WA.L409.2

ADAPTIVE AND ROBUST MMWAVE-BASED 3D HUMAN MESH ESTIMATION FOR DIVERSE POSES

2023.10.12

†Kotaro Amaya, †‡Mariko Isogawa

†Keio University ‡JST sakigake

Keio University
1858
CALAMVS
GLADIO
FORTIOR

Existing Human Pose/Mesh Estimation Methods

- Methods for measuring, estimating, and digitizing human posture
- Most methods use RGB images/videos captured by conventional cameras, and there are also optical methods that utilize infrared cameras or inertial sensors

Input: RGB Image

Model

[Wang+ arXiv:2207.02696]

Output: **Human Pose**
2D/3D joint positions

Input: RGB Image

Model

[Kanazawa+ CVPR2018]

Output: **Human Mesh**
Mesh model parameters
(e.g., Skinned Multi-Person Linear model (SMPL model))

Disadvantages of Using RGB Images/Videos (1/2)

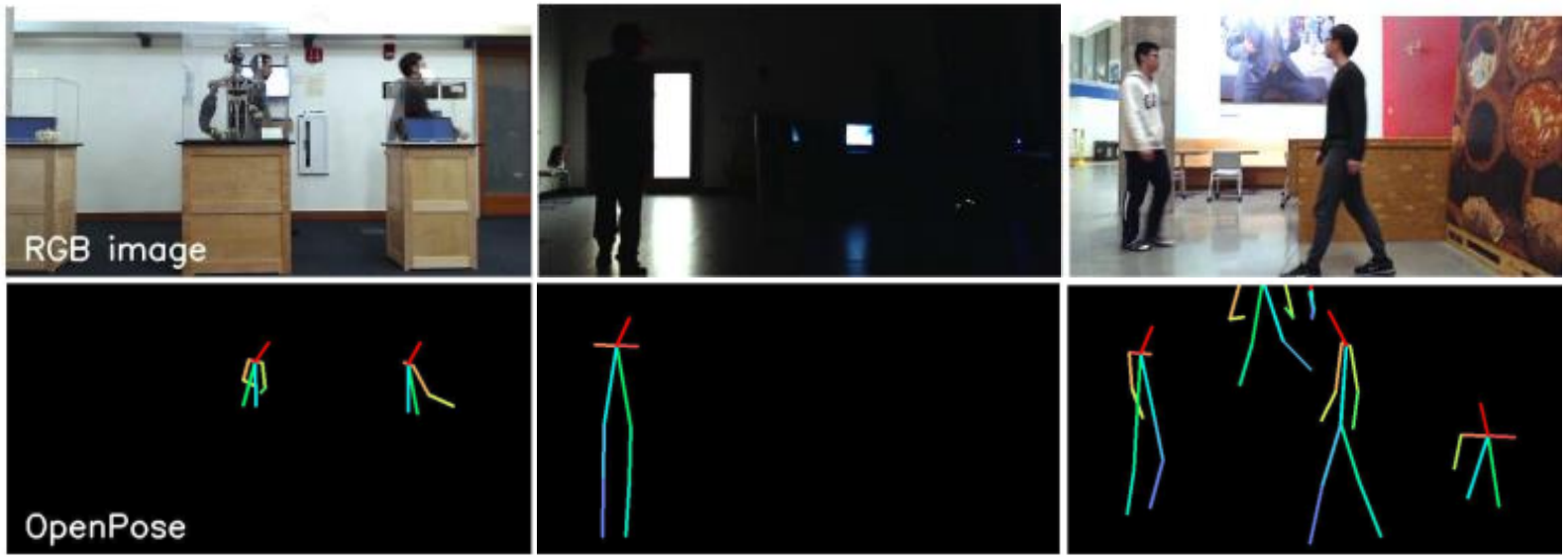
Privacy considerations are challenging with RGB camera implementations.

- Using Images that can identify individuals in detail
- The location can be determined from the background

Disadvantages of Using RGB Images/Videos (2/2)

Sensitive to

- Light conditions (e.g., night road, bedroom)
- Occlusions (e.g., feet are hidden under the table and not visible)



Occluded image | Low-light condition | False Positive Due to Human Posters

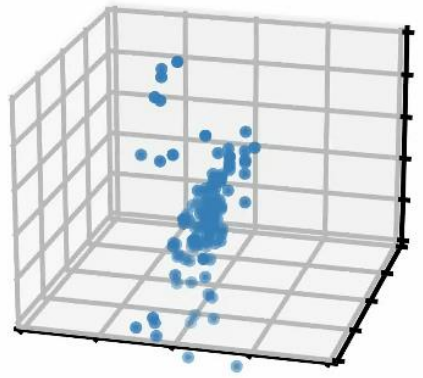
Examples of inference failures with RGB video imaging[1]

[1] Zhao, Mingmin, et al. "Through-wall human pose estimation using radio signals." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Human Pose/Mesh Estimation Using Wireless signals

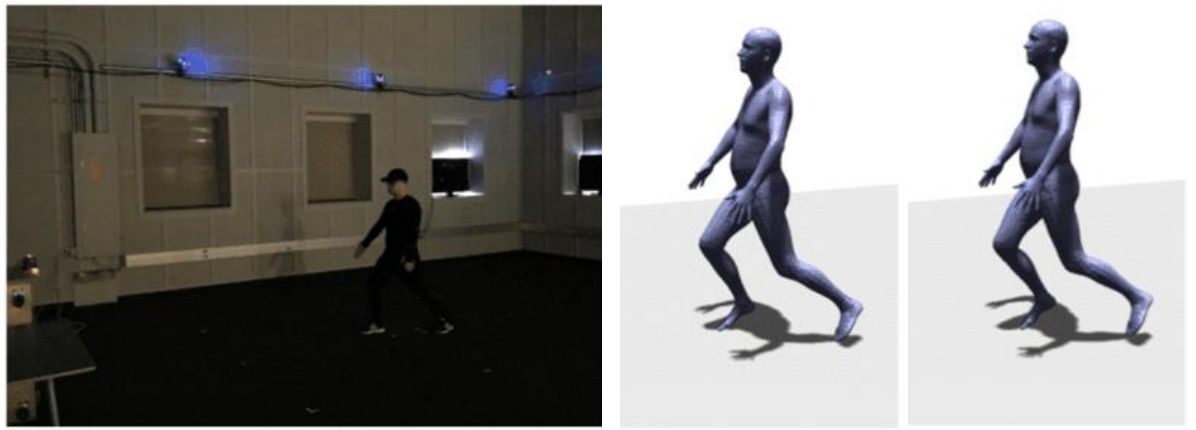
Methods using Wi-Fi(2~5GHz), Millimeter Wave(76~81GHz)

