Virtual Humans – Winter 23/24

Lecture 11_1 – Human Behavior Capture

Prof. Dr.-Ing. Gerard Pons-Moll

University of Tübingen / MPI-Informatics

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

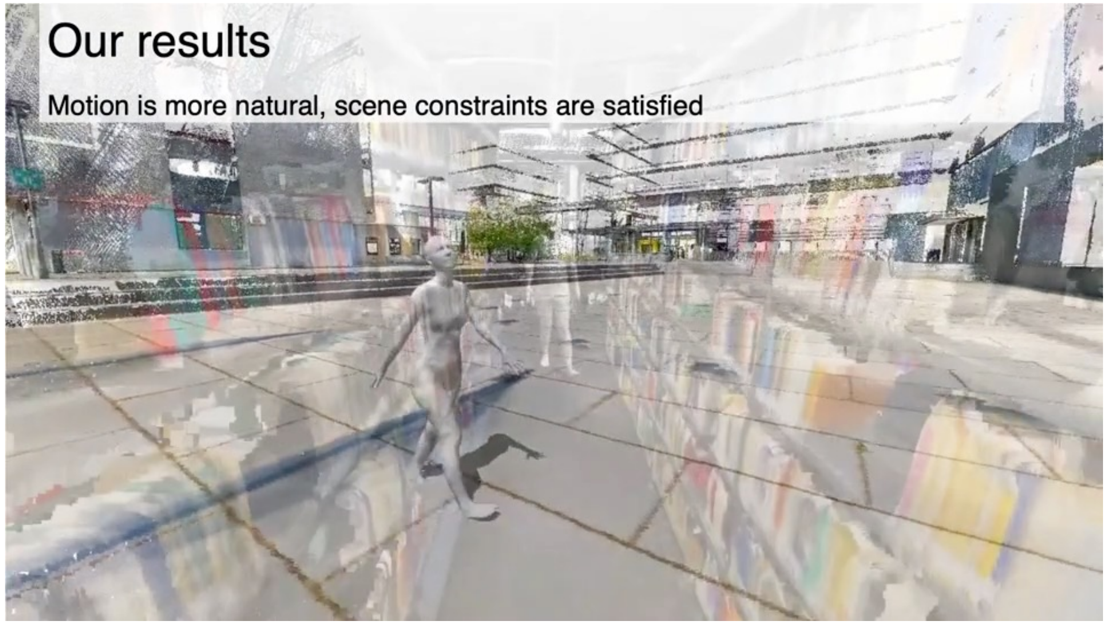
In this lecture…

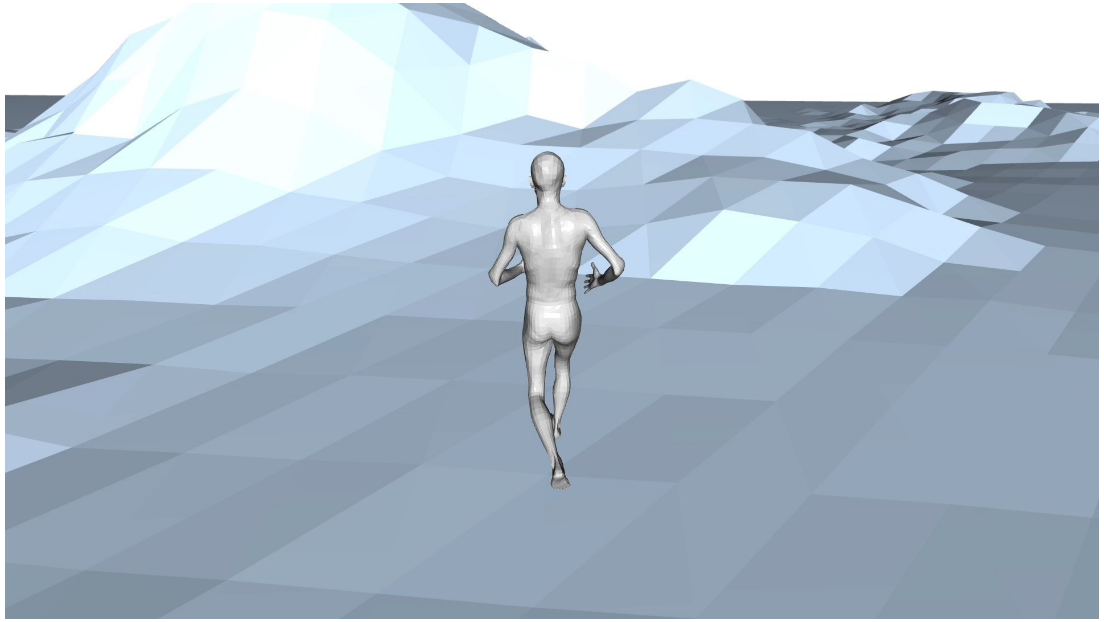
- Capturing humans in static scenes.

- Capturing dynamic human object interactions.

- Large scale, long term humans in scene.

Goal: **Awaken** Virtual Humans



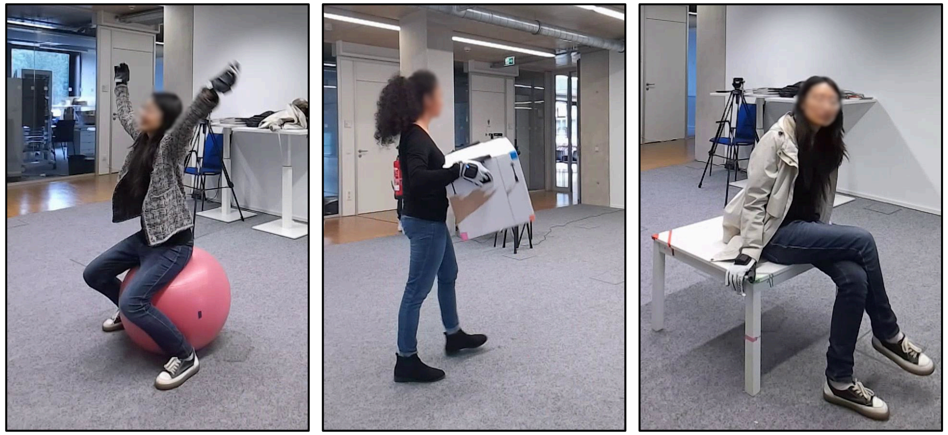


Perceive: We should be able to reconstruct **real** 3D humans jointly with the objects and the scene they interact with.

