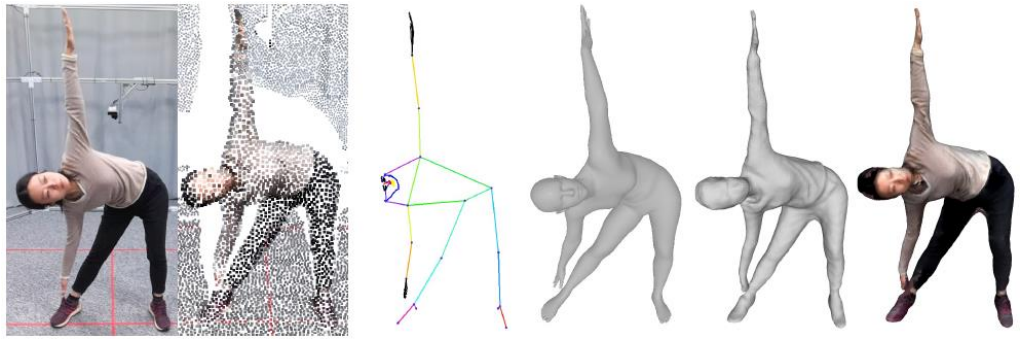
3D Human Data Is Expensive to Acquire [1]

[1] Y. Rong et al., Delving deep into hybrid annotations for 3d human recovery in the wild, ICCV 2019

Solution

Playing for 3D Human Recovery

Large-scale Generation of Synthetic Data



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

Large-scale Collection of Real Data

Playing for 3D Human Recovery

Zhongang Cai*, Mingyuan Zhang*, Jiawei Ren*, Chen Wei, Daxuan Ren,
Zhengyu Lin, Zhao Haiyu, Lei Yang, Chen Change Loy, Ziwei Liu

S-Lab, Nanyang Technological University,
SenseTime Research

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Playing for 3D Human Recovery

Overview

TABLE 1: 3D human dataset comparisons. We compare GTA-Human with existing real datasets with SMPL annotations and synthetic datasets with highly realistic setups. GTA-Human has competitive scale and diversity. Datasets are divided into three types: real, synthetic and mixed. GTA-Human samples character action sequences from a large in-game database that allows a unique action to be assigned to each video sequence. Note that EFT [20] re-annotates 2D human pose estimation datasets where the number of subjects are difficult to trace. *: 3DPW and Panoptic Studio only have general descriptions of scene activities

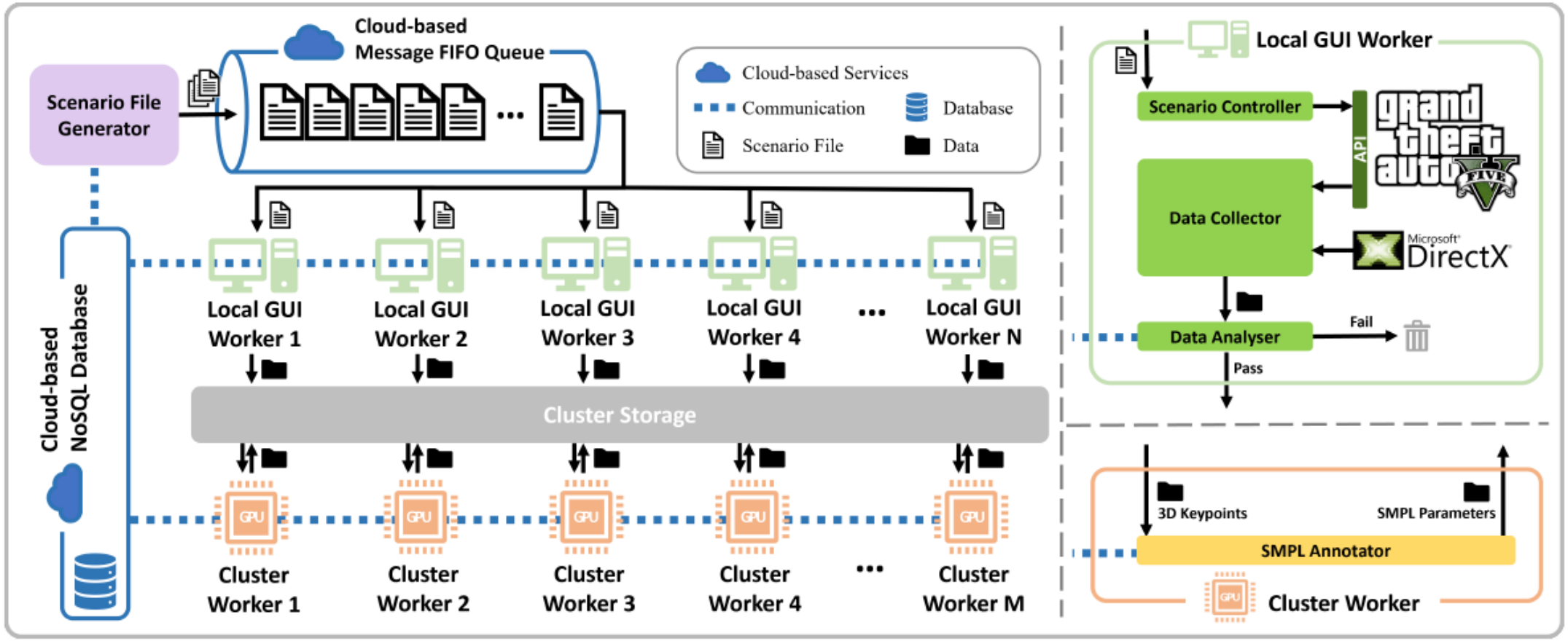
Dataset	Year	Type	In-the-Wild	Video	#SMPL	#Sequence	#Subject	#Action
HumanEva [5]	2009	Real	-	✓	NA	7	4	6
Human3.6M [8]	2013	Real	-	✓	312K	839	11	15
MPI-INF-3DHP [21]	2017	Mixed	✓	✓	96K	16	8	8
3DPW [6]	2018	Real	✓	✓	32K	60	18	*
Panoptic Studio [9]	2019	Real	-	✓	736K	480	∼100	*
EFT [20]	2020	Real	✓	-	129K	NA	Many	NA
SMPLy [7]	2020	Real	✓	✓	24K	567	742	NA
AGORA [22]	2021	Synthetic	✓	-	173K	NA	>350	NA
GTA-Human	2022	Synthetic	✓	✓	1.4M	20K	>600	20K




NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Toolchain & Data

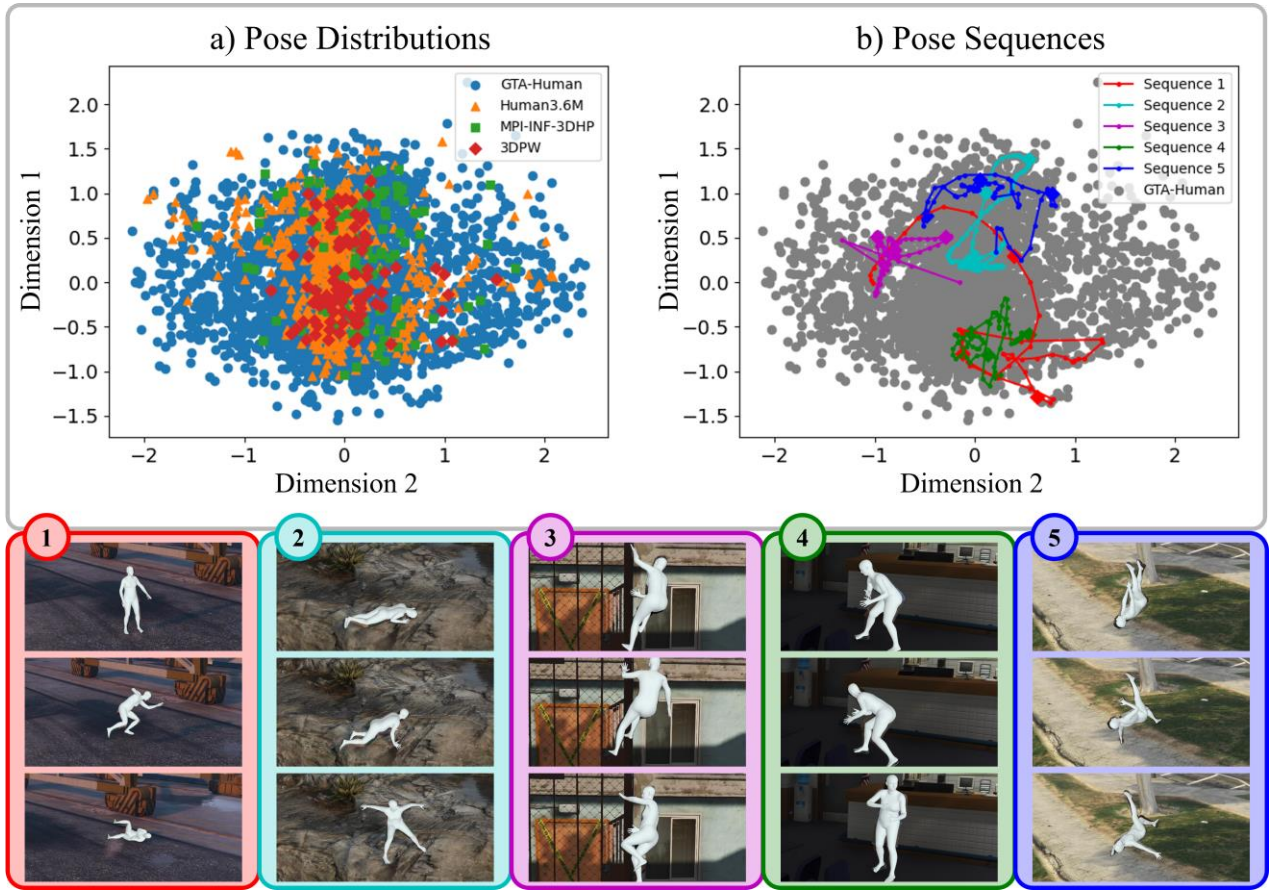
Toolchain


Cloud-based
Message FIFO Queue
Scenario File
Generator
Cloud-based Services
Communication
Database
Scenario File
Data
Cloud-based
NoSQL Database
Local GUI
Worker 1
Local GUI
Worker 2
Local GUI
Worker 3
Local GUI
Worker 4
Local GUI
Worker N
Cluster Storage
GPU
Cluster
Worker 1
Cluster
Worker 2
Cluster
Worker 3
Cluster
Worker 4
Cluster
Worker M
Local GUI Worker
Scenario Controller
API
Data Collector
Microsoft
DirectX
Data Analyser
Fail
Pass
3D Keypoints
SMPL Parameters
SMPL Annotator
Cluster Worker

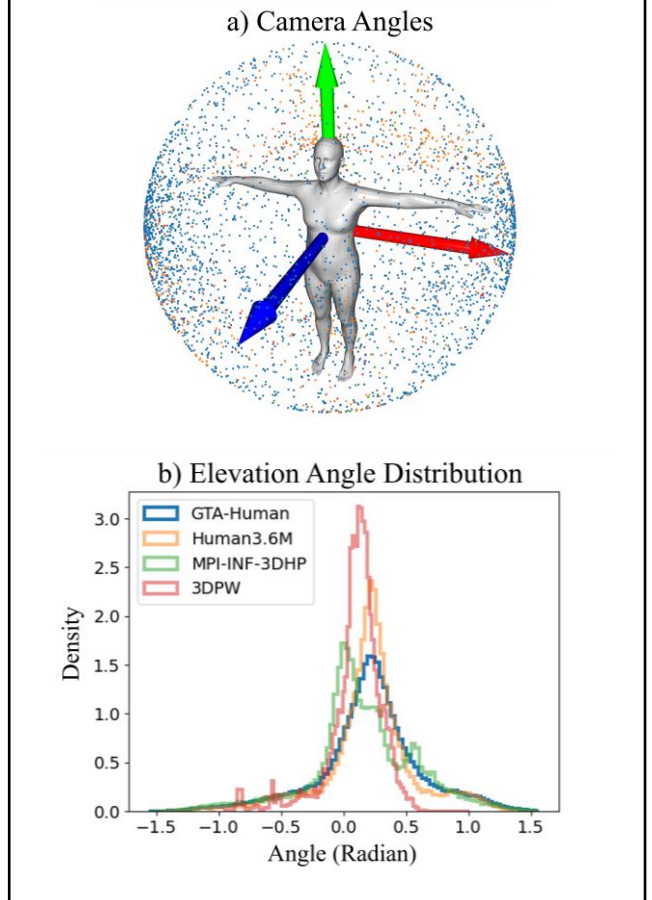
Data Diversity

Subjects, Locations, Weathers, and Light Conditions

Data Diversity



Actions



Camera Angles

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Experiments

Experiments

TABLE 2: GTA-Human's impact on model performance. The values are reported on 3DPW test set in mm. We employ two strategies: **blended training (BT)** that directly mixes GTA-Human data with real data to train an HMR model; **finetuing (FT)** that finetunes pretrained models with mixed data. Significant performance improvements are achieved with both settings. Including GTA-Human in the training boosts the HMR [23] baseline to outperform much more sophisticated methods such as SPIN [24] that leverages in-the-loop optimization (Registration) and VIBE [25] that utilizes temporal information (Video); State-of-the-art method PARE [26] also benefit from data mixture. We also conduct further experiments on video-based human recovery with VIBE in Table 3. Mixture: data mixture strategies. Real: real datasets.

Method	Mixture	Registration	Video	Pretrain	Train	Finetune	MPJPE ↓	PA-MPJPE ↓
HMR	-	-	-	ImageNet	Real	-	112.3	67.5
HMR+	-	-	-	ImageNet	Real	-	98.5	61.7
SPIN	-	✓	-	ImageNet	Real	-	96.9	59.2
VIBE	-	-	✓	ImageNet	Real	-	93.5	56.5
PARE	-	-	-	ImageNet	Real	-	82.0	50.9
HMR	BT	-	-	ImageNet	Mixed	-	98.7 (-13.6)	60.5 (-7.0)
HMR	FT	-	-	HMR	-	Mixed	91.4 (-20.9)	55.7 (-11.8)
HMR+	BT	-	-	ImageNet	Mixed	-	88.7 (-9.8)	56.0 (-5.7)
HMR+	FT	-	-	HMR+	-	Mixed	91.3 (-7.2)	55.5 (-6.2)
SPIN	FT	-	-	SPIN	-	Mixed	83.2 (-13.7)	52.0 (-7.2)
PARE	FT	-	-	PARE	-	Mixed	77.5 (-4.5)	46.8 (-4.1)

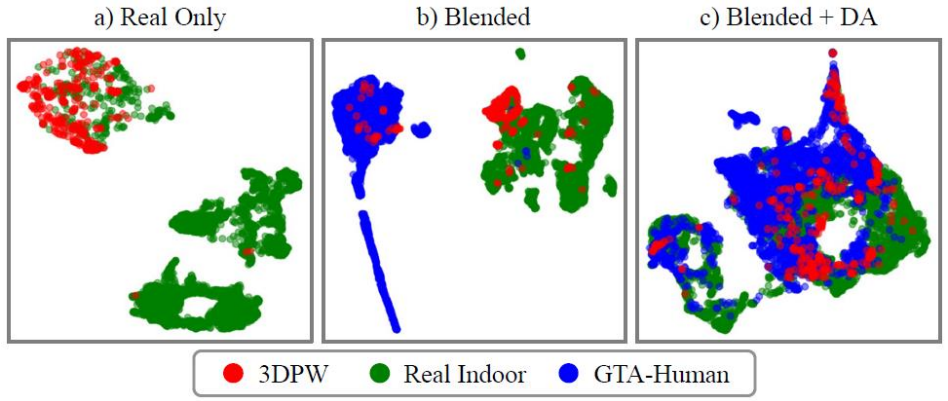
TABLE 3: Video-based 3D human recovery. The values are reported on 3DPW [6] test set with VIBE as the base model. MI3: MPI-INF-3DHP. GTA: GTA-Human. PA: PA-MPJPE. Accel: acceleration error (mm/s^2). *: downsampled GTA-Human data to match the size of MPI-INF-3DHP (96K SMPL poses).

MI3	3DPW	GTA-Human	MPJPE ↓	PA ↓	Accel ↓
✓	-	-	95.0	56.5	27.1
-	-	✓*	93.7	55.0	26.3
-	✓	-	87.9	54.7	**23.2**
-	-	✓	91.3	54.1	24.7
-	✓	✓	**85.2**	52.4	24.2
✓	✓	✓	86.0	**51.9**	23.3

Adding synthetic data is effective

- **Image-based: 4~10 mm PA-MPJPE improvements**
- **Video-based: almost on-par with real data**

Experiments



Domain Gap Analysis

TABLE 6: Domain adaptation with equal amount real and synthetic data. PA: PA-MPJPE.

Method	Real	GTA-Human	PA-MPJPE ↓
HMR	✓	-	76.7
HMR (1×)	-	✓	65.7
HMR (BT, 1×)	✓	✓	58.6
CycleGAN [68]	✓	✓	61.6
Chen *et al*. [70]	✓	✓	57.9
JAN [69]	✓	✓	56.5
Ganin *et al*. [71]	✓	✓	**55.5**

Domain Adaption

Experiments

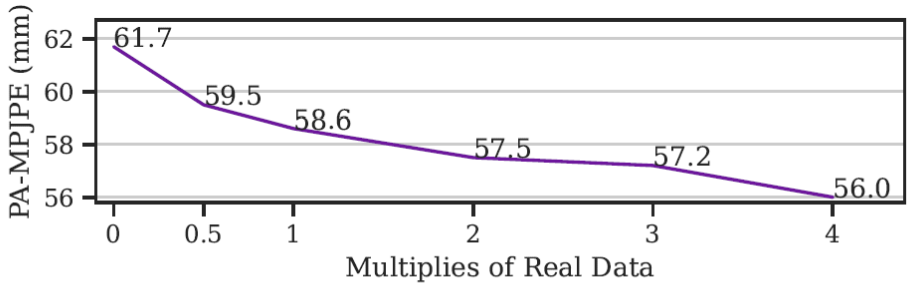


Fig. 8: Amount of GTA-Human Data. The horizontal axis indicate the amount of GTA-Human data used as multiples of the amount of real data. HMR+ is used as the base method.

Data Scale Matters!

TABLE 7: Synthetic Data as a Supplement. Different total data amount with different real data ratio are shown. Values are PA-MPJPE (mm) on 3DPW test set. Synthetic data are sampled from 4× set during training. N/A: this ratio cannot be sustained beyond 300K data due to insufficient real data. HMR+ (BT) is used as the base method.

Real Ratio	100K	200K	300K	400K	500K
0%	70.6	64.5	65.7	65.0	64.9
25%	62.4	60.9	58.0	57.6	57.3
50%	61.7	58.9	57.9	56.3	55.6
75%	62.4	58.4	56.8	55.7	N/A
100%	65.8	62.7	61.7	N/A	N/A

Supplementing Synthetic Data to Real Data

Experiments

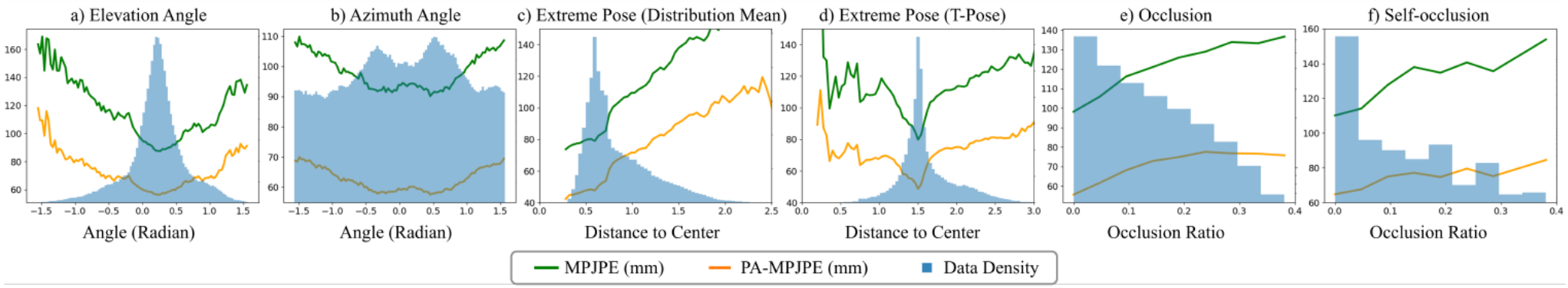


Fig. 9: Impact of data scarcity. We show that model performance is sensitive to data scarcity, and this observation is consistent on factors such as camera angles, poses, and occlusion. For c) and d), we follow [74] to encode pose as a set of 3D coordinates of the 24 key joints, and plot the distance from the mean pose and T-pose respectively. The data density of e) and f) are in log scale.

Impact of Data Scarcity: Severe Performance Degradation where Data is Scarce

Experiments

TABLE 8: Strong supervision is key. The first row is the HMR+ baseline without any GTA-Human data added.

Keypoints	SMPL	MPJPE ↓	PA-MPJPE ↓
-	-	98.5	61.7
✓	-	93.4	60.9
-	✓	92.0	56.3
✓	✓	88.7	56.0

Strong Supervision is Key

TABLE 9: Big data benefits big models. Real: training with only the real datasets. +GTA: blended training setting is used with GTA-Human. Values in green indicate the error reduction in PA-MPJPE (mm) with blended training.

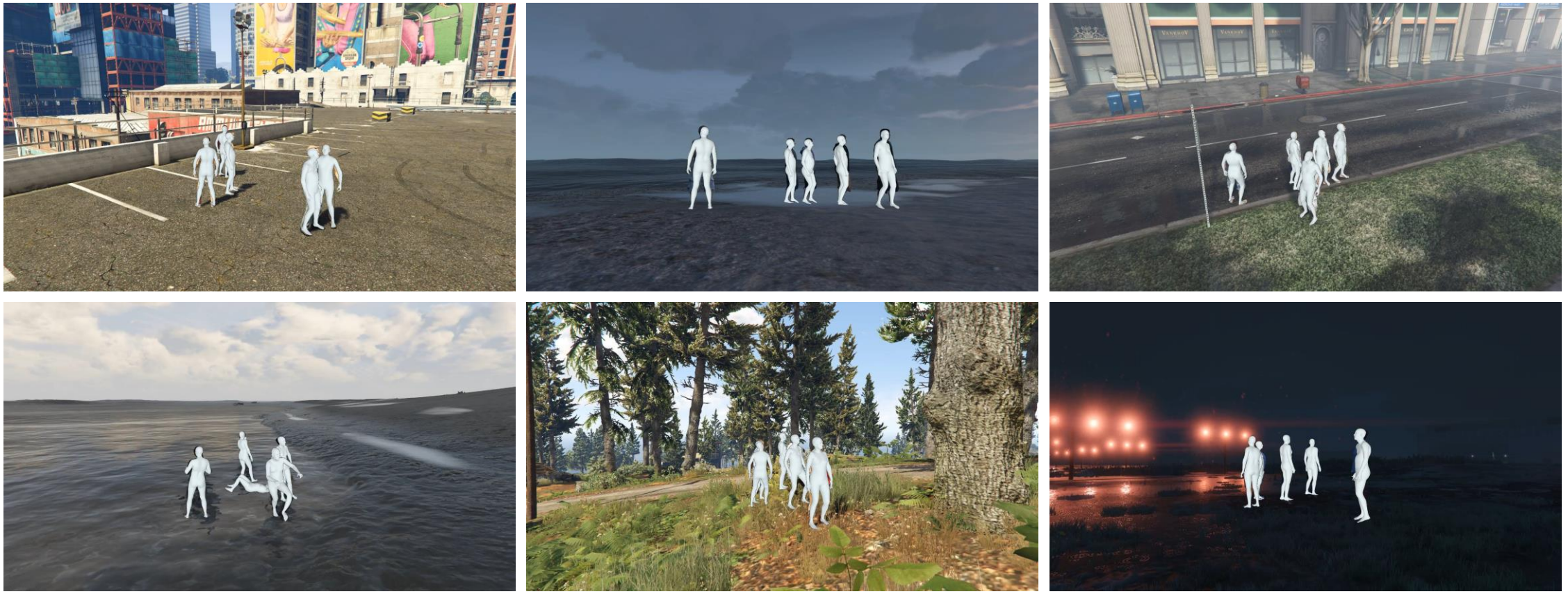
Backbone	#Param	Real ↓	+GTA-Human ↓
ResNet-50	26M	61.7	56.0 (-5.7)
ResNet-101	45M	60.1	54.5 (-5.6)
ResNet-152	60M	58.4	54.3 (-4.1)
DeiT-Small	22M	66.5	60.7 (-5.8)
DeiT-Base	86M	61.2	56.2 (-5.0)

Big Data Benefits Big Models

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

What's Next?

GTA-Human++

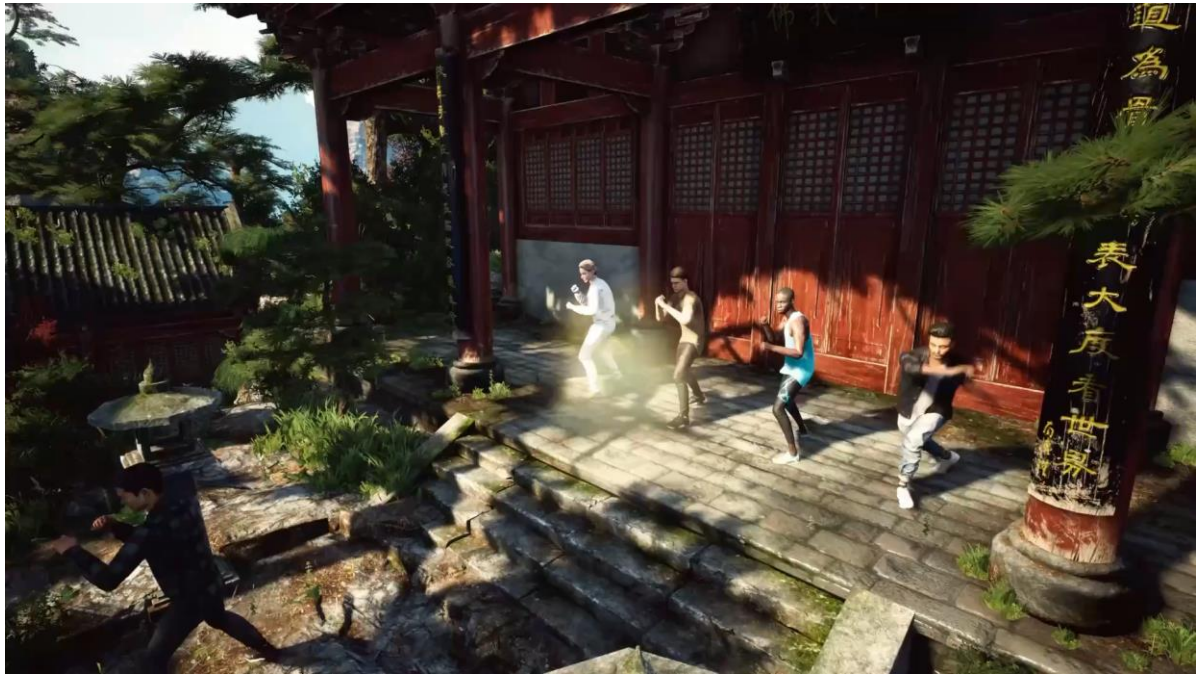


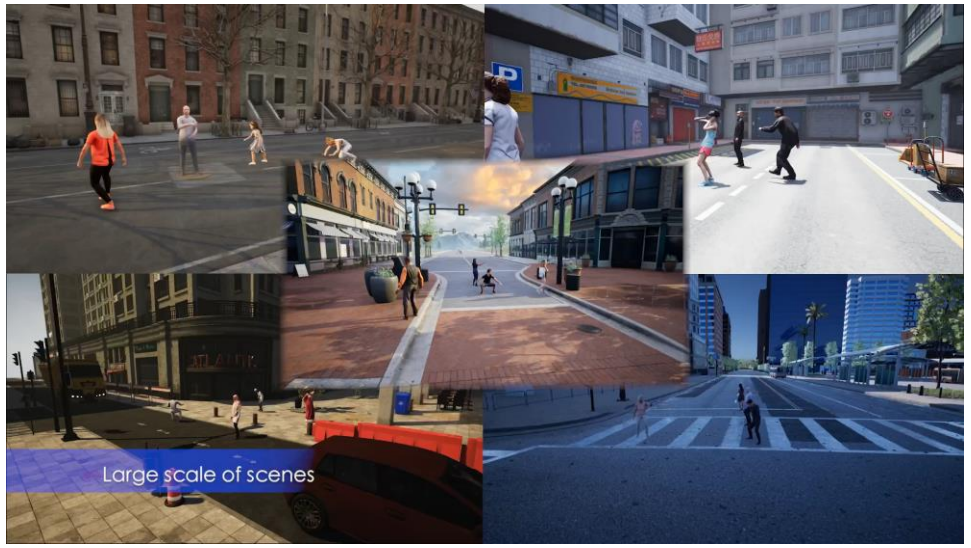
Multi-person

SMPL-X

More modalities
e.g. Point Clouds

SynBody (Full Set Coming Soon!)





https://openxdlab.org.cn/details/SynBody

That’s all for GTA-Human

GTA-Human

(Homepage)

MMHuman3D

(Perception Toolbox)

XRMocap

(Toolchain)

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

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

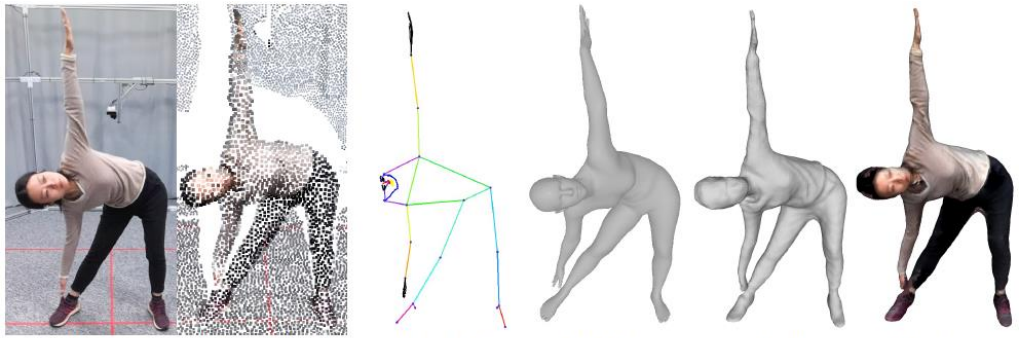
商汤 sensetime

香港中文大學
The Chinese University of Hong Kong

TSINGHUA UNIVERSITY 1911

Overview

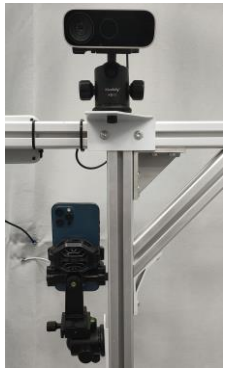
Dataset	#Subj	#Act	#Seq	#Frame	Video	Mobile	Modalities							
							RGB	D/PC	Act	K2D	K3D	Param	Mesh	Txtr
UCF101 [85]	-	101	13k	-	✓	-	✓	-	✓	-	-	-	-	-
AVA [20]	-	80	437	-	✓	-	✓	-	✓	-	-	-	-	-
FineGym [82]	-	530	32k	-	✓	-	✓	-	✓	-	-	-	-	-
HAA500 [14]	-	500	10k	591k	✓	-	✓	-	✓	-	-	-	-	-
SYSU 3DHOI [26]	40	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	-	-	-	-	-	-	-	✓	✓	✓
HuMMan (ours)	1000	500	400k	60M	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

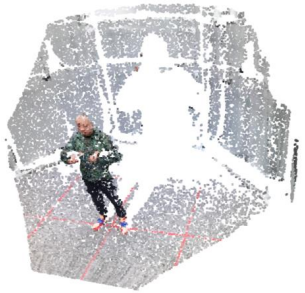


a) Color Image b) Point Cloud c) Keypoints d) SMPL e) Mesh f) Texture

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

Complete and Unambiguous Action Set (500)





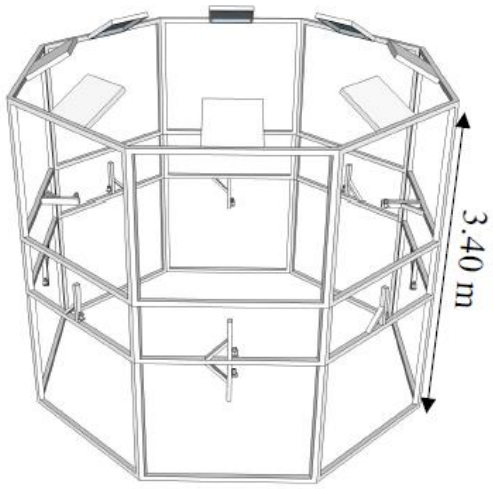
a) Kinect (ID 0) b) iPhone

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

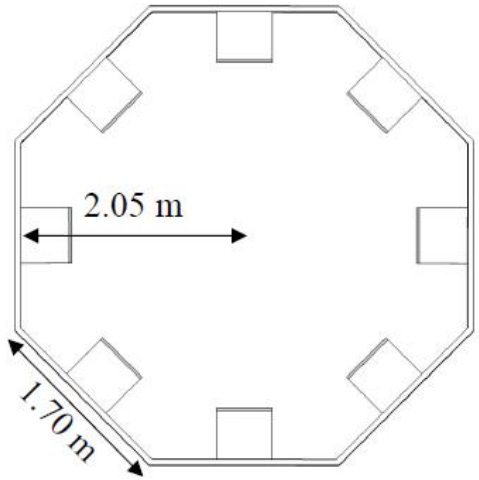
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Hardware

Hardware



a) Perspective view



b) Top view

c) Sensors

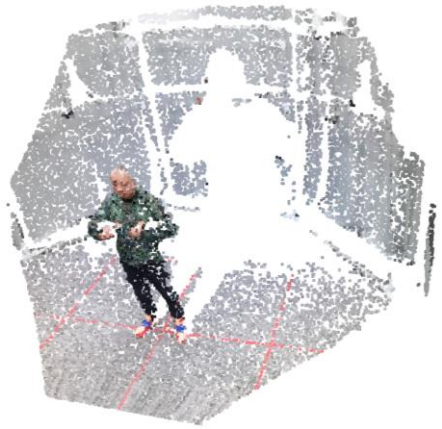
Hardware Solution

Microsoft Azure Kinect

Artec Eva High-resolution Scanner

iPhone 12 Pro Max (with LiDAR)

Sensor Suite



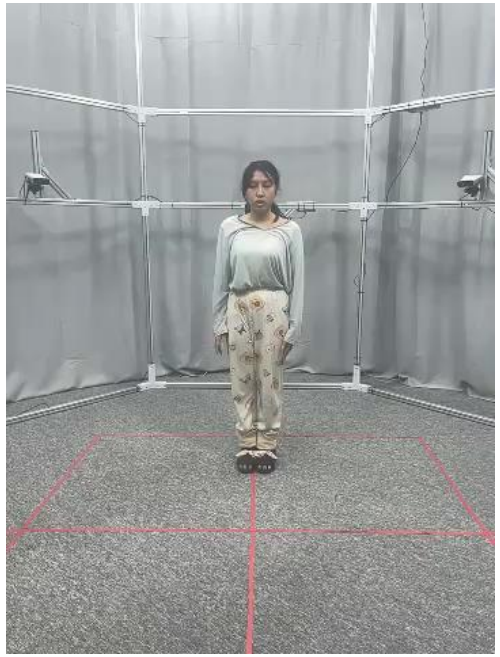
a) Kinect (ID 0)

b) iPhone

Visualization

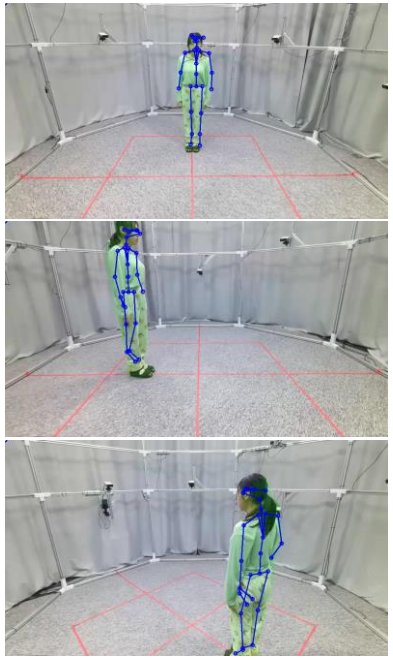
Data Collection

Artec Eva

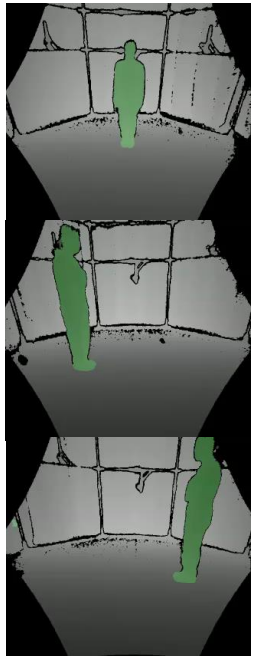


iPhone RGB

iPhone Depth



Kinect RGB



Kinect Depth

0.1 mm

Scan Accuracy

11

Views

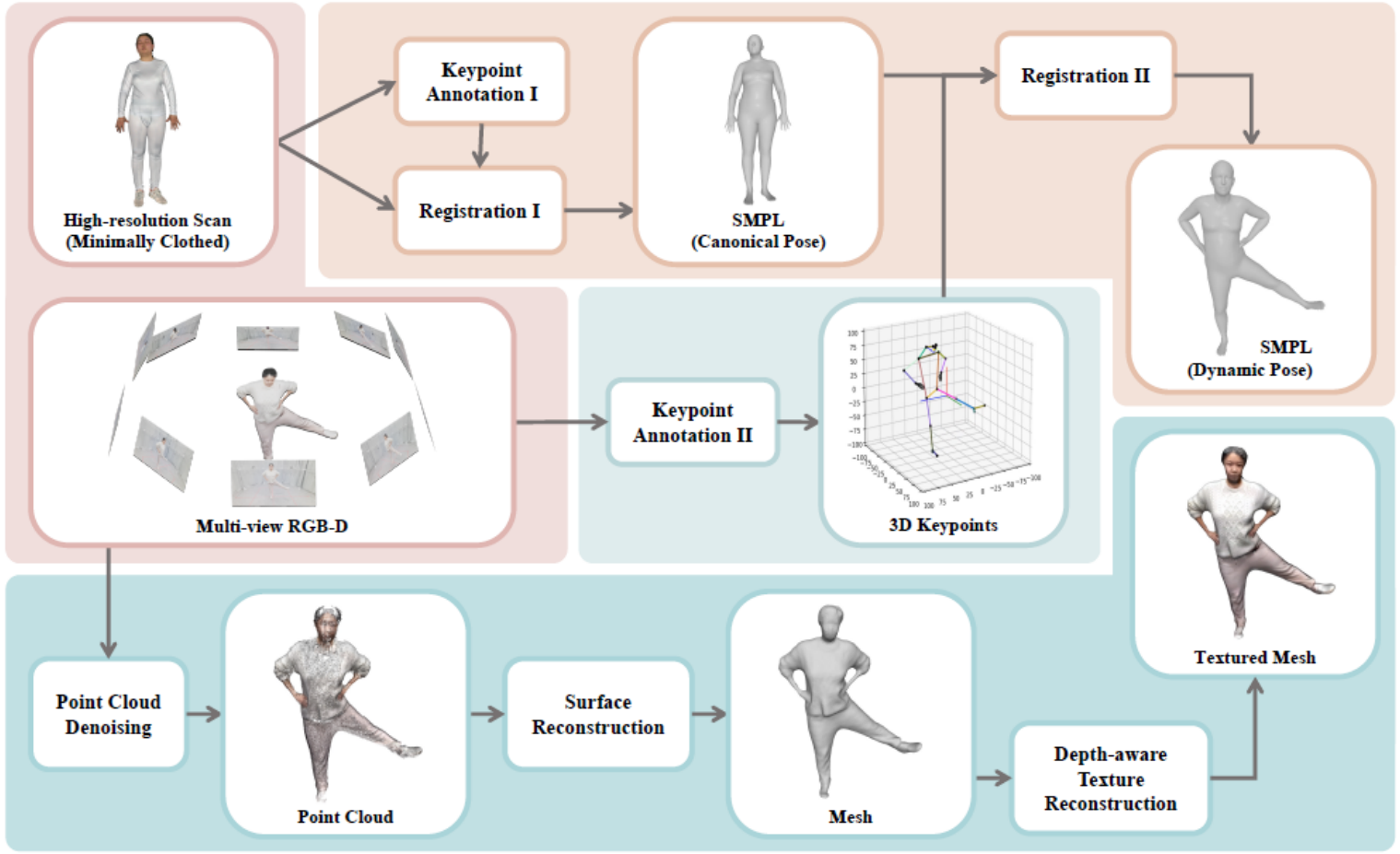
1G

Data / second

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Toolchain

Toolchain

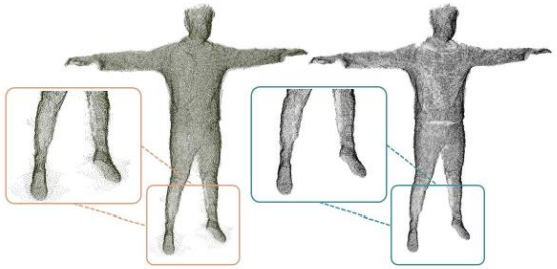


Toolchain

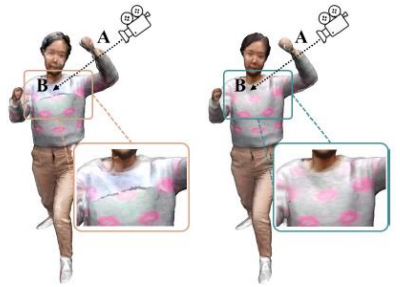
Mesh Geometry

Toolchain

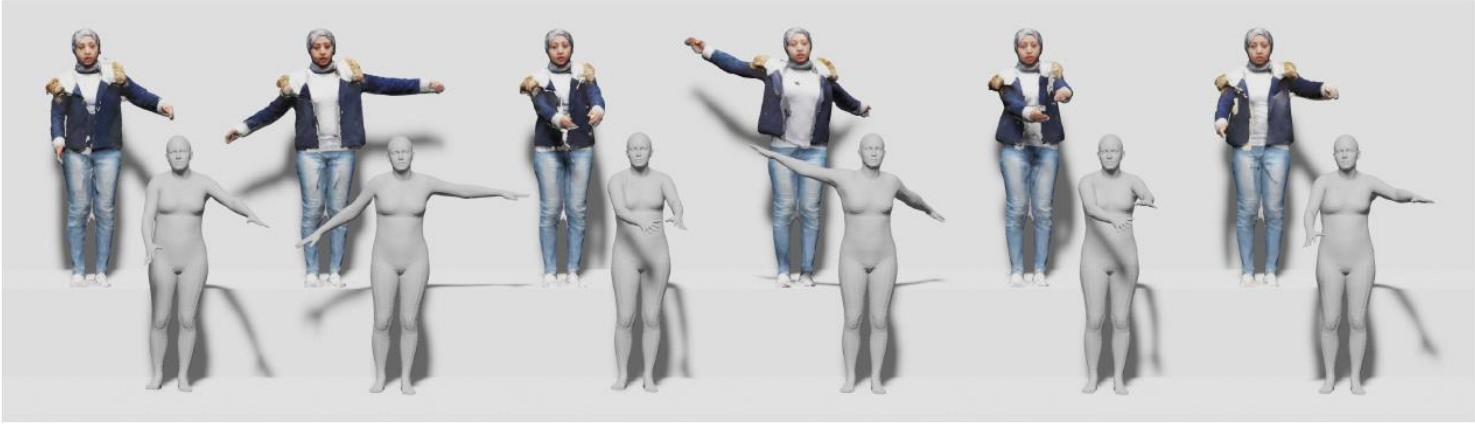
Shape Registration



Point Cloud Denoising



Depth-Aware Texture Reconstruction

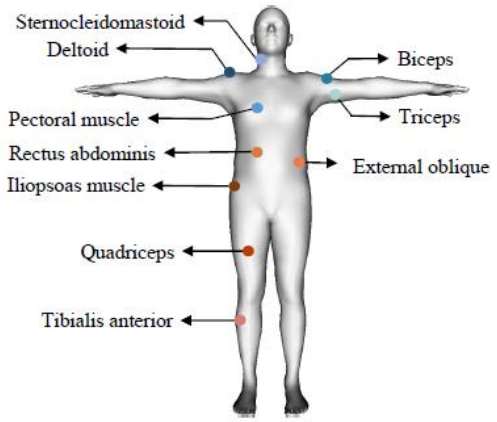


Parametric Model and Textured Mesh Sequences

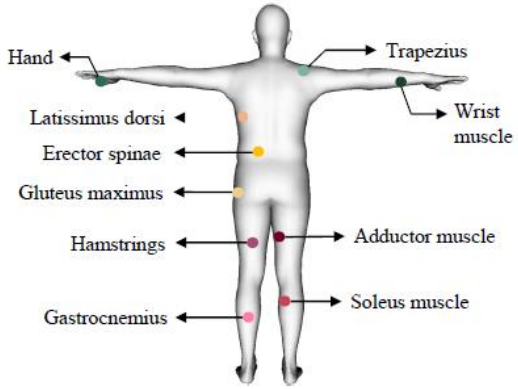
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Action Set

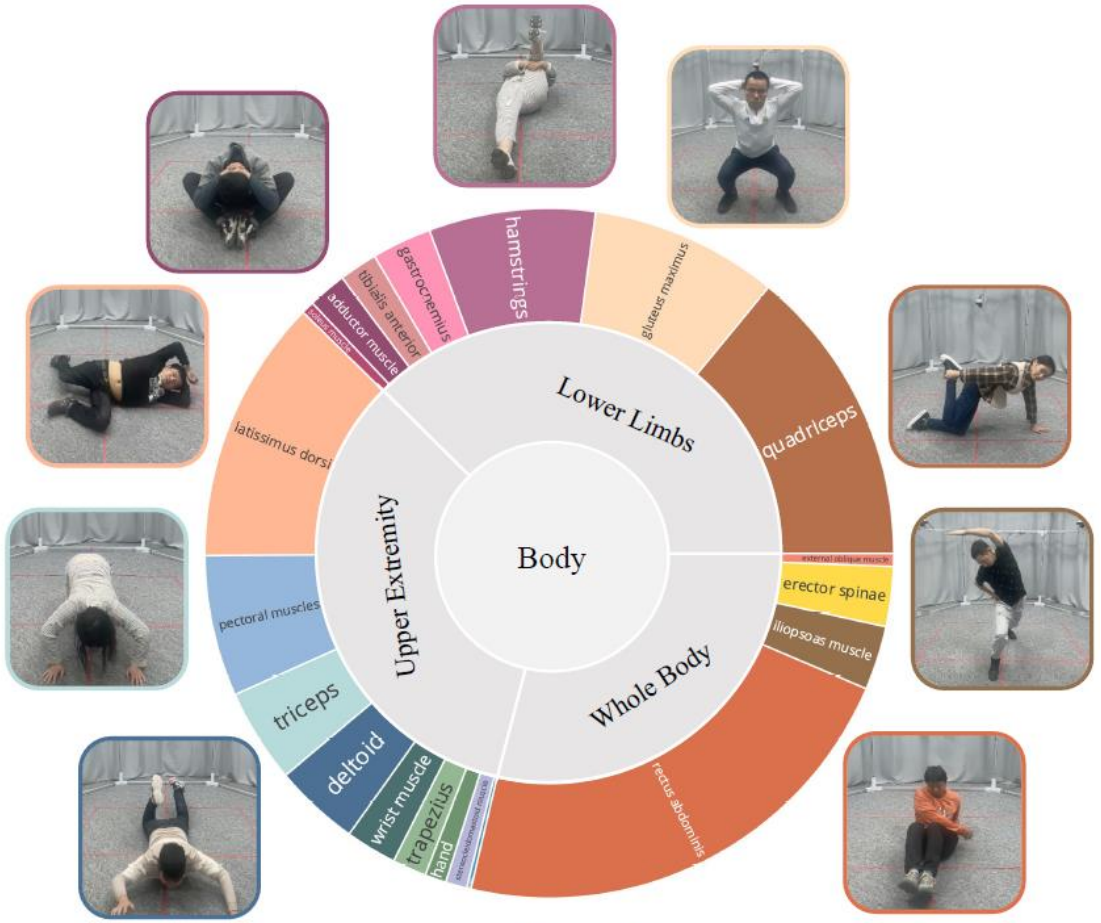
Action Set



(a) Schematic Diagram (Front View)



(b) Schematic Diagram (Back View)



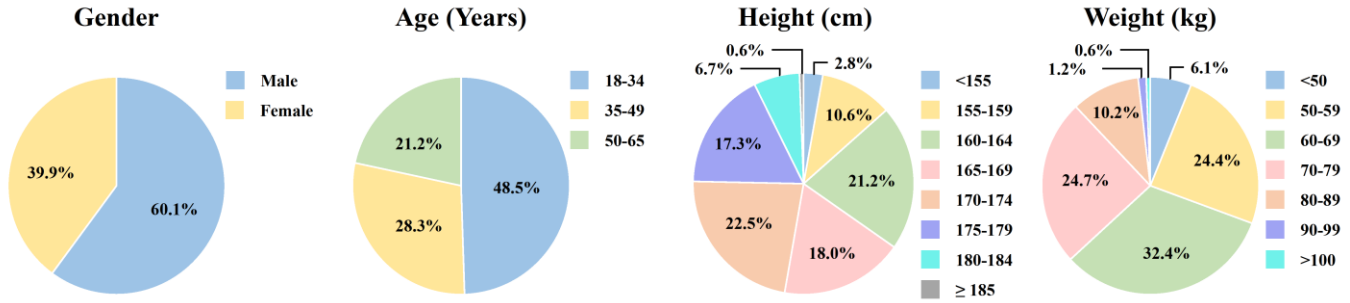
(c) Action Hierarchy Diagram

Hierarchical, Complete and Unambiguous

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Subjects

Subjects



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

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Updates on HuMMan v1.0

Reconstruction Subset (Just Released!)