- Able to infer through obstructions like walls and fabric
- Can be used in dark places
- So sparce signals that can consider privacy



Captured Signals

Human Pose estimation using radio signals in 2~5GHz[1]



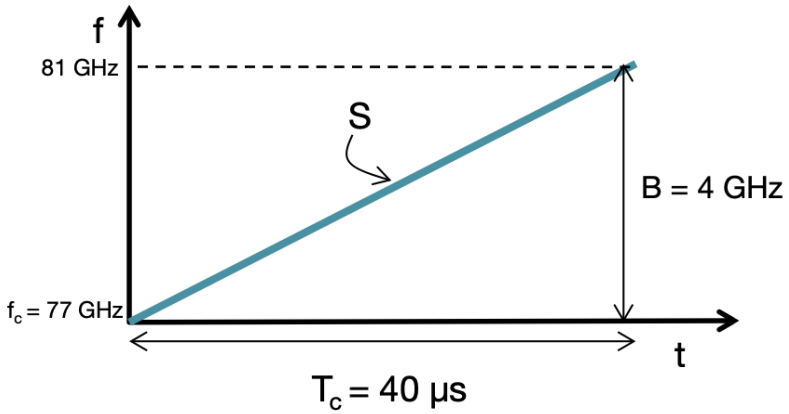
Human mesh estimation using millimeter waves in dark places[2]

[2] Xue et al., "mmMesh: Towards 3D Real-Time Dynamic Human Mesh Construction Using Millimeter-Wave", In MobiSys, pp.269-282, 2021.

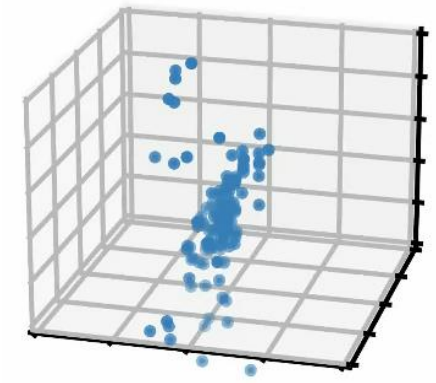
Millimeter Wave Sensor

- Transmit millimeter-wave chirp signals from multiple transmission antennas
- Receive reflected signals with multiple receiving antennas

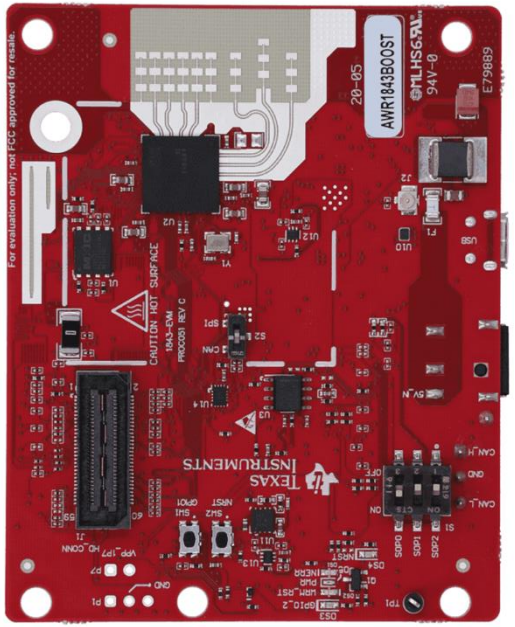
3D position, velocity, reflection intensity,
and azimuth angle can be obtained



Chirp signal[6]



Captured Signal



Millimeter wave sensor[7]

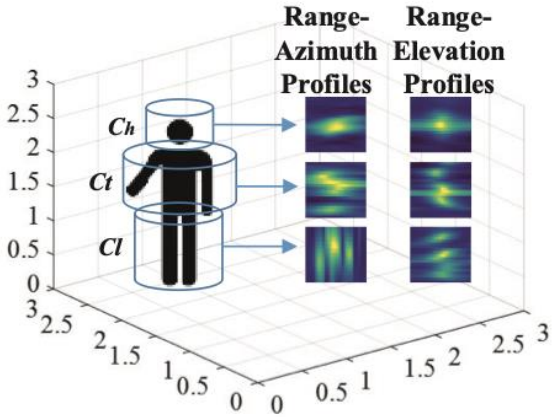
[7] Texas Instruments AWR1843BOOST, https://www.ti.com/tool/ja-jp/AWR1843BOOST, (Jan, 2023)

Previous Research

mmMesh[2], m^{3}Track[3]：Methods using millimeter wave sensor

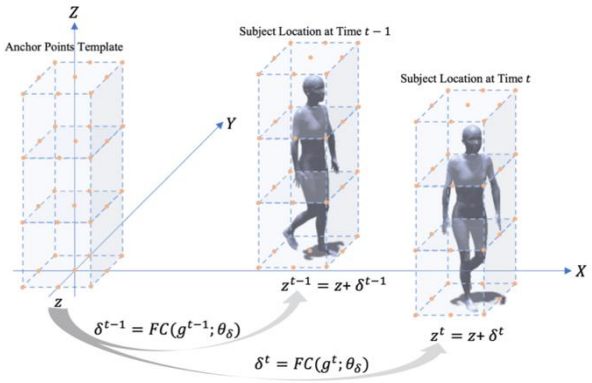
- mmMesh uses a cuboid, m3Track uses a cylinder for feature extraction
  → They assume that the person's posture is upright
- They use data where only the person exists in the measurement space
  → They do not consider the presence of objects other than the person

