InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction

[1] Yinghao Huang [1] Omid Tehari [1]Michael J. Black [2] Dimitrios T[1]zionas

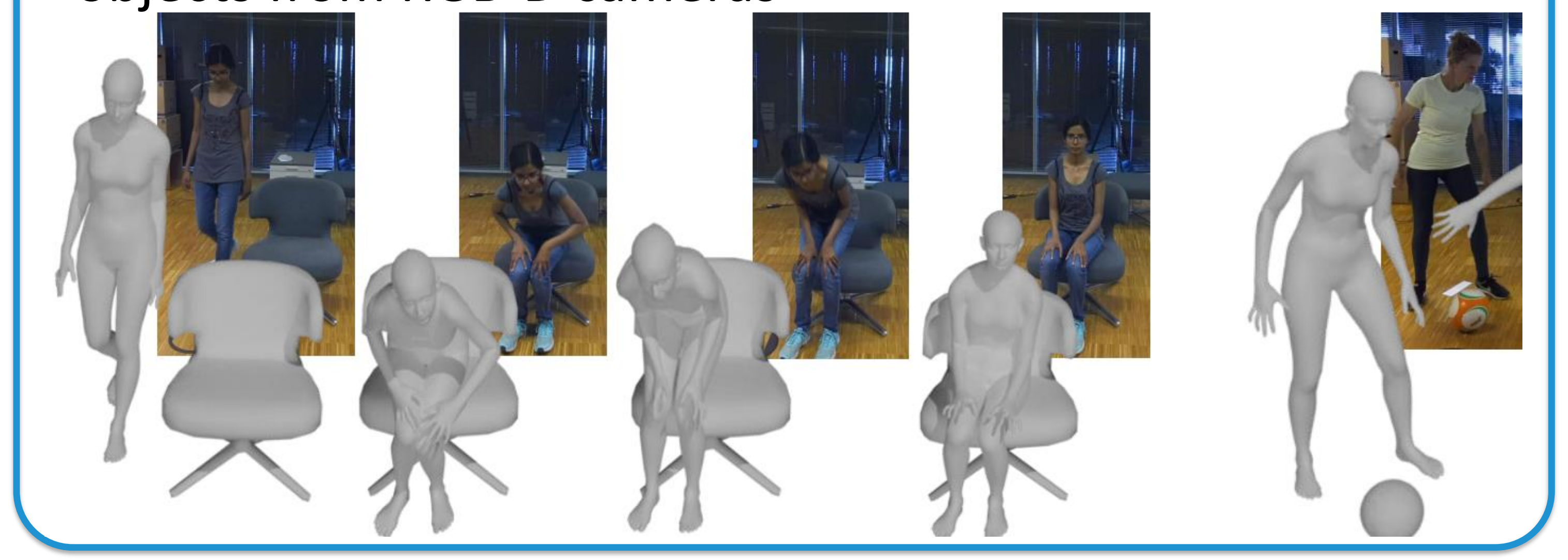
[1] Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]University of Amsterdam

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

UNIVERSITY OF AMSTERDAM

Goal

Markerless reconstruction of 3D humans and interacting objects from RGB-D cameras

