Biomechanically Accurate Vision and Geometry based Human mesh Recovery with a single RGBD sensor

Victor Nogues[1], Hugo Rodet[1], Julia Irvine[2], Elise K. Laende[3], Manuela Kunz[4], Lama Seoud[1]
[1]Polytechnique Montréal, QC Canada [2]Queen's University, ON Canada [3] University of Waterloo, ON Canada [4] National Research Council Canada

ECCV 2026 | MALMÖ | SEPT 8–12

Problems

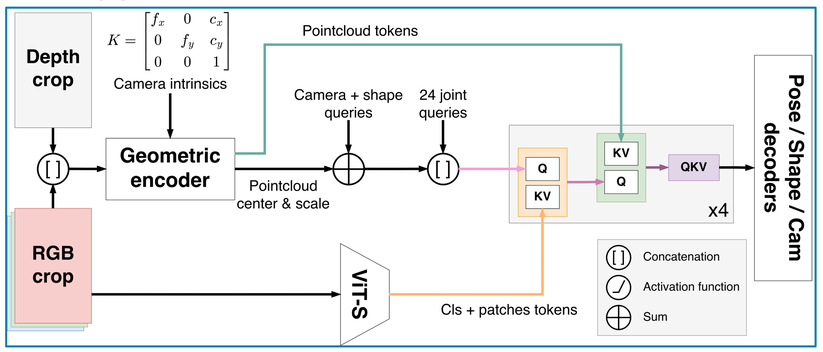
- State-of-the-art human pose and shape recovery methods rely on **huge, data-intensive** neural networks designed for universal generalization.
- Existing models are **rarely tested or validated for real-world deployment** in actual clinical environments.
- Traditional pipelines neglect the advantages of simple hardware configurations like **affordable RGBD sensors.**

Objectives

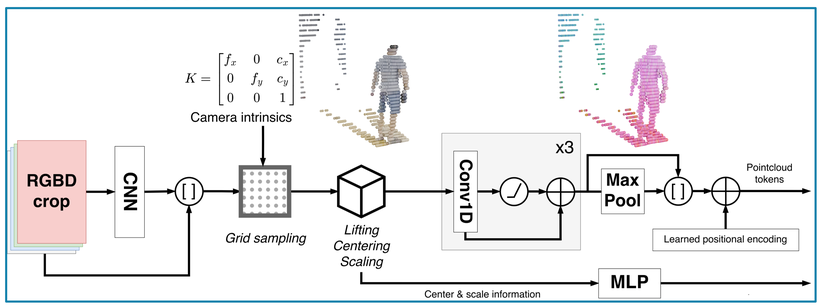
- Trade unnecessary global generality for a **lightweight,** geometry-aware architecture optimized specifically for fixed-camera clinical settings.
- Combine the streamlined design with **low-cost RGBD sensors** to maximize practicality and cost-effectiveness.
- Validate **real-world performance** directly within clinical workflows to ensure practical reliability and adoption.
- Propose an **open-source pipeline** for RGBD human pose and shape estimation to offer a baseline

Proposed fusion algorithm

Global pipeline



Geometric encoder



Dual-branch encoder - decoder architecture

- **Image Encoder:** ViT-S[1] features extractor.
- **Geometric Encoder:** Extracts features via shallow CNN, lifts pixel samples to a normalized 3D metric point cloud (32×32 grid) and processes them using a 3-layer PointNet[2] with global pooling.
- **Decoder:** transformer decoder where learned queries iteratively cross-attend to image and geometric tokens

Evaluation protocol

Pose-and-shape datasets

BEHAVE[3]: Evaluated using a split of 38,904 training, 6,484 validation, and 18,216 test images (63,604 total samples). The dataset consists of Kinect Azure RGBD frames featuring 8 subjects performing interactions with 20 objects across 5 locations.

EgoBody[4]: Partitioned into 845,708 training, 248,506 validation, and 617,330 test samples (1,711,544 total samples). Experiments discard the HoloLens subset, retaining only the Kinect data across 125 video sequences and 36 subjects.

InterCap[5]: Utilized exclusively for testing purposes, comprising 400,249 individual frames of 10 subjects recorded in a controlled environment performing motions and object interactions with severe occlusions.