m^{3}Track



Feature extraction using cylinder

mmMesh



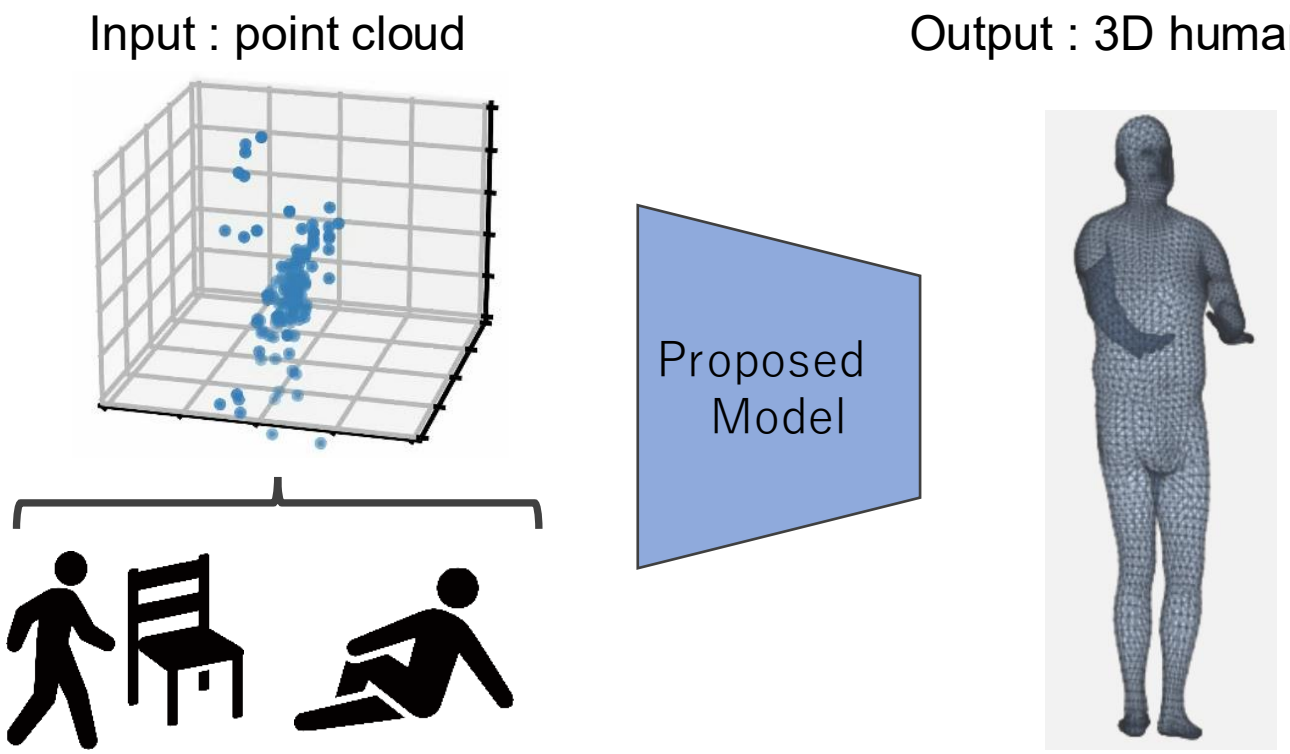
Feature extraction using cylinder

Common postures and environments

[3] Kong et al., “m3track: mmwave-based multi-user 3d posture tracking”, in Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services, pp. 491–503, 2022

Purpose of this research

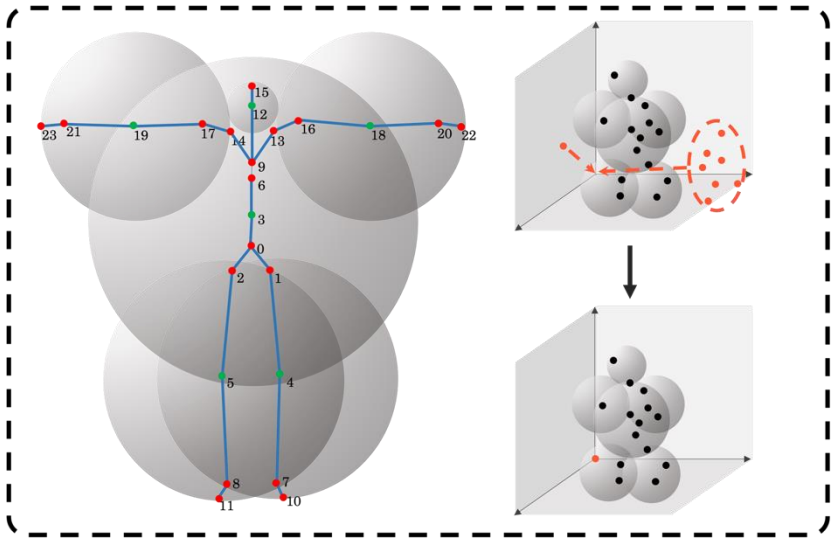
- Achieving robust methods that can handle the presence of objects other than the person in the measurement space
- Achieving methods that are robust to postures other than upright, such as lying horizontally or sitting on the ground



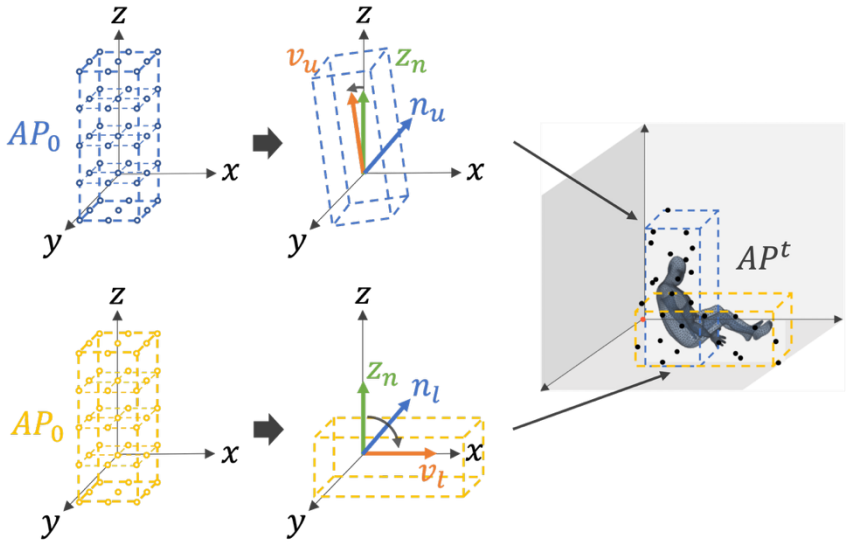
The focal point of this research

Key Point 1：
Denoising process using a spherical set that fully encompasses the human range of motion in any posture