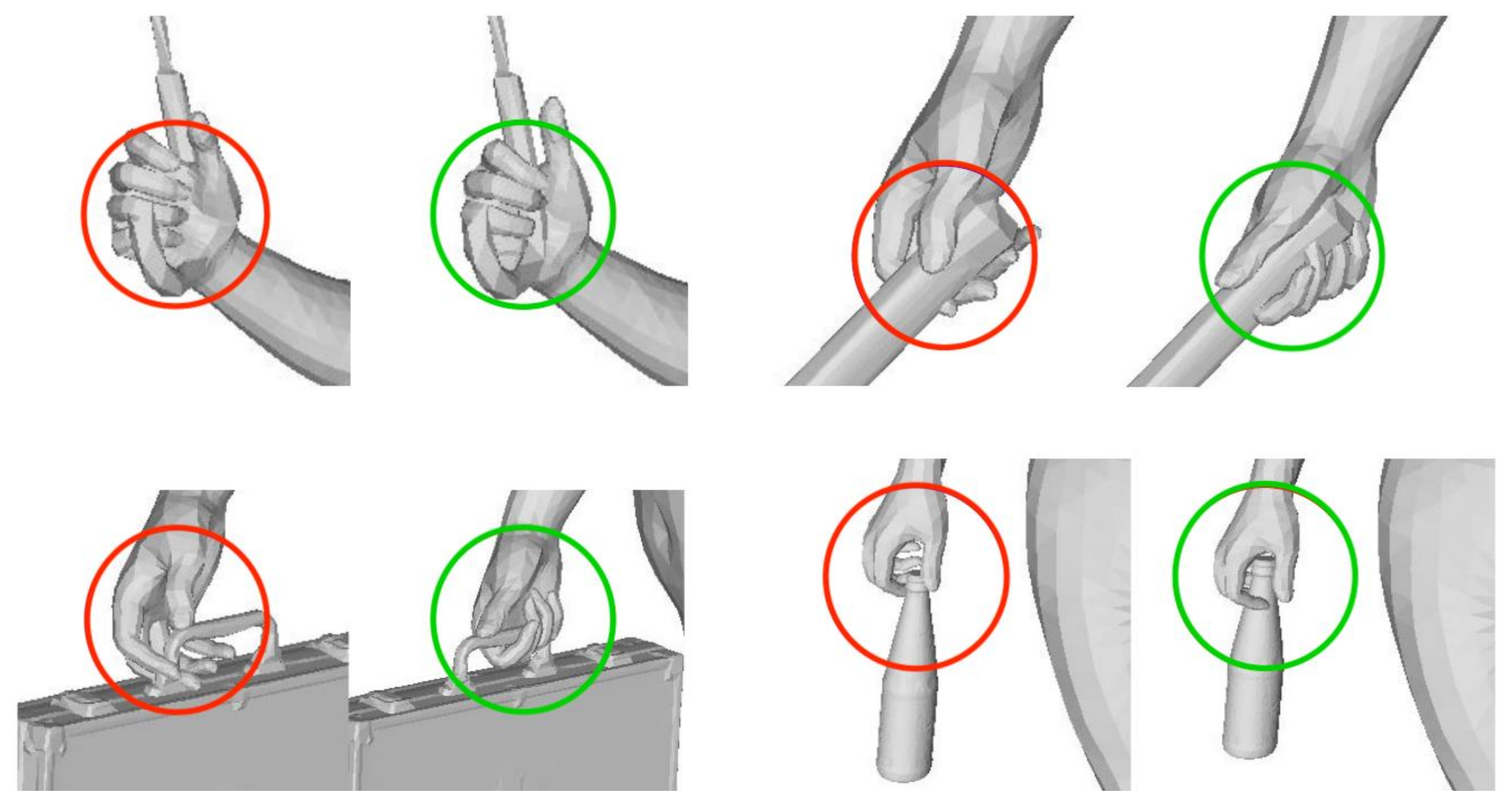


Problem

1. Humans are intrinsically complex articulated creatures, estimating accurate human pose is challenging even with special devices like dense markers
2. Occlusion between the subject and the object during interaction is heavy and common, making tracking hard
3. Previous marker/IMUs-based solutions cannot easily be applied in daily scenarios, thus are not so practical