Kinematics PoC dataset

AzureTheia : controlled lab environment using two Femto Bolt cameras, featuring recordings of 5 subjects (1 male, 4 females) performing varied walking motions with simultaneous ground truth acquisition from a Theia3D[6] markerless reference system.

Results

Results on public datasets and pose-and-shape metrics

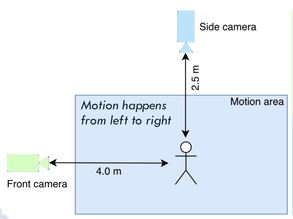
Dataset	Method	MPJPE	PA-MPJPE	MVE	MTE	MJLTC (%)
BEHAVE	CameraHMR[7]	51.9	38.2	64.3	83.6	-
	HSMR[8]	94.6	51.3	108.0	-	11.3
	Ours	**45.5**	**36.5**	**56.8**	**38.3**	**10.5**
EgoBody	CameraHMR	48.8	34.3	57.8	144.3	-
	HSMR	92.3	48.2	109.6	-	23.5
	Ours	**42.2**	**33.9**	**52.7**	**73.8**	**8.9**
InterCap	CameraHMR	51.2	**32.3**	62.5	116.2	-
	HSMR	82.5	48.9	95.9	-	18.1
	Ours	**45.7**	32.8	**57.4**	**55.0**	**16.9**

Compute cost compared to other monocular RGB methods

Model	GFLOPS	Number of parameters (M)
CameraHMR	245.17	670.71
HSMR	245.17	670.46
Ours	**10.14**	**26.27**

Kinematics accuracy compared to the Azure Kinect Body Tracking toolkit

Top view of the acquisition setup



	Kinect ABT			Ours		
Flexion angle	MAE	RMSE	$ICC_{3,1}$	MAE	RMSE	$ICC_{3,1}$
Left hip	27.08	32.96	0.75 [0.73, 0.77]	10.17	11.54	0.91 [0.91, 0.92]
Right hip	29.35	35.15	0.76 [0.74, 0.77]	10.16	11.43	0.92 [0.92, 0.93]
Left knee	8.17	12.36	0.70 [0.68, 0.72]	6.45	8.38	0.84 [0.83, 0.85]
Right knee	8.44	13.14	0.68 [0.66, 0.70]	6.18	8.11	0.86 [0.85, 0.87]
Left hip	28.64	34.27	0.75 [0.72, 0.77]	9.31	10.12	0.94 [0.93, 0.95]
Right hip	30.66	36.08	0.73 [0.71, 0.76]	9.84	10.64	0.94 [0.93, 0.94]
Left knee	7.91	11.58	0.77 [0.74, 0.79]	4.64	5.86	0.94 [0.94, 0.95]
Right knee	7.57	10.83	0.79 [0.77, 0.81]	4.75	6.18	0.92 [0.91, 0.93]

Discussion

Our work would benefit from integrating temporal modeling to ensure motion continuity. However, high inference speed makes it an efficient initialization tool, ideally suited to serve as a strong kinematic prior for optimization-based refinement pipelines, or with the integration of motion models

[1] A. Dosovitskiy et al., 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', arXiv: arXiv:2010.11929
[2] R. Q. Charles, et al., 'PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation', CVPR 2017, doi: 10.1109/CVPR.2017.16.
[3] B. L. Bhatnagar et al., 'BEHAVE: Dataset and Method for Tracking Human Object Interactions', CVPR 2022, doi: 10.1109/CVPR52688.2022.01547.
[4] S. Zhang et al., 'EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices', ECCV 2022, doi: 10.1007/978-3-031-20068-7_11.
[5] Y. Huang et al., 'InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction', in Pattern Recognition, doi: 10.1007/978-3-031-16788-1_18.
[6] Theia Markerless. Available at https://www.theiamarkerless.ca/
[7] P. Patel et al., 'CameraHMR: Aligning People with Perspective', doi: 10.48550/arXiv.2411.08128.
[8] Y. Xia et al., 'Reconstructing Humans with a Biomechanically Accurate Skeleton', doi: 10.48550/arXiv.2503.21751.