Key Point 2：
Feature extraction capable of handling various postures by using rectangular boxes divided into upper and lower body segments



Denoising using spheres

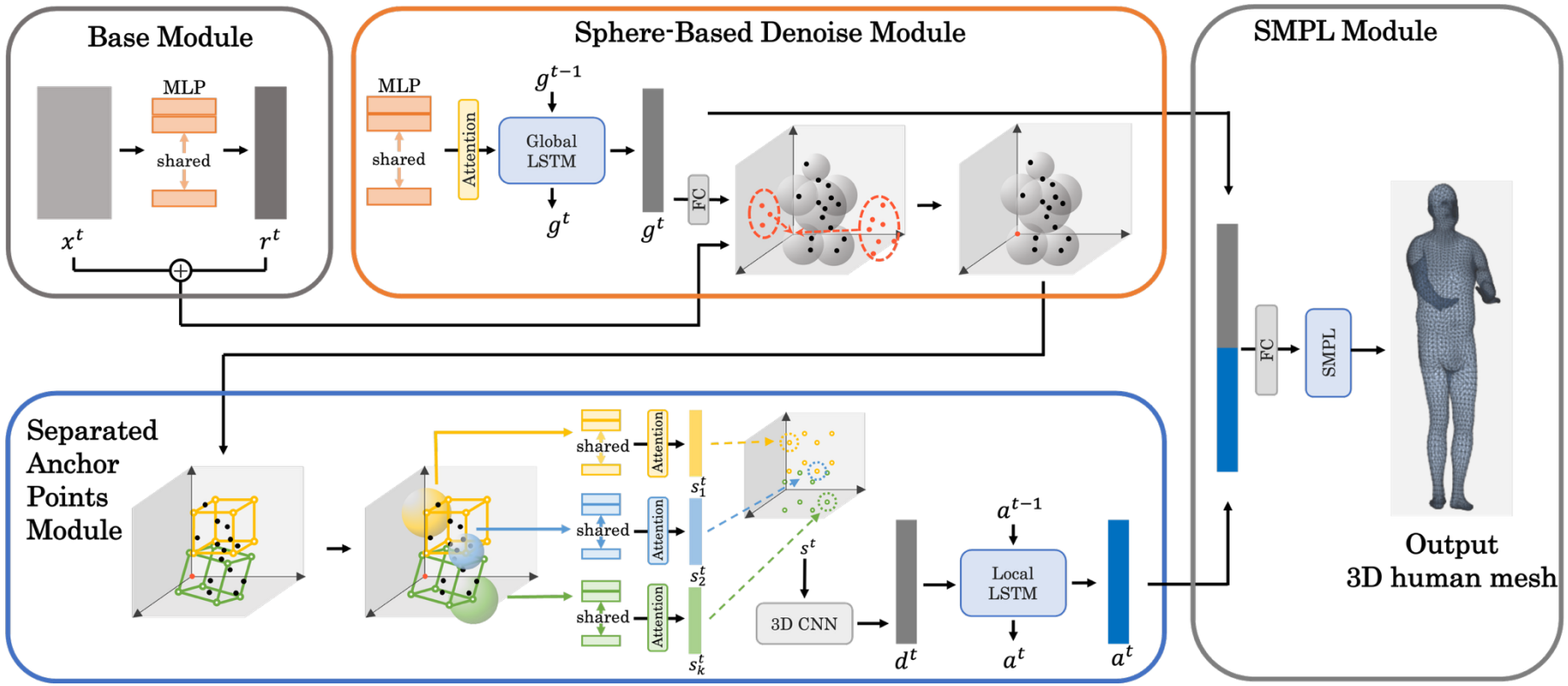


Feature extraction
using segmented rectangular boxes

Model Architecture

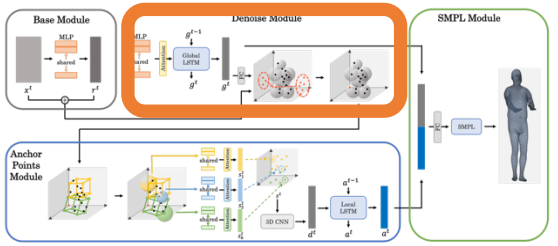
Proposed method consists of Base Module, Sphere-Based Denoise Module, Separated Anchor Points Module, SMPL Module