Generation: **Virtual** humans should be able to move and interact with objects and scenes like real humans

Why model human-object interactions?

We interact with our surroundings constantly.



We understand the world by interacting with it.

Applications of human-object interactions.

Robots can learn from humans

Smith et al. RSS'20

Humans can learn in virtual environments

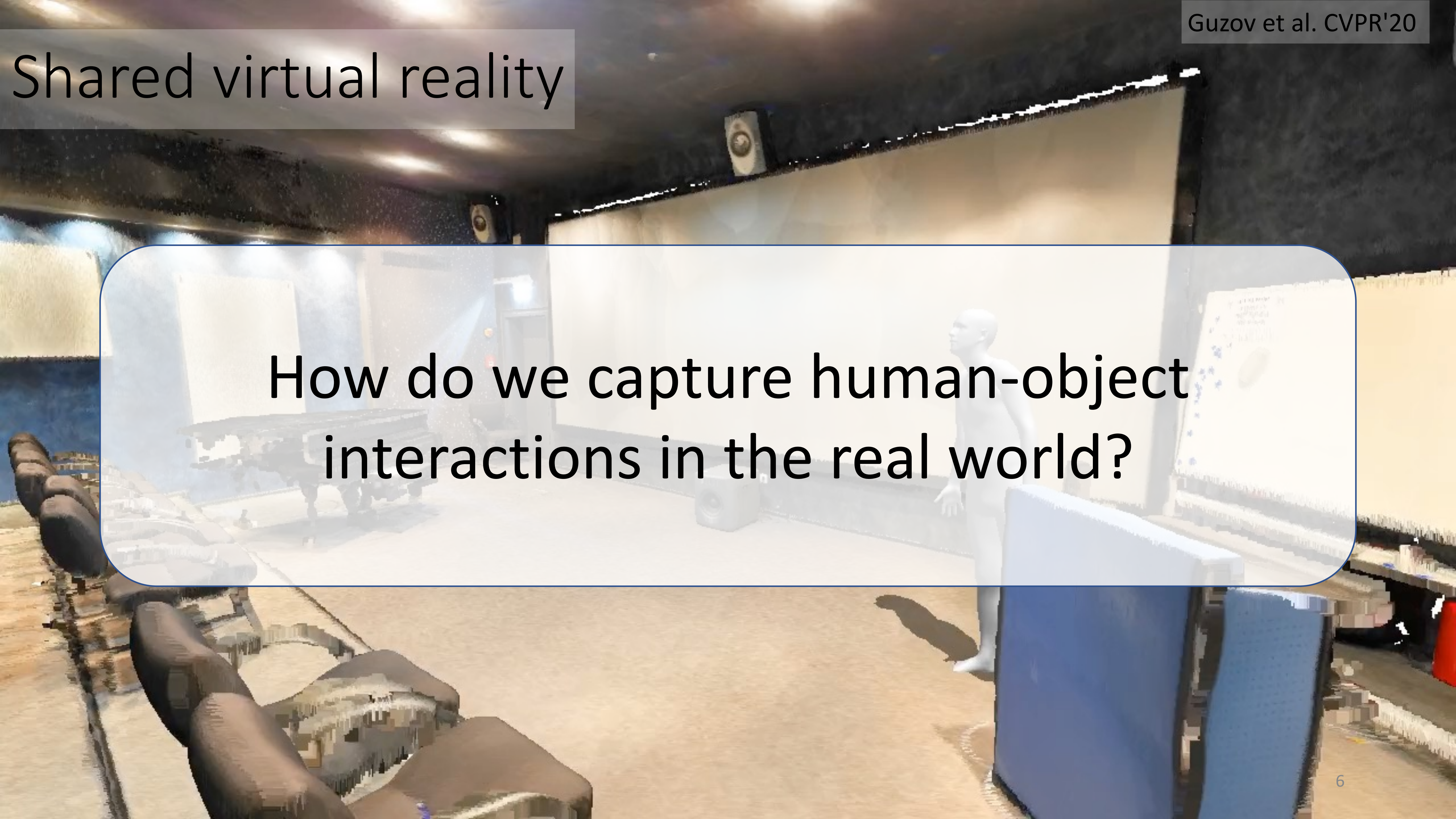
F.A.S.T. VR Qualcomm'19

Virtual assistants

Guzov et al. CVPR'20

Shared virtual reality

Guzov et al. CVPR'20



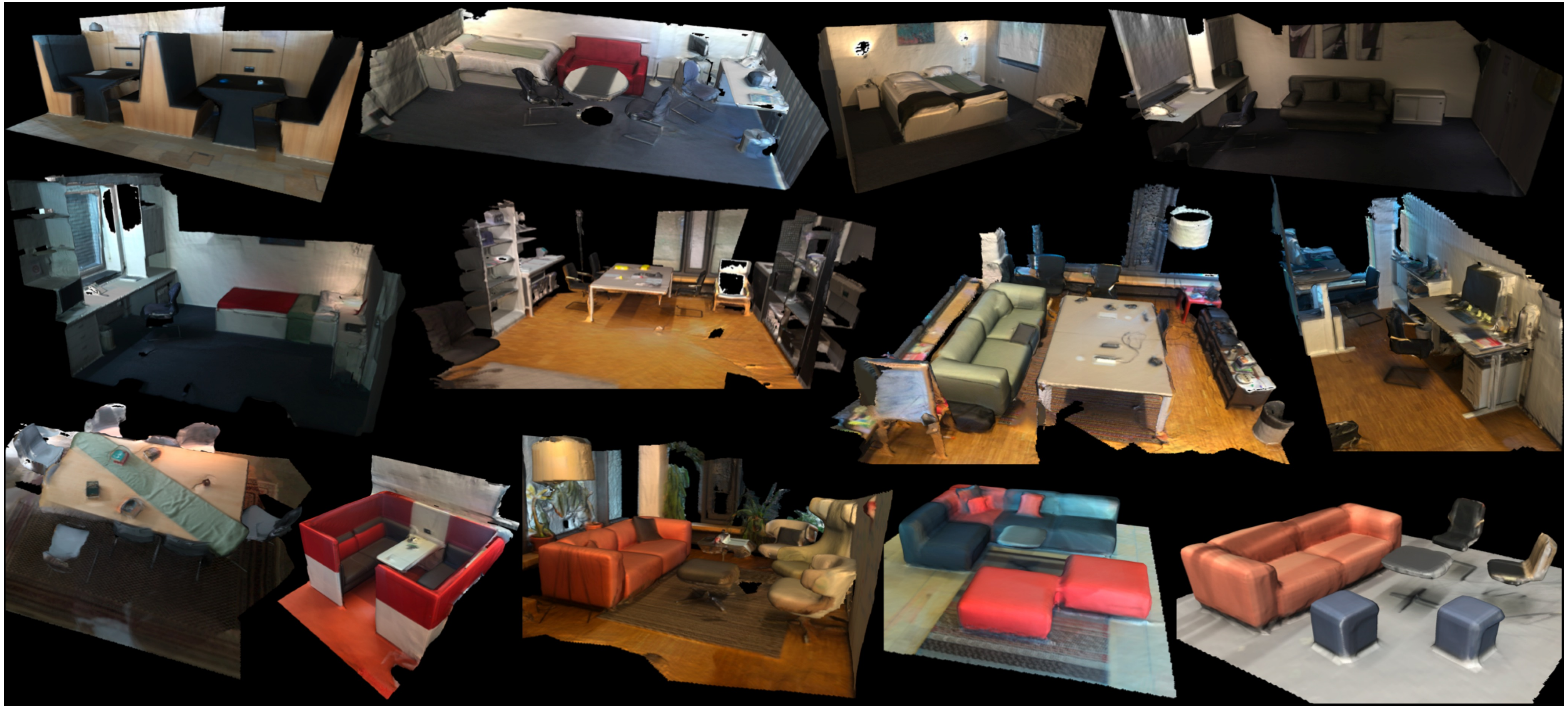
How do we capture human-object interactions in the real world?

