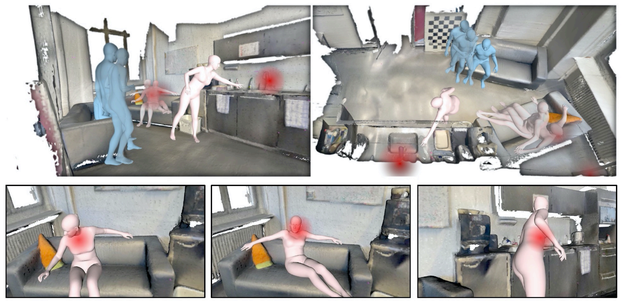
EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices

Siwei Zhang[1], Qianli Ma[1], Yan Zhang[1], Zhiyin Qian[1], Taein Kwon[1], Marc Pollefeys[1,2], Federica Bogo[2*], Siyu Tang[1]
1 ETH Zürich 2 Microsoft
* Now at Meta Reality Labs Research

Overview

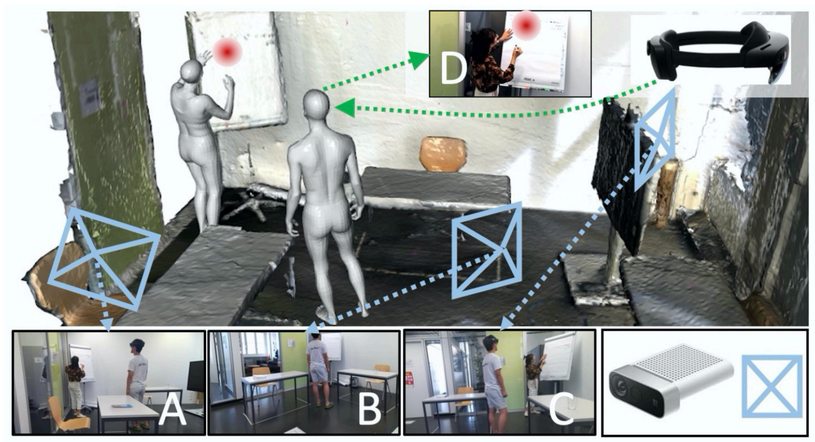


The first step towards **egocentric human behavior understanding**:
- 3D body pose, shape (3DHPS) and motion estimation of the social interaction partner ("*interactee*") from egocentric views

Limitations of existing datasets:
- most focus on the third-person view
- limited social interaction diversity / 3D scene information / only skeleton information

EgoBody is the **first large-scale dataset** of:
- accurate 3D human body shape, pose and motion of humans interacting in 3D scenes with multi-modal streams from third-person and egocentric views

Capture Setup



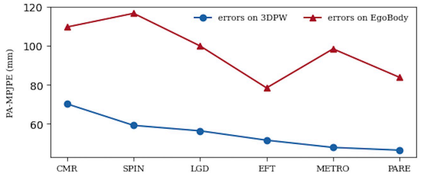
3-5 Azure Kinects (A, B, C): multi-view third-person view RGBD videos
1 HoloLens2 headset (D): egocentric RGB, depth, eye gaze, hand/head tracking

Category	Interaction Scenarios
Cooperation	Guess by Action game, catching and tossing, searching for items, etc.
Social exchange	Teaching to dance/workout, giving a presentation, etc.
Conflict	Arguing about a specific topic
Conformity	One subject instructs the other to perform a task
Others	Haggling, negotiation, promotion, self-introduction, casual chat, etc.
Action Types	Sitting, standing, walking, dancing, exercising, bending, lying, grasping, squatting, drinking, passing objects, catching, throwing, etc.

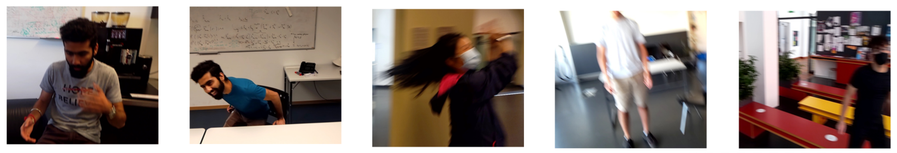
EgoBody Dataset

- **125** sequences
- **36** subjects
- **15** 3D indoor scenes
- **220k multi-view third-person view RGBD frames**
- **199k egocentric view RGB frames**
- **Eye gaze, hand/head tracking**
- **3D human shape and motion annotations** for both interacting subjects

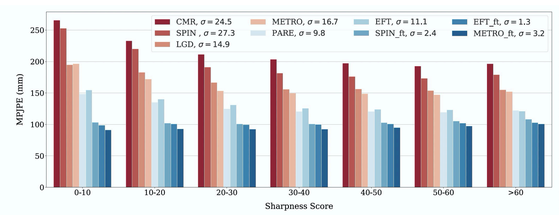
Egocentric 3DHPS Estimation Benchmark



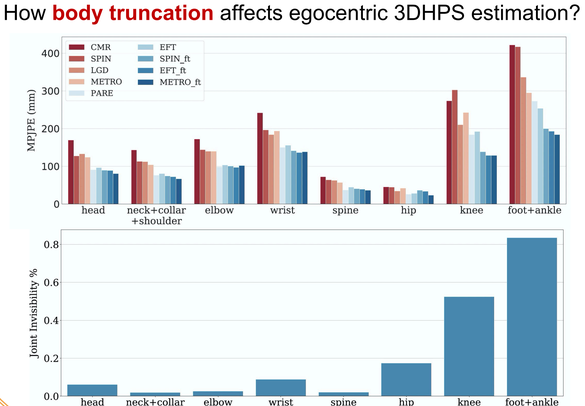
Task: 3D human pose and shape estimation of the interactee from the egocentric view
Challenges: **motion blur, body truncations**, etc.



How **motion blur** affects egocentric 3DHPS estimation?



How **body truncation** affects egocentric 3DHPS estimation?



Baseline Evaluation / Improvements

Method	MPJPE ↓	PA-MPJPE ↓	V2V ↓	PA-V2V ↓
CMR [49]	200.7	109.6	218.7	136.8
SPIN [48]	182.8	116.6	187.3	123.7
LGD [81]	158.0	99.9	168.3	106.0
METRO [57]	153.1	98.4	164.6	106.4
PARE [46]	123.0	83.8	131.5	89.7
EFT [36]	123.9	78.4	134.9	86.0
SPIN-ft (Ours)	106.5	67.1	120.9	78.3
METRO-ft (Ours)	**98.5**	70.0	**110.5**	76.8
EFT-ft (Ours)	102.1	**64.8**	116.1	**74.8**

test on **EgoBody** test set

Cross-dataset Evaluation

Method	PA-MPJPE ↓	
SPIN	152.8	
SPIN-ft (Ours)	87.9	-42%
METRO	117.7	
METRO (Ours)	88.2	-25%
EFT	95.8	
EFT-ft (Ours)	85.6	-11%

test on **You2Me** dataset

- Fine-tuning on **EgoBody** training set (-ft, Ours) improves existing methods' performance and robustness on both **EgoBody** test set and **You2Me** dataset.