Global feature extraction and denoising using a spherical set



Feature extraction using segmented rectangular boxes

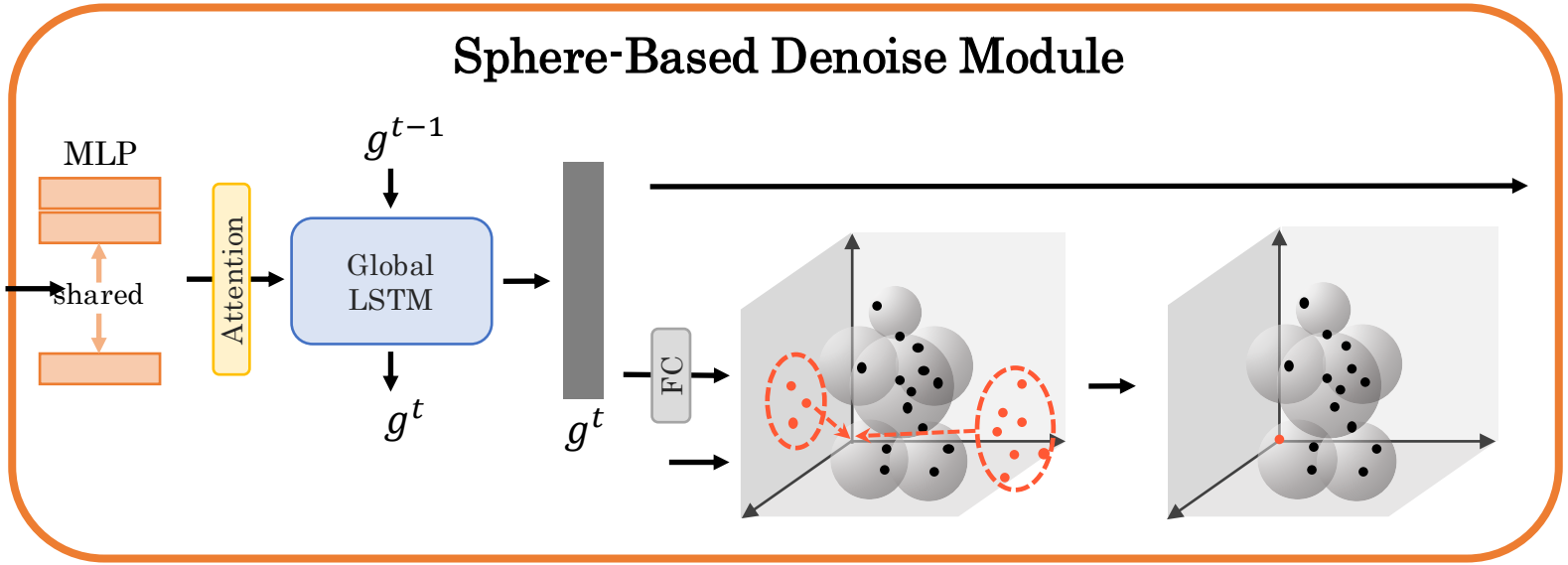
Key Point 1: Sphere-Based Denoise Module



- Extracting features g^t with Shared-MLP, Attention and LSTM

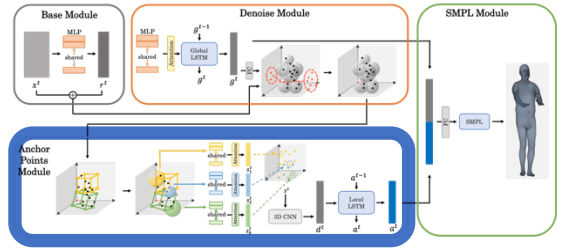
  $\rightarrow g^t$ is used as input to SMPL Module
- g^t is used to infer six joint positions j^t and radius r^t each through fully connected layers

  $\rightarrow$ Points not present within the six spheres are considered noise and set to 0

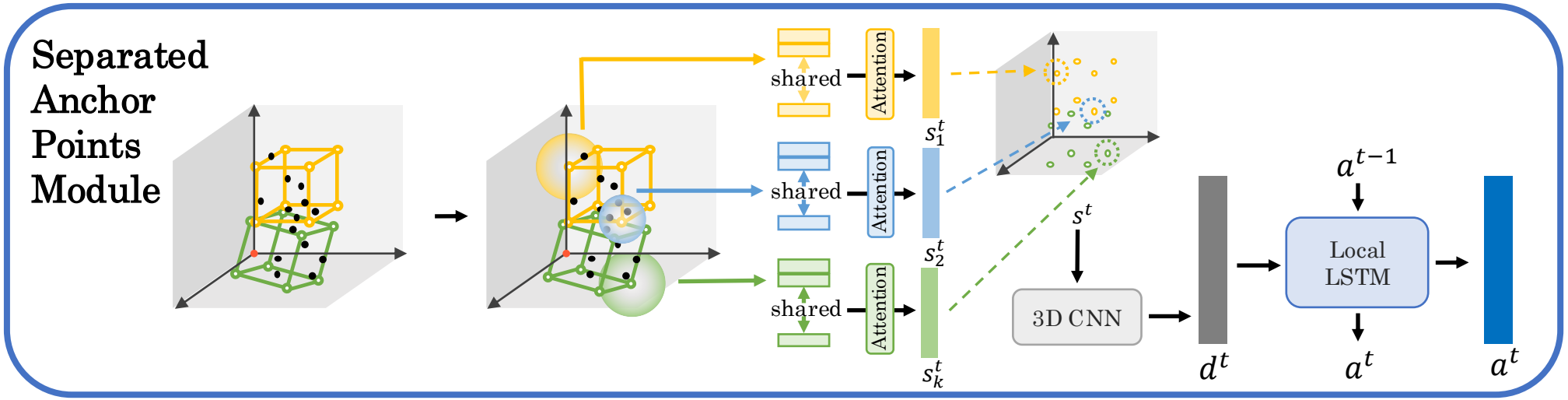


Model architecture of
Sphere-Based Denoise Module

Key Point 2: Separated Anchor Points Module



- From j^t and r^t estimated in Denoise module, define Separated Anchor Points
- Arrange two rectangular to align with the upper and lower body
- After feature extraction at each anchor point, perform position encoding using 3D CNN