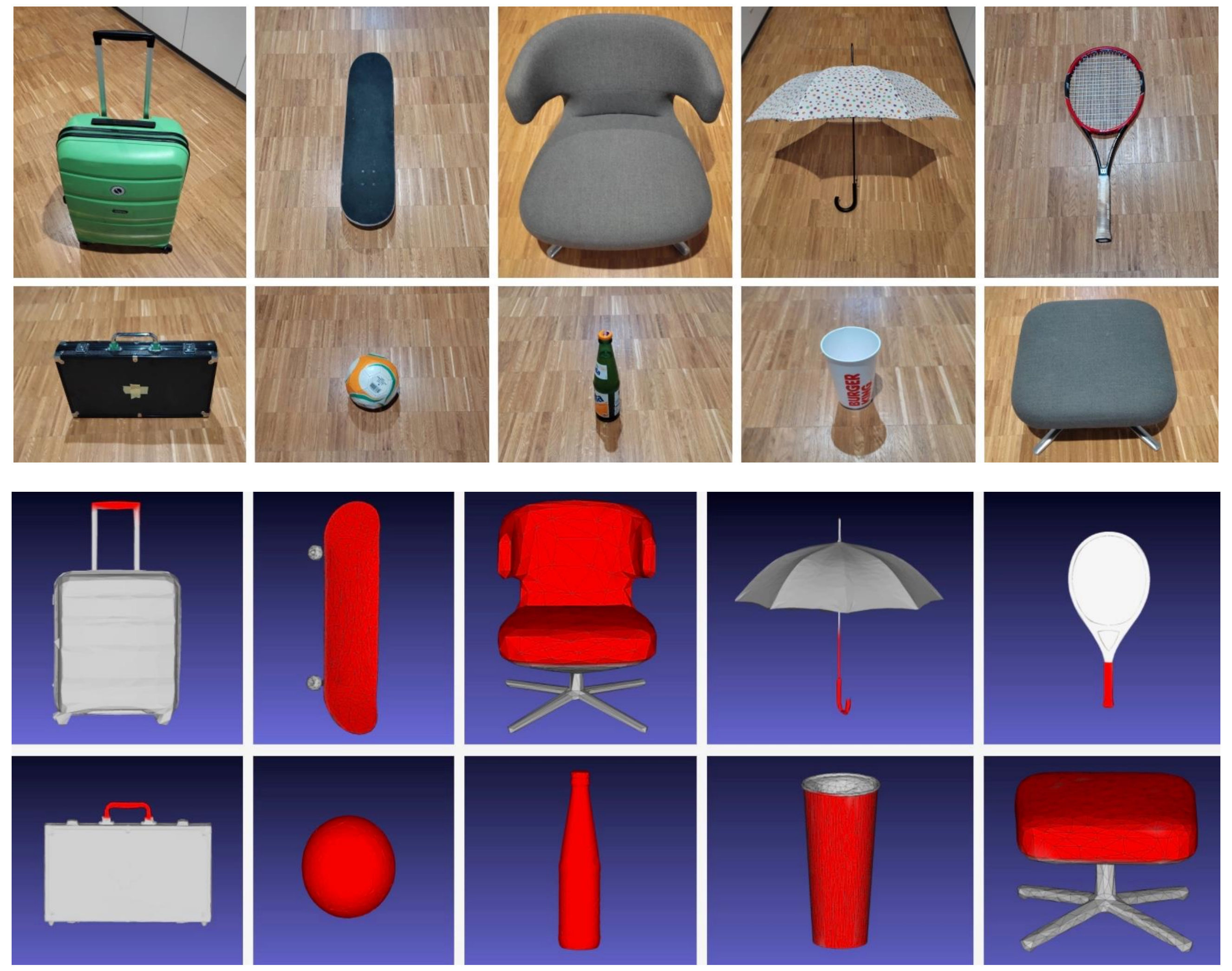
Key Observation

Body-object interactions provide strong constraints about how the human and the object move