Capturing humans in **static** **scenes**

[PROX, ICCV'19]

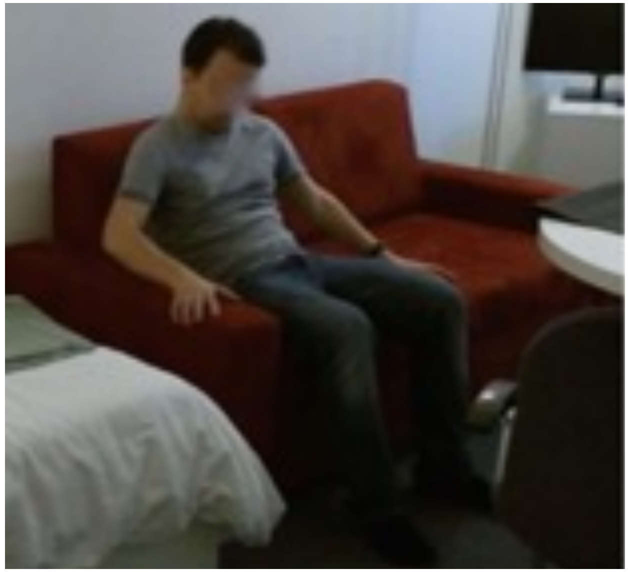
Pre-reconstruct 3D scene with RGBD cameras.

Record “empty (w/o human)” and “static” scenes.



Fit SMPL to the scene and image jointly.

Capture images of the person in the scene.



Retrieve corresponding part of 3D scene.

Fit SMPL to scene and image.

Challenges.

- Fitting SMPL to image does not give correct depth.
- There are interpenetrations.

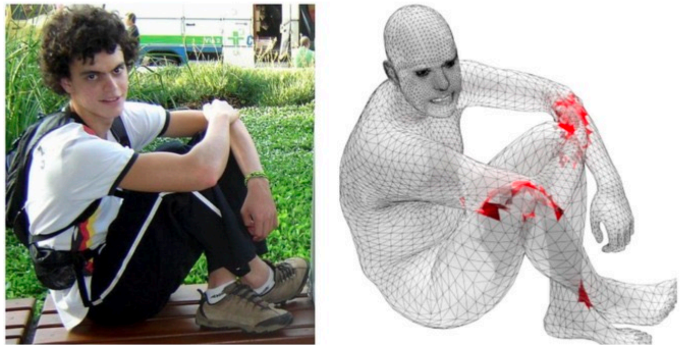
Input Image

SMPL aligns in camera view

Incorrect in 3D

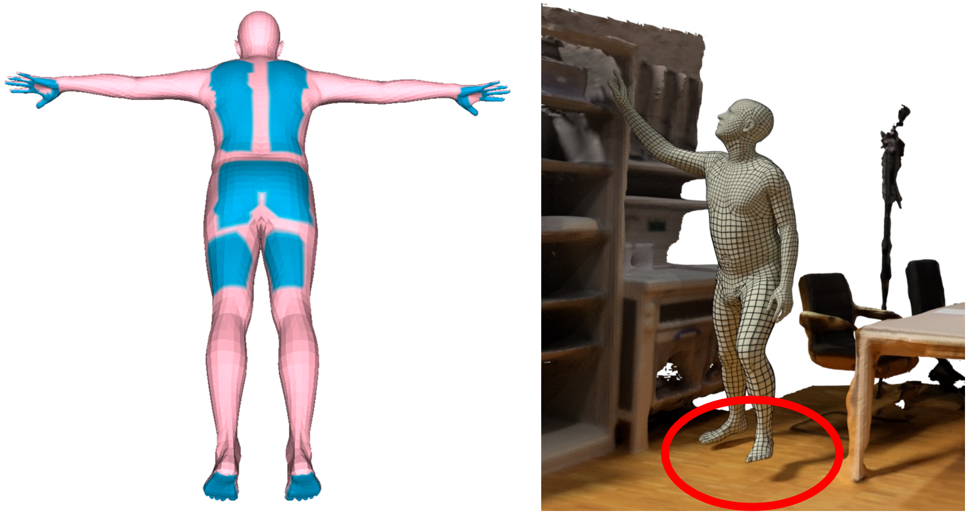
[PROX, ICCV'19]

Penalise penetrations.