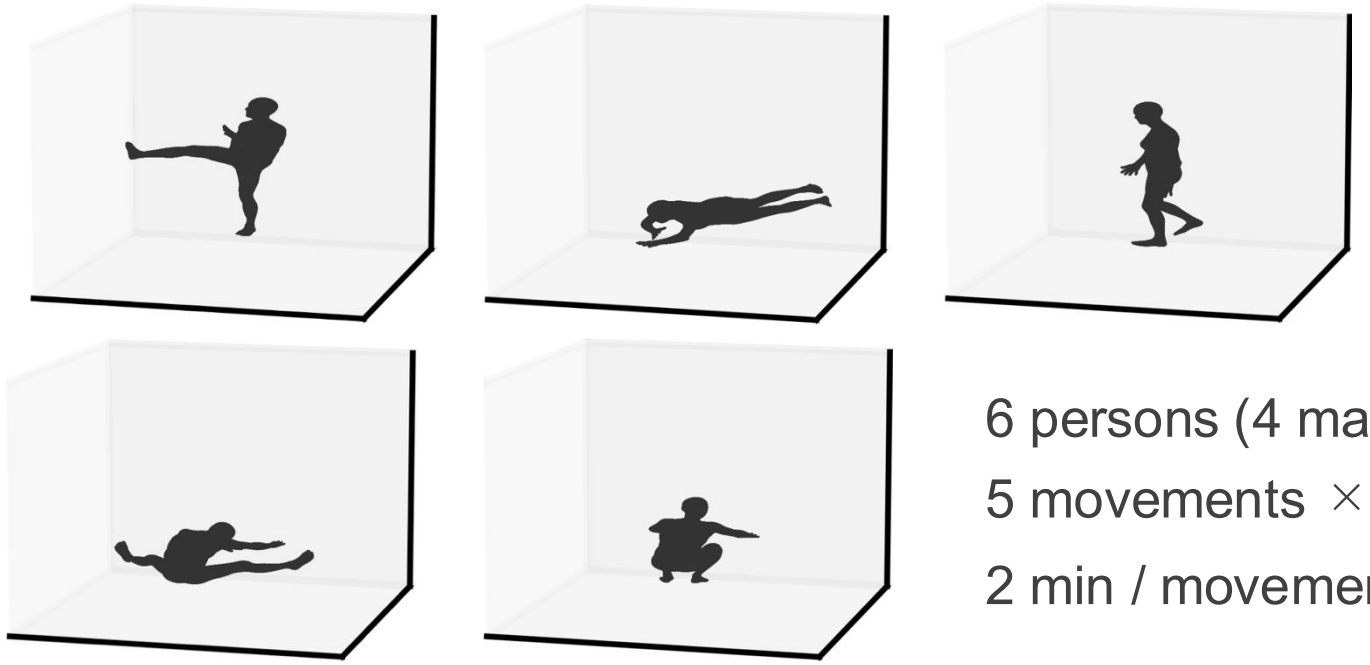


Model architecture of
Separated Anchor Points Module

Dataset

We have created a custom dataset that includes the following features

- When objects other than a person are present in the measuring space
- Postures of a person other than standing, such as lying horizontally or sitting on the ground



6 persons (4 male, 2 female)

5 movements × beside a box or not

2 min / movement

Postures in the dataset

Quotative results

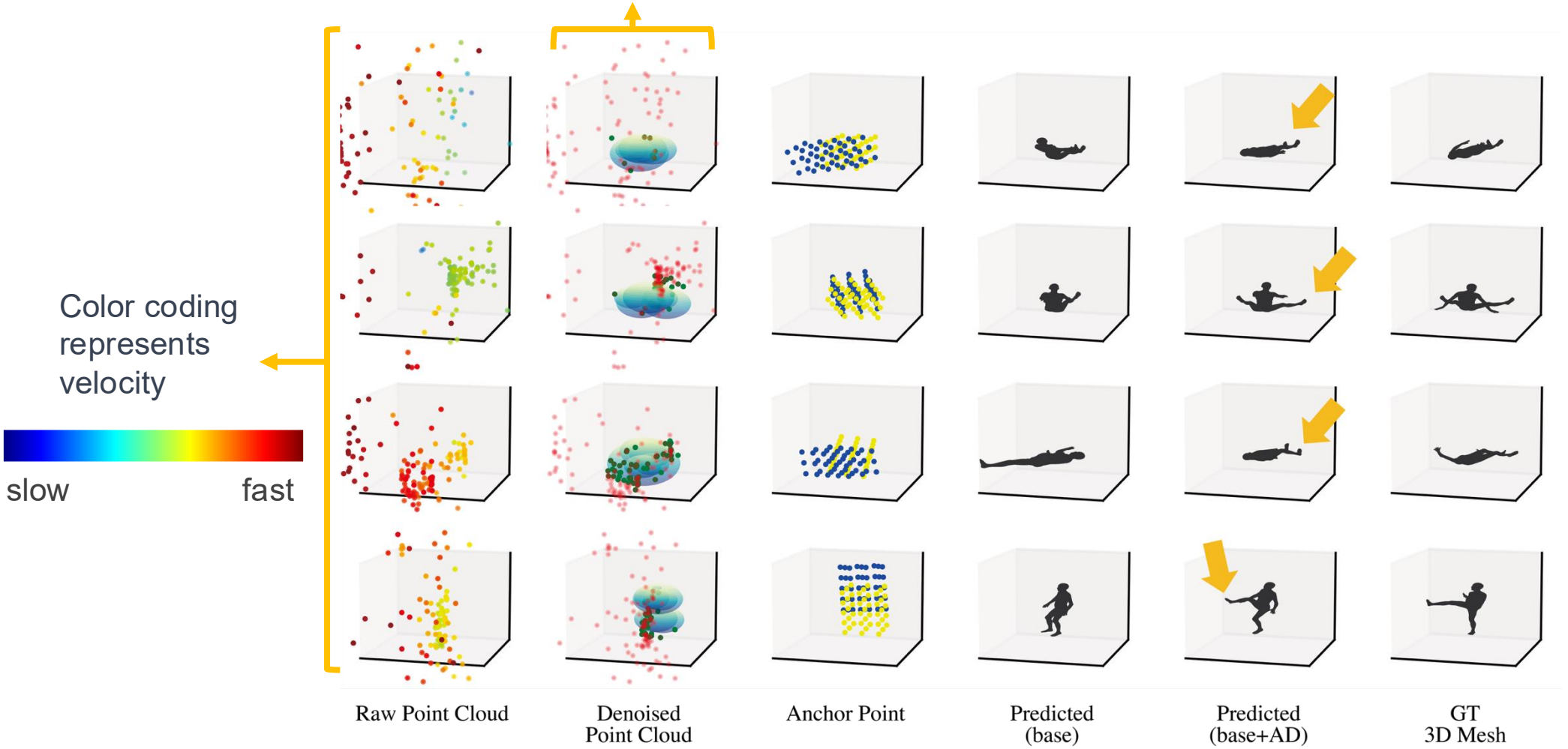
The proposed method demonstrates equivalent or superior across all evaluation metrics