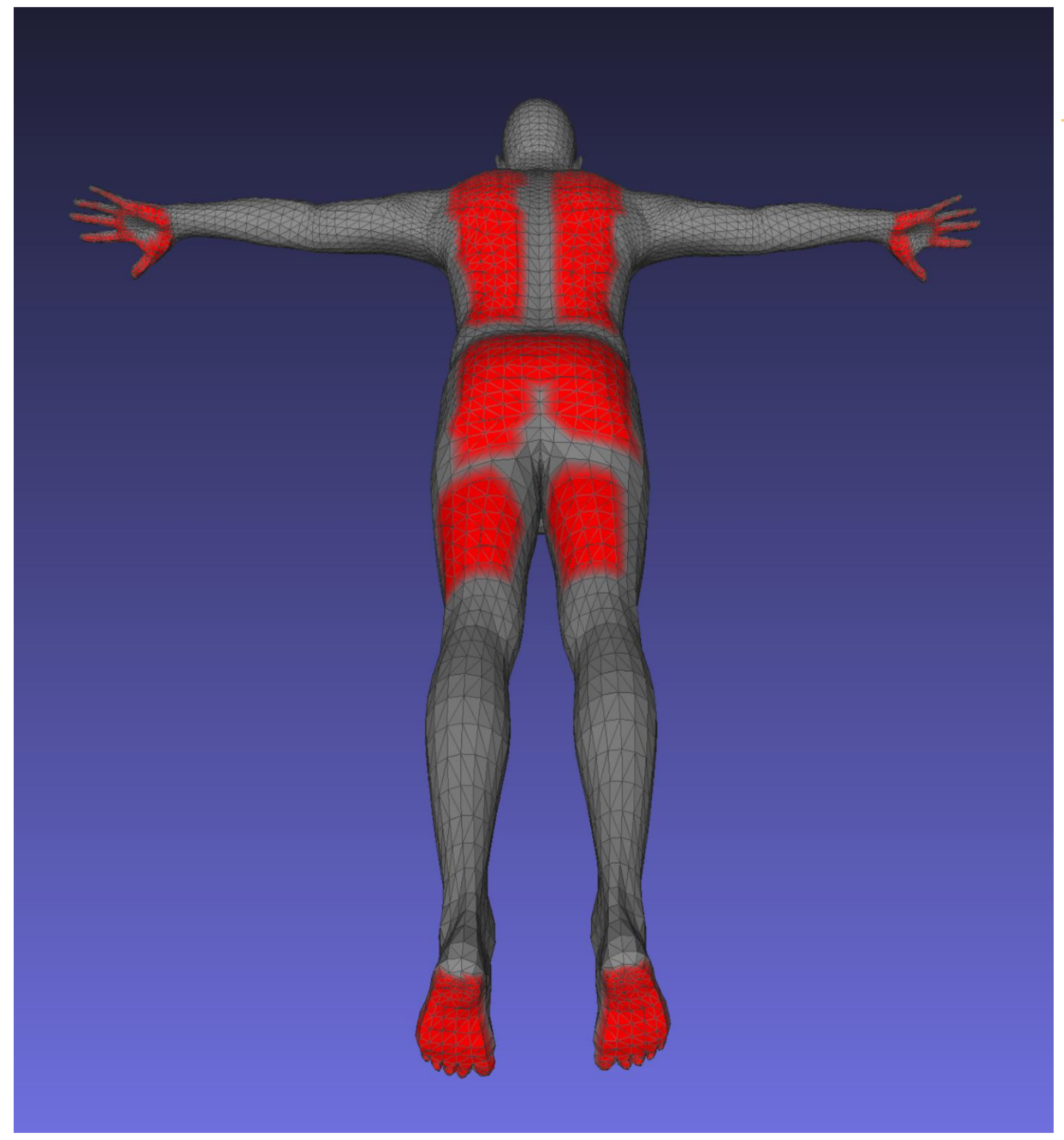


InterCap: Optimization Approach

Objects & Annotations



Body Annotation



Objective Function

$$
\begin{aligned}
E = \; & \frac{1}{T}\sum_{\text{frame } t}\Big[E_{O}(\Xi_t;\mathcal{S}_t,\mathcal{D}_t)+E_{B}(\beta^*,\Theta_t,\Psi_t,\Gamma_t;\mathcal{J}_{est})\Big]+ \\
& \frac{1}{T}\sum_{\text{frame } t}\Big[E_{\mathcal{P}}(\Theta_t,\beta^*,\Gamma_t)+E_{\mathcal{C}}(\beta^*,\Theta_t,\Psi_t,\Gamma_t,\Xi_t,M)\Big]+ \\
& \frac{\lambda_{\mathcal{G}}}{T}\sum_{\text{frame } t}\Big[E_{\mathcal{G}}(\beta^*,\Theta_t,\Psi_t,\Gamma_t)+E_{\mathcal{G}'}(\Xi_t,M)\Big]+ \\
& \frac{\lambda_{\mathcal{Q}}}{T}\sum_{\text{frame } t}\Big[Q_t * E_{\mathcal{C}}(\beta^*,\Theta_t,\Psi_t,M',\Xi_t)\Big]+ \\
& \quad \lambda_{\mathcal{S}}E_{\mathcal{S}}(\Theta,\Psi,\Gamma,A;\beta^*,T)+ \\
& \quad \lambda_{\mathcal{A}}E_{\mathcal{A}}(\Xi,T,M),
\end{aligned}
$$