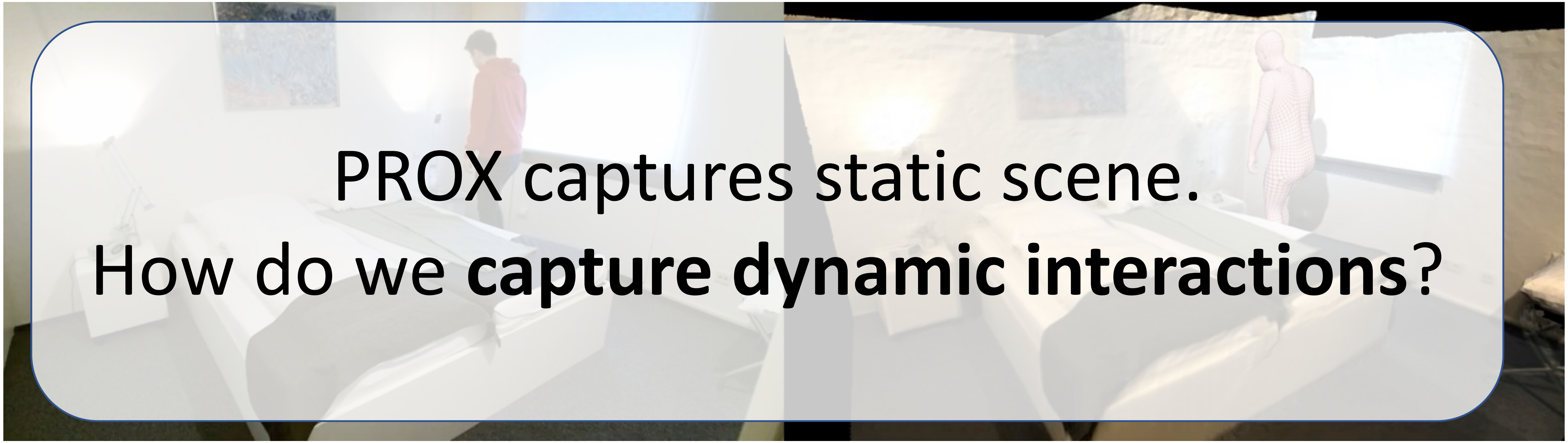


Use contacts.

- Manually annotate likely contact vertices.
- Encourage proximity if contact vertices close to scene.

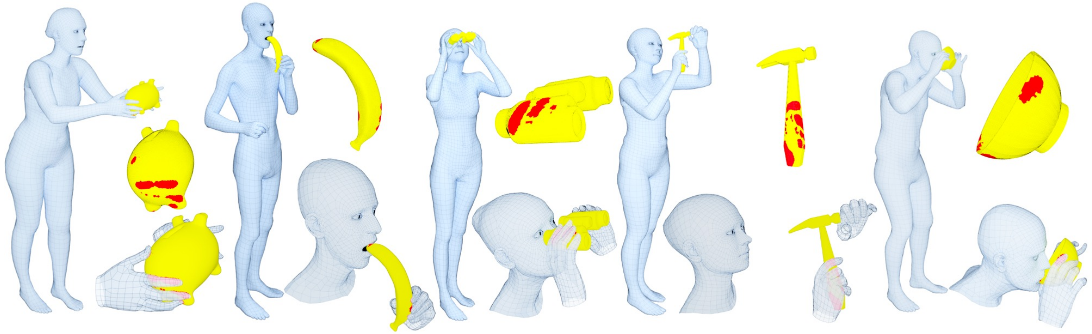


PROX dataset.



GRAB, Taheri et. al., ECCV'20
HPS, Guzov et. al., CVPR'21

Traditional capture setups are not suitable…



Marker based capture methods:

- ✓ High quality data
- – Expensive
- – Limited in recording volume
- – Not easy to scale recording locations
- – Markers get occluded during human-object interactions.

IMU based capture methods:

- ✓ Easy to scale
- ✓ No restriction on recording volume
- – Prone to sensor drift. Quality of data is poor.

BEHAVE: Dataset and Method for Tracking Human Object Interactions, CVPR'22

Bharat Lal Bhatnagar[1,2], Xianghui Xie[2], Ilya Petrov[1], Cristian Sminchisescu[3], Christian Theobalt[2], Gerard Pons-Moll[1,2]

[1]University of Tübingen, Germany

[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

[3]Google Research

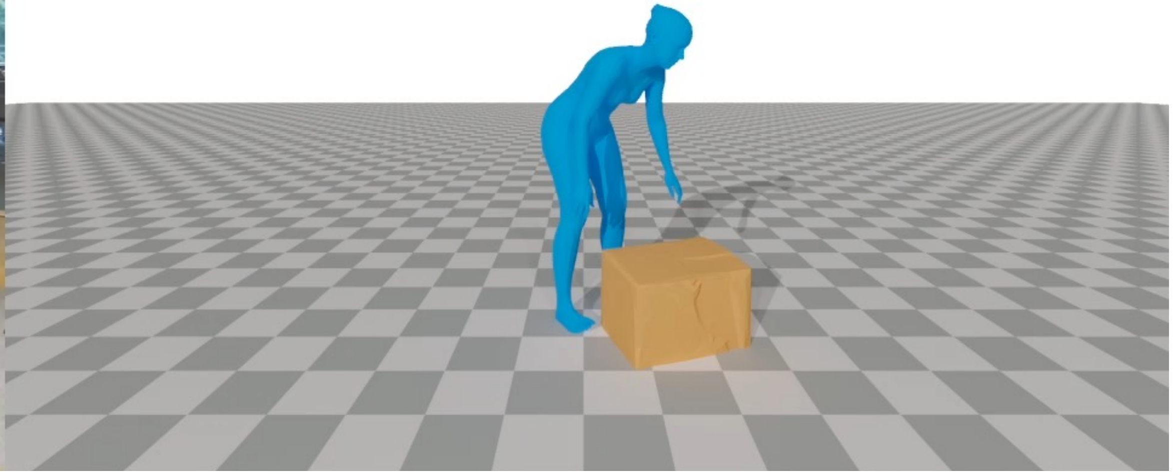
BEHAVE capture setup.

- Captured using **4 calibrated Kinects**
- 8 subjects (5 male, 3 female)
- 5 locations
- 20 common daily-use objects and interactions
  - Object: boxes, chairs, tables, backpack, monitor, exercise ball,…
  - Interactions: sit, lift, drag, pull, push…

BEHAVE dataset.

RGB sequence

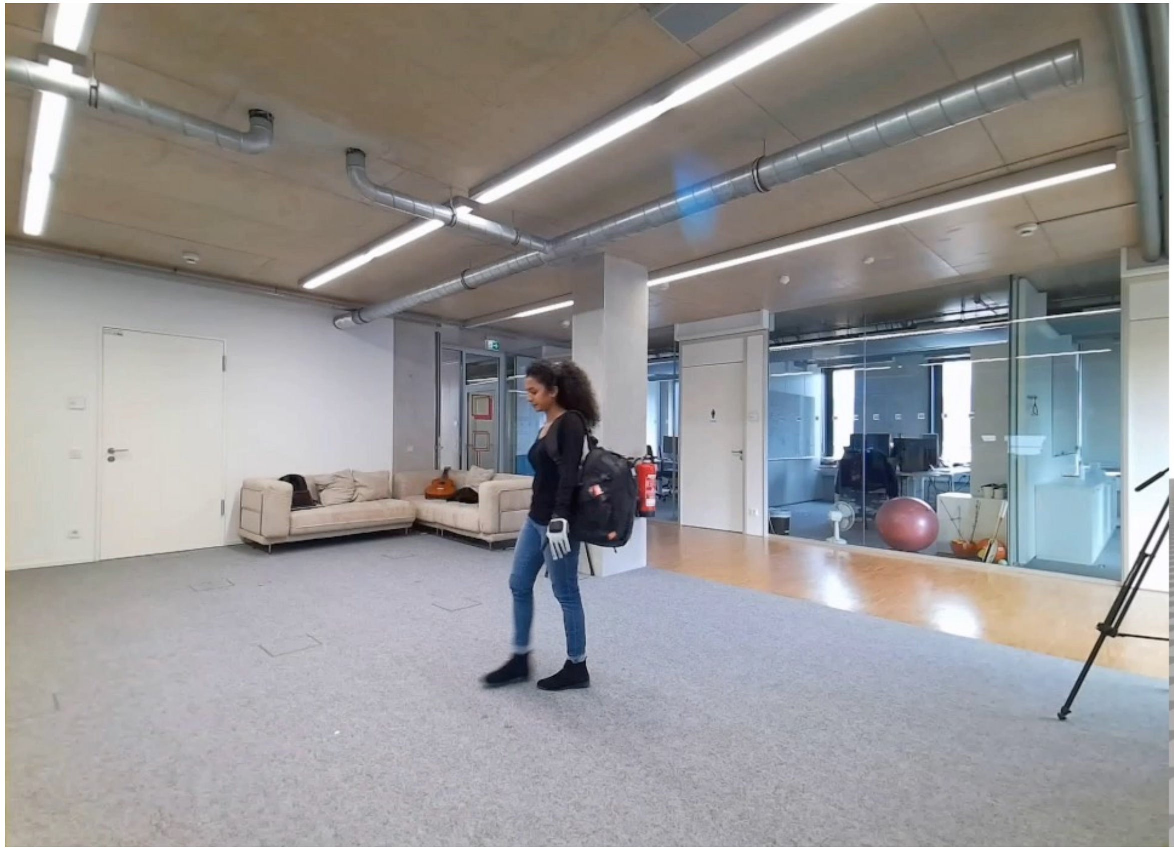
SMPL, object and contacts



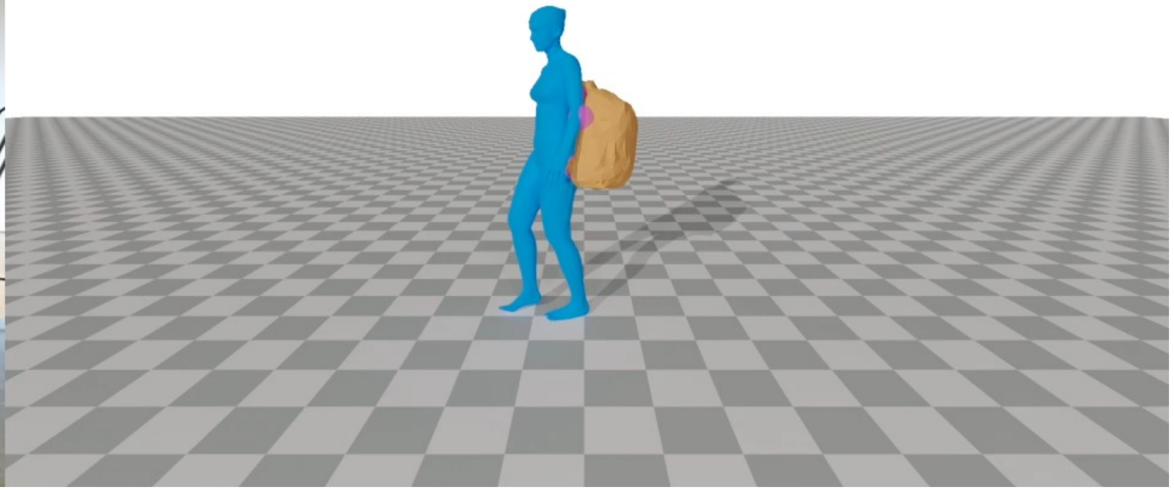
BEHAVE, CVPR’22

BEHAVE dataset.

RGB sequence

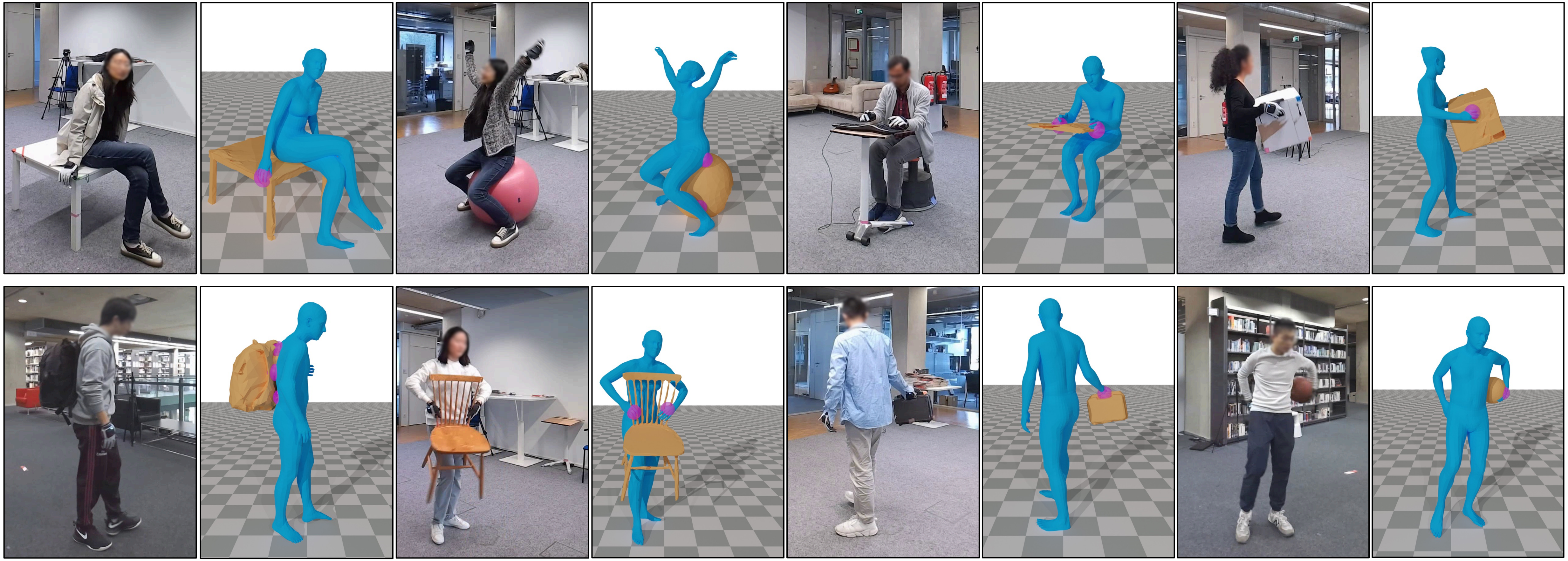


SMPL, object and contacts



BEHAVE, CVPR’22

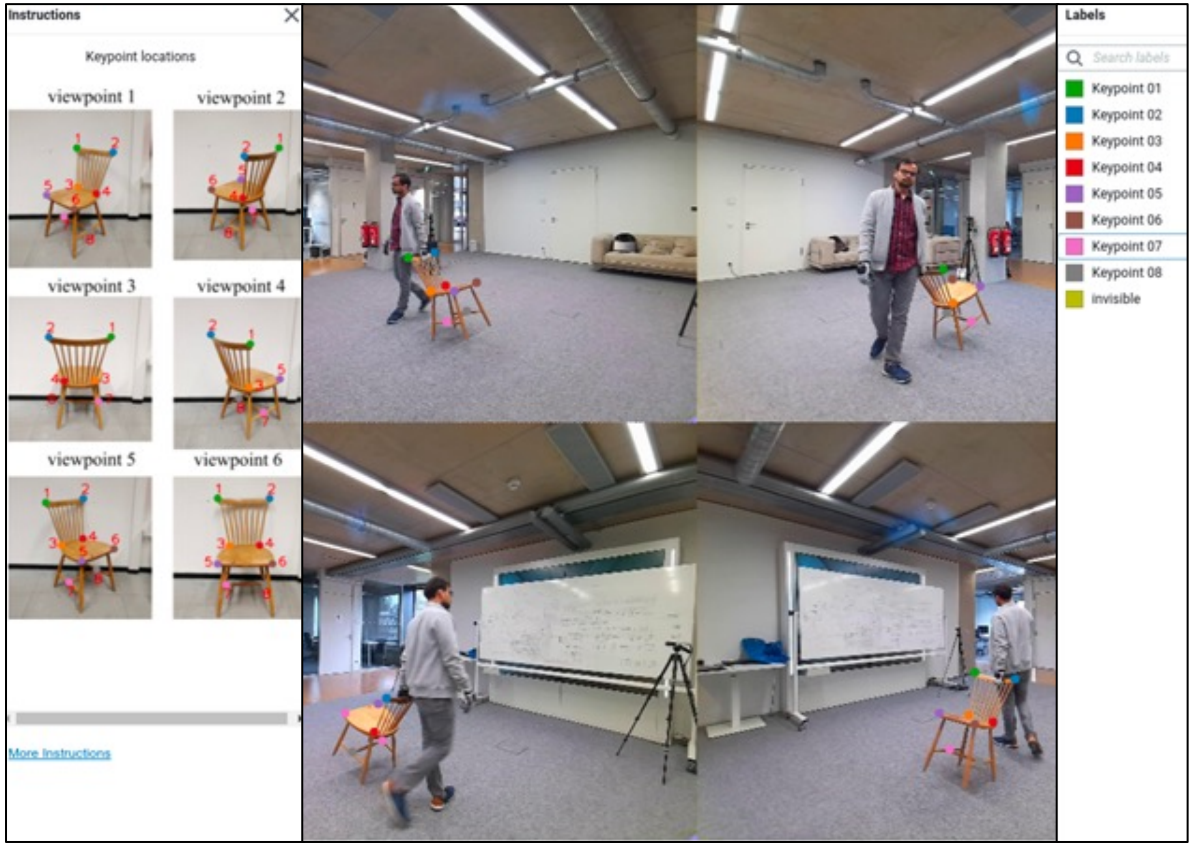
BEHAVE dataset



[1] https://github.com/facebookresearch/detectron2
[2] https://www.mturk.com/
[3] Cao et. al., PAMI'19

BEHAVE annotations.

- Human segmentation:
  - DetectronV2[1] + Manual correction with AMT[2]
- Human fits (SMPL)
  - OpenPose[3] keypoints + optimization
- Object fits and segmentation:
  - Keypoint annotation from AMT + optimization
  - Project fit object to image for segmentation

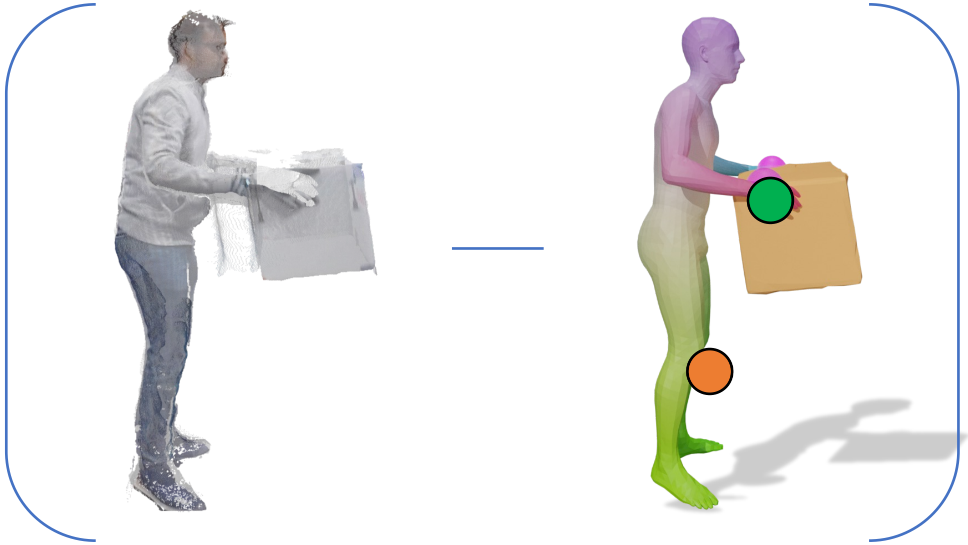


BEHAVE, CVPR’22

Fitting Models to Data (the classical way)

$$\arg\min_{\mathbf{x}} \operatorname{dist}(f(I), f_m(M(\mathbf{x})))$$

Model Parameters



Data

World Model

With standard features (edges, keypoints, sillhouettes) it is prone to local minima due to occlusions, matching ambiguities, missing data.

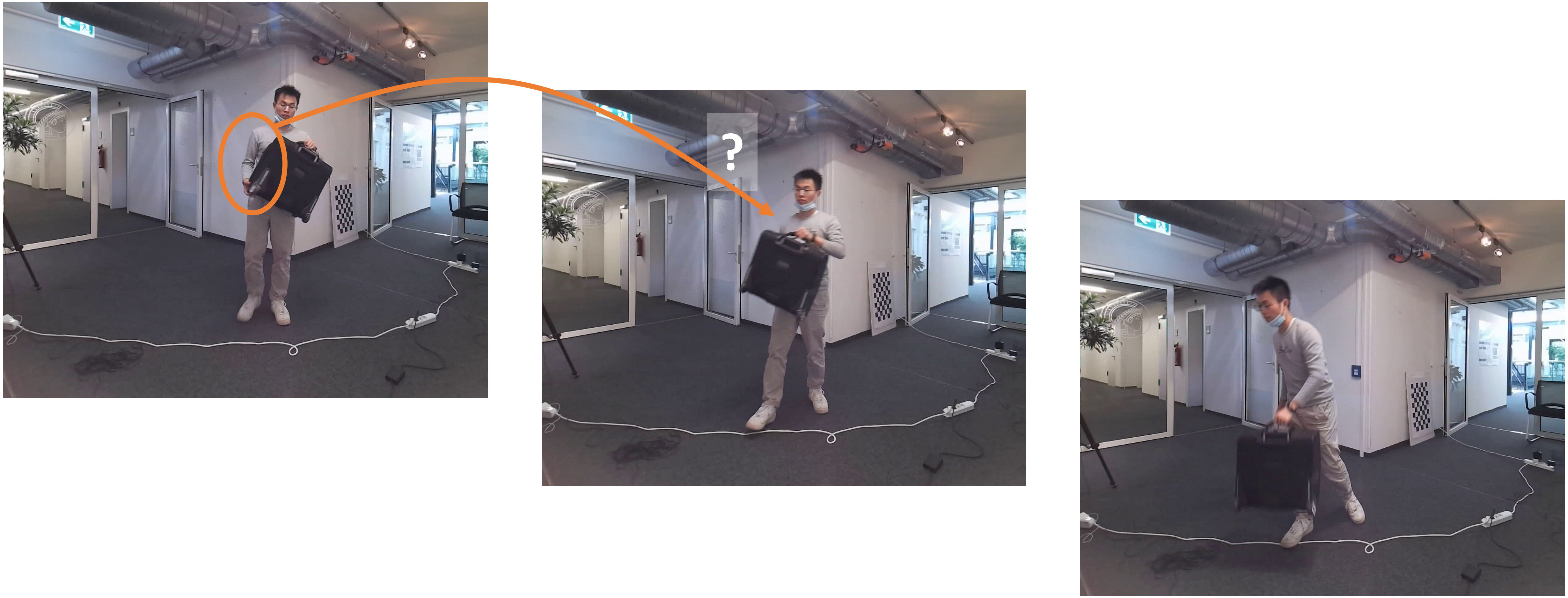
Prone to local minima due to occlusions, matching ambiguities, missing data.

Kinect data is noisy and incomplete.

Contacts are fine-grained and often barely visible.

Prone to local minima due to occlusions, matching ambiguities, missing data.

Prone to local minima due to occlusions, matching ambiguities, missing data.

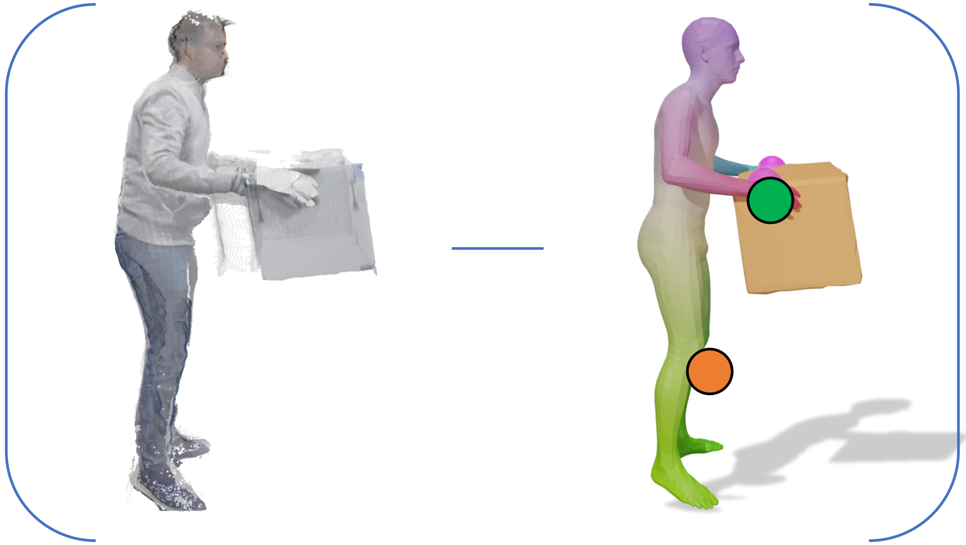


Prone to local minima due to occlusions, matching ambiguities, missing data.

Fitting Models to Data (BEHAVE).

$$\arg\min_{\mathbf{x}} \mathrm{dist}(f(I), f_m(M(\mathbf{x})))$$

Model Parameters

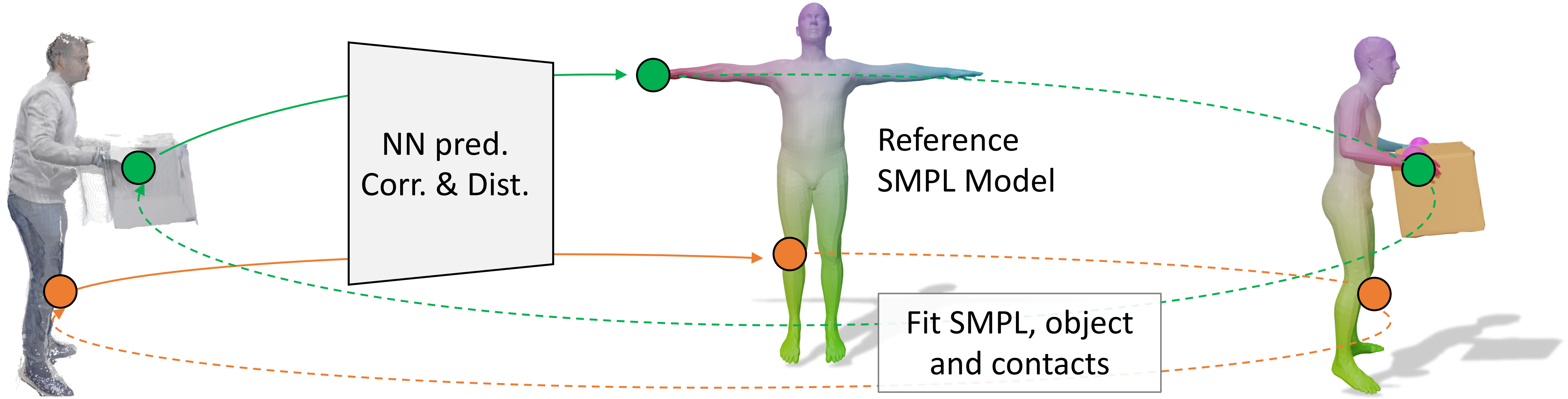


Data

World Model

Key idea: Use neural fields to make optimization well behaved

Overview.

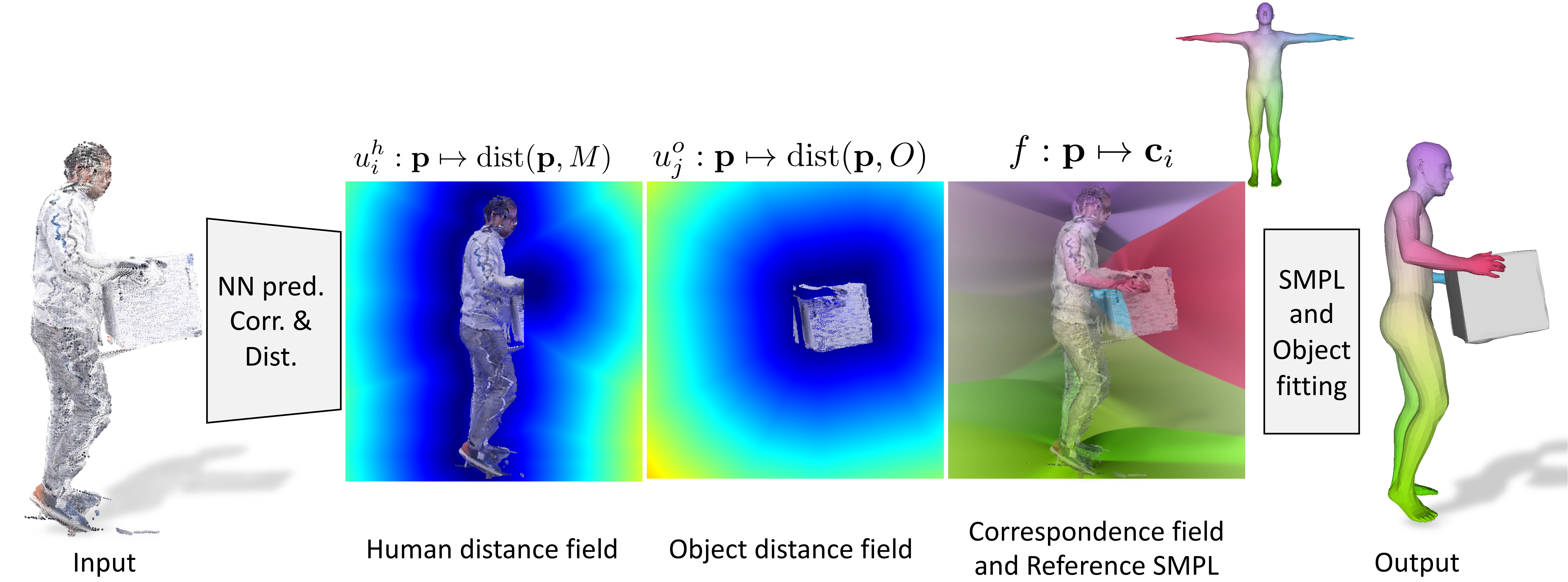


Input: Segmented human and object multi-view PC

Output: Registered SMPL, object and contacts

BEHAVE, CVPR’22

BEHAVE predictions.



BEHAVE, CVPR'22