153 Subjects

339 Sequences

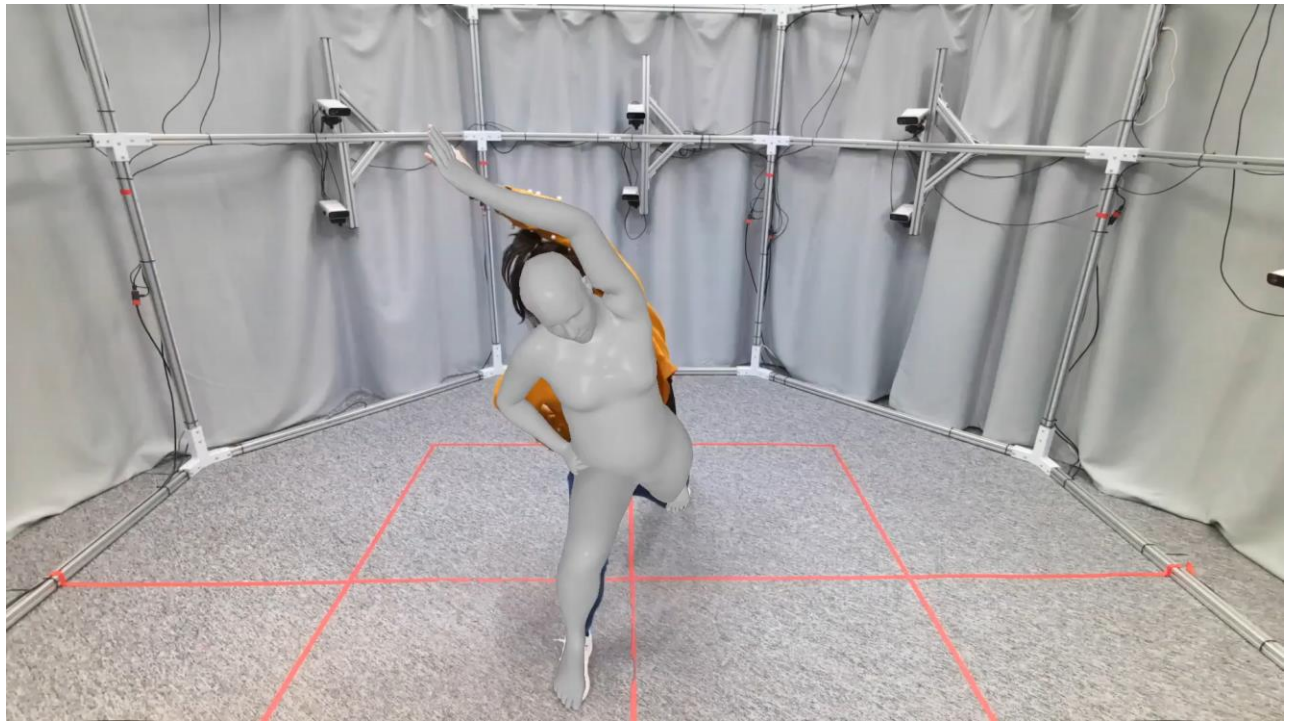
Realistic Challenges

Methods	Novel Pose	Novel View
NHP [1]	<18	<18
MPS-NeRF [2]	<18	<18
? [?]	~20	~20

Generalizable Animatable Avatar from Single Image
(Metric: PSNR)

[1] NHP Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering
[2] MPS-NeRF: Generalizable 3D Human Rendering from Multiview Images

Action Understanding Subset (Coming Soon!)

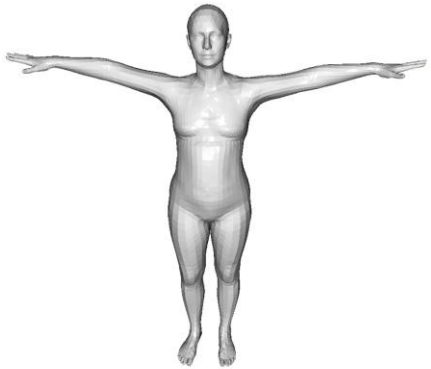


Lunge Twist Stretch L (ID=11):

Put your **right leg** in the front and your **left leg** behind. Bend your **right knee** about 90 degrees while keep your **left knee** straight with **left toes** on the ground. Keep your **body** in the upright position and tilt to your right. Straighten your **left arm** and keep it close to your **left ear**. Put your **right hand** on the right side of the **waist**, with **right elbow** bent at 90 degrees.

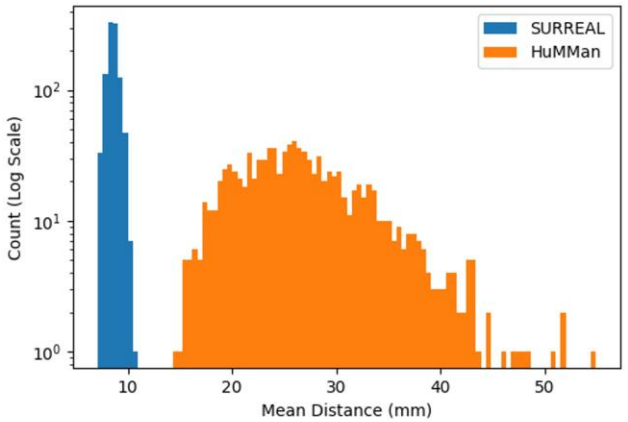
Perception Subset (Coming Soon!)

a) i) HuMMan Color Image

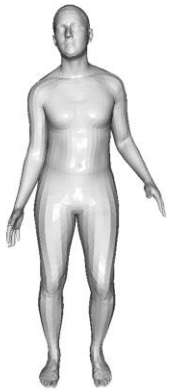


a) ii) HuMMan SMPL

a) iii) HuMMan Point Cloud



b) Point Distribution Analysis



c) i) SURREAL SMPL

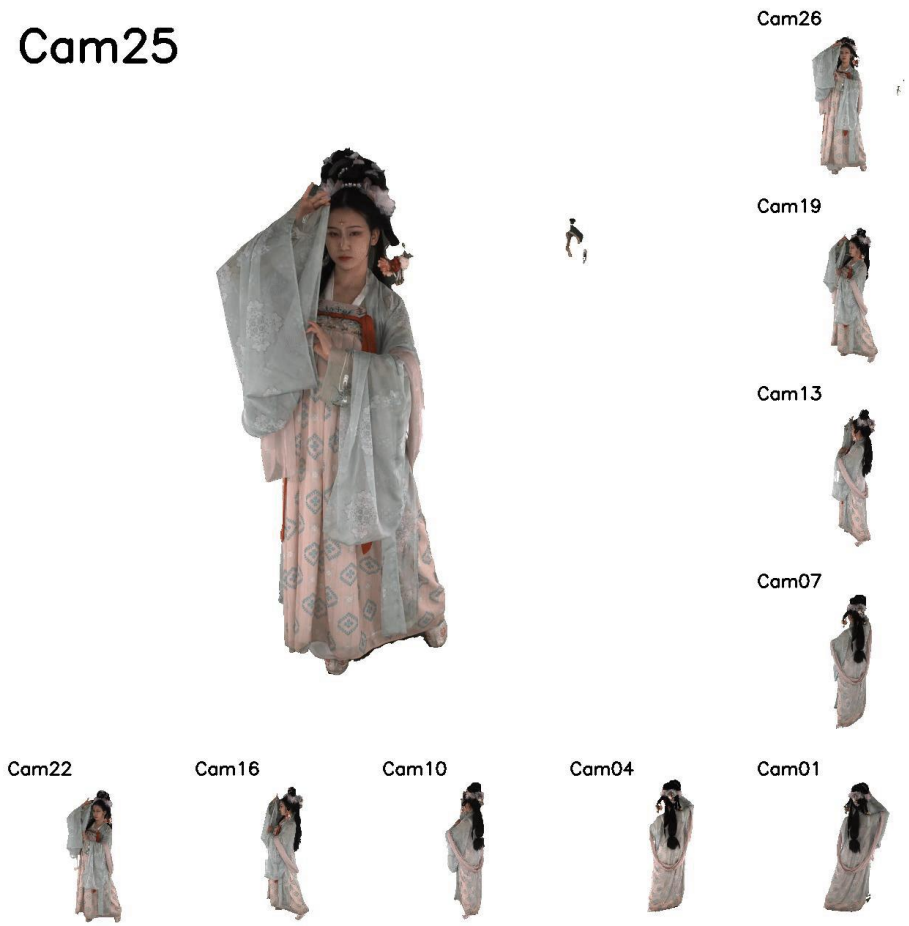
c) ii) SURREAL Point Cloud

Point Cloud-Based Human Pose and Shape Estimation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

What's Next?

RenBody (Full Set Coming Soon!)



https://openxdlab.org.cn/details/RenBody

Thank you!

HuMMan (Homepage)

GTA-Human (Homepage)

MMHuman3D (Perception Toolbox)

XRMocap (Toolchain)