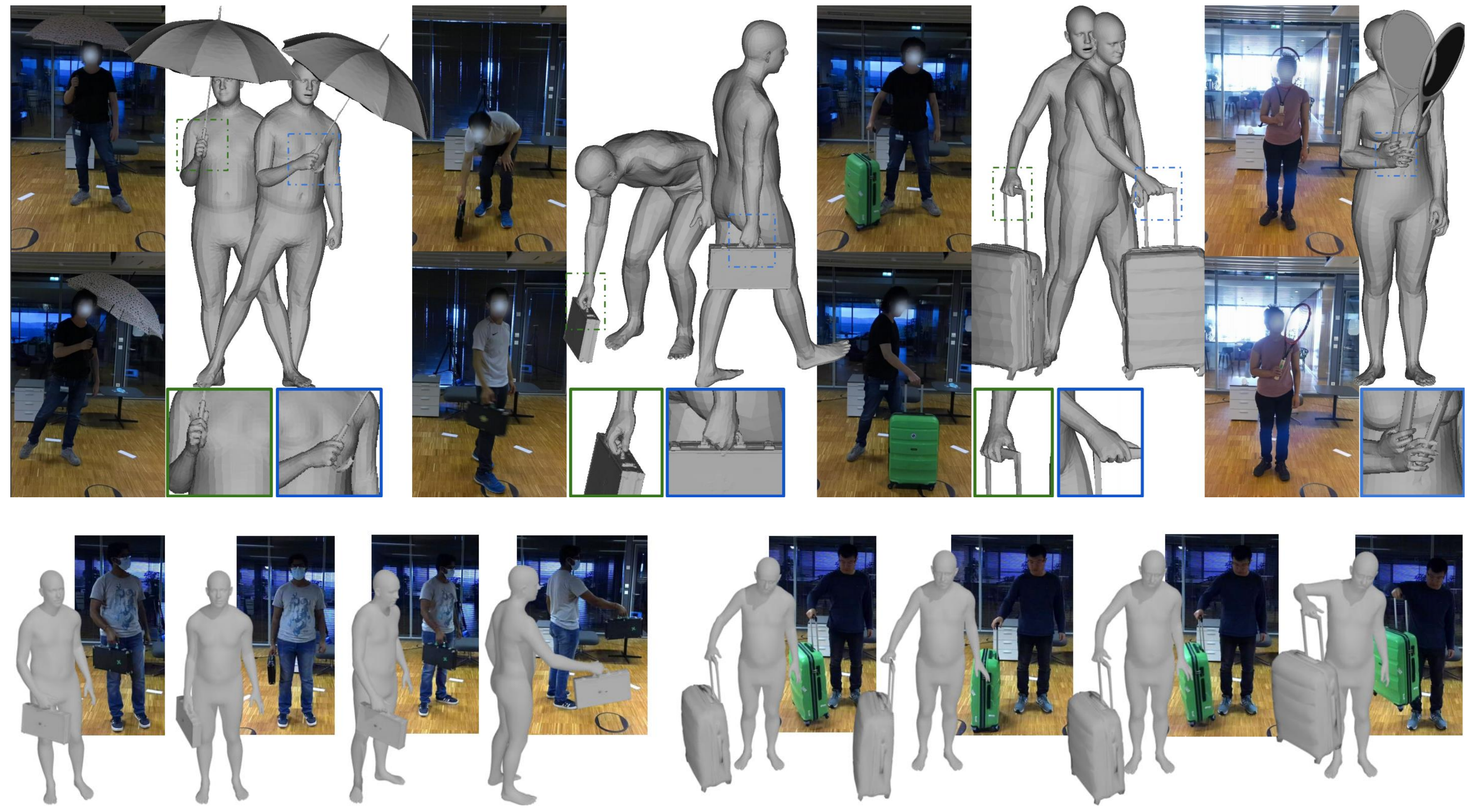
Contour & Landmarks

Prior loss & Contact

Above-ground term

Smoothness constraint

Results



Comparisons

Name	# of Seq.	Natural Appear.	Moving Objects	Accurate Motion	With Image	Artic. Hands
HumanEva [52]	56	✓	✗	✓	✓	✗
Human3.6M [23]	165	✓	✗	✓	✓	✗
AMASS [35]	11265	✓	✗	✓	✗	✗
GRAB [54]	1334	✓	✓	✓	✗	✓
3DPW [36]	60	✓	✗	✓	✓	✗
GTA-IM [5]	119	✗	✗	✓	✓	✗
SAIL-VOS [20]	201	✗	✗	✗	✗	✗
PiGraphs [51]	63	✓	✗	✓	✓	✗
PROX [15]	20	✓	✗	✗	✓	✗
RICH [21]	142	✓	✗	✓	✓	✗
BEHAVE [3]	321	✓	✓	✓	✓	✗
InterCap (ours)	223	✓	✓	✓	✓	✓

References

[1] SMPL: a skinned multi-person linear model, Loper et al., SIGGRAPH Asia 2015

[2] Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image, Bogo et al., ECCV 2016

[3] End-to-end recovery of human shape and pose, Kanazawa et al., CVPR 2018

[4] Expressive body capture: 3d hands, face, and body from a single image, Pavlakos, CVPR 2019

[5] BEHAVE: Dataset and Method for Tracking Human Object Interactions, Bhatnagar et al., CVPR 2022

[6] Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion, Sigal et al., IJCV 2010

[7] Resolving 3D human pose ambiguities with 3D scene constraints, Hassan et al., CVPR 2019

[8] Learning motion priors for 4d human body capture in 3d scenes, Zhang et al, ICCV 2021

[9] Human-aware object placement for visual environment reconstruction, Yi et al., CVPR 2022

[10] GRAB: A dataset of whole-body human grasping of objects, Taheri et al., ECCV 2020

Code & Data: intercap.is.tue.mpg.de