- Approximate 10% improvement in terms of $MPJPE$ for joint position errors
- Approximate 7% improvement for E_{ver} for vertices errors

	Evaluation metrics					
Method	E_{rot} (°)↓	E_{ver} (cm)↓	MPJPE (cm)↓	PCKh @0.5↑	E_{loc} (cm)↓	E_{body} (cm)↓
Xue *et al.* [2]	33.8	35.7	33.2	0.23	30.1	**0.88**
Ours	**32.3**	**33.3**	**30.4**	**0.25**	**23.3**	**0.88**

Qualitative result

Red : Denoised point cloud outside the sphere
Green : point clouds inside the sphere



Summary

For a robust millimeter-wave-based method of estimating a person's 3D mesh in dark or obstructed areas, with consideration for privacy,
we propose the following to address the traditional challenges, such as vulnerability to environmental noise and limited versatility in estimating postures.

- 'Sphere-based Denoise Module' for extracting the human region in any posture
- 'Separated Anchor Points Module' to accommodate a wider variety of postures

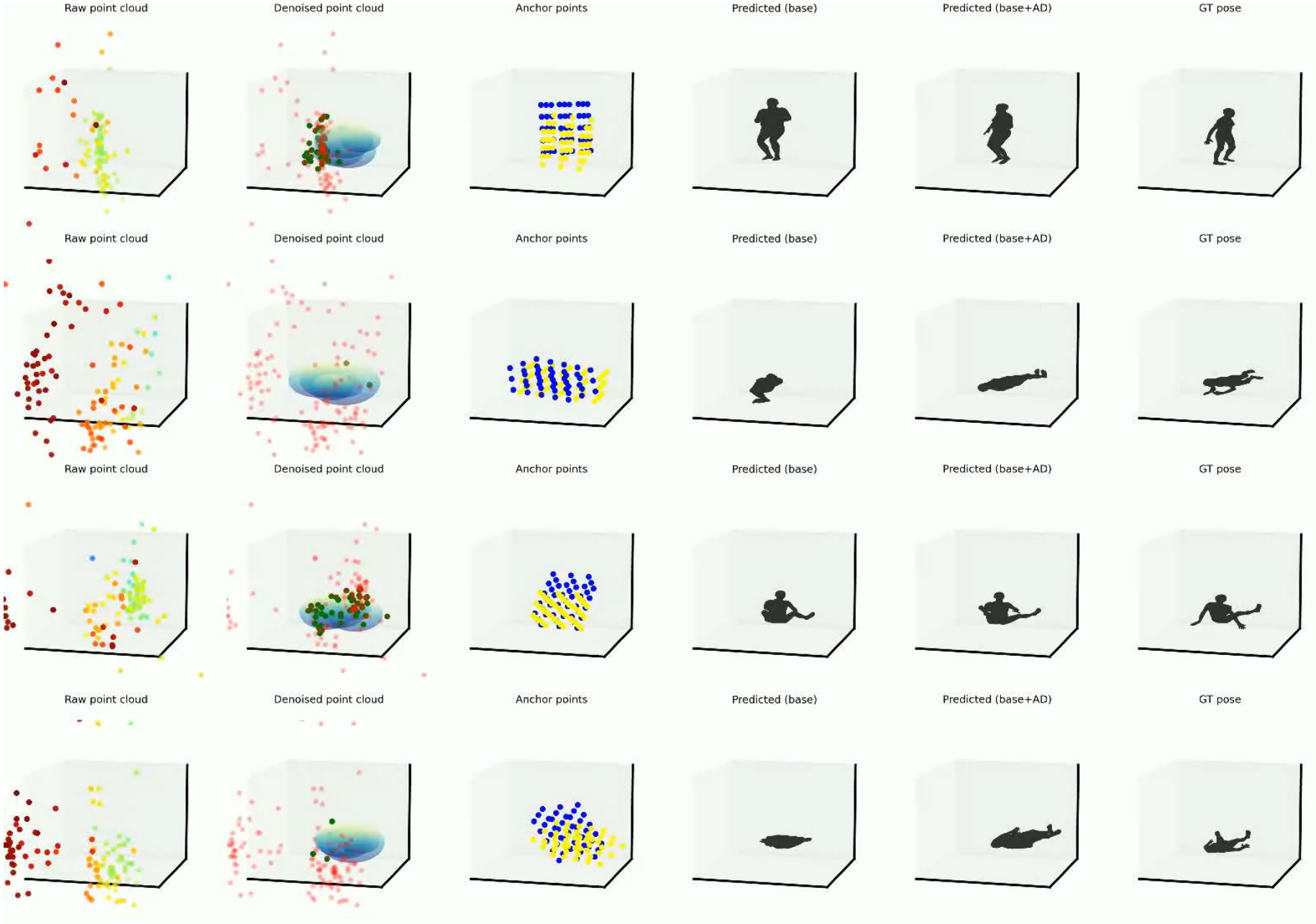
Future Challenges and Prospects

- Further improvement in the versatility of postures that can be accurately inferred
- Introduction of Anchor Points with higher expressiveness

Appendix

Appendix:

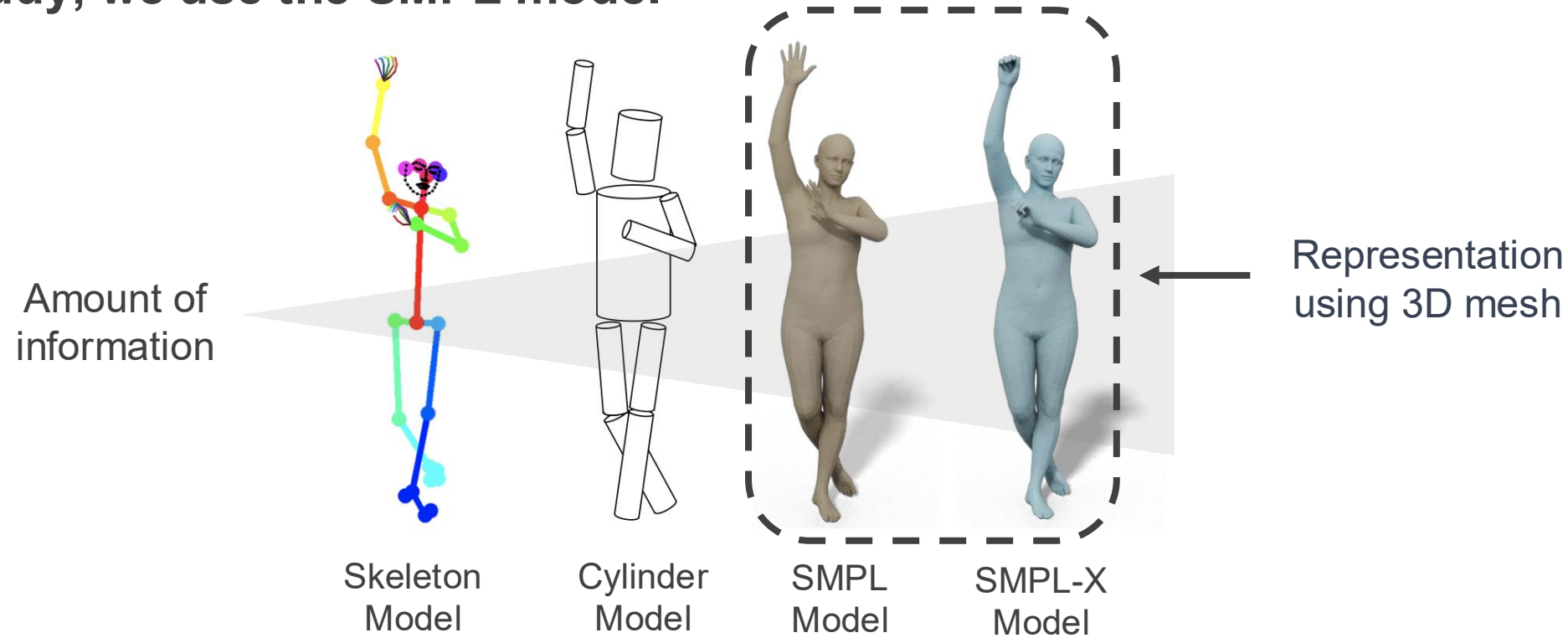
Sample Movie of result



Raw Point Cloud | Denoised Point Cloud | Anchor Point | Base+A+D (Ours) | Base (mmMesh) | Groud Truth

Appendix: Human Models

- **In human models, not only the skeleton but also the shape can be estimated**
- **The more detailed the representation of a person, the wider the applications. In this study, we use the SMPL model**



Human Models[3]

[3] Tian et al., "Recovering 3D Human Mesh from Monocular Images: A Survey", CVPR 2022

Appendix: Loss Function

$$Loss = \sum_{K \in (\boldsymbol{V},\boldsymbol{S},\boldsymbol{B},\boldsymbol{j})} \alpha_K * \sum_t^T ||K^t - GT(K^t)||_{L_1} + \alpha_G * \sum_t^T H(G^t, GT(G^t)) + L_l + L_r$$

V : Vertices position of human mesh

S : Position of skeleton joints position

B : Parameters of SMPL model

j : Root joint points

G : gender

H : Hinge loss

$$\boldsymbol{L_l}^t = \frac{1}{6}(||\boldsymbol{j}^t_{xyz} - \boldsymbol{T}^t_1||_{L1} + \sum_{i=2}^{6} ||\boldsymbol{j}^t_{xyz} - \boldsymbol{Sk}^t_i||_{L1})$$

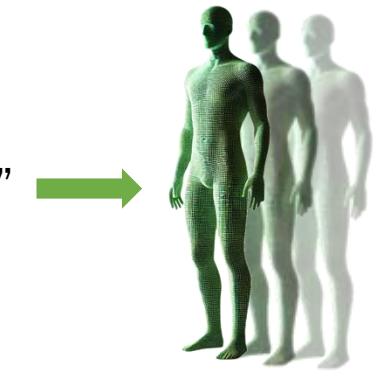
$$\boldsymbol{L_r}^t = \frac{1}{6}\sum_{i=1}^{6}(r_i^{sk} - r_i^j)$$

Supervised learning of joint positions using the Denoise Module

Appendix: Evaluation Metrics

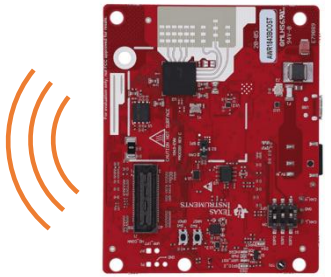
Below, T denotes time, J represents the number of joints, v is the vertex position, θ is the rotation vector of the root joint, x stands for joint positions, and l is the position of the root joint

- Average Vertex Error (E_{ver})
  - $\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}||v_t^j - \hat{v}_t^j||_2$
- Average Joint Rotation Error (E_{rot})
  - $\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}|\theta_t^j - \hat{\theta}_t^j|$
- Mean Per Joint Position Error (MPJPE)
  - $\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}||x_t^j - \hat{x}_t^j||_2$
- Mesh Localization Error (E_{loc})
  - $\frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J}||l_t^j - \hat{l}_t^j||_2$

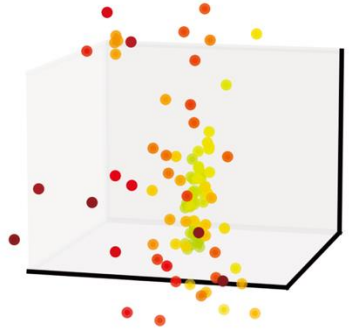


ストから
作を生成

(2) 3D環境に
人物動作を生成



(3)シミュレーションを用いた
ミリ波人物形状推定用の
大量データ生成



(4) ミリ波
人物形状推定

目1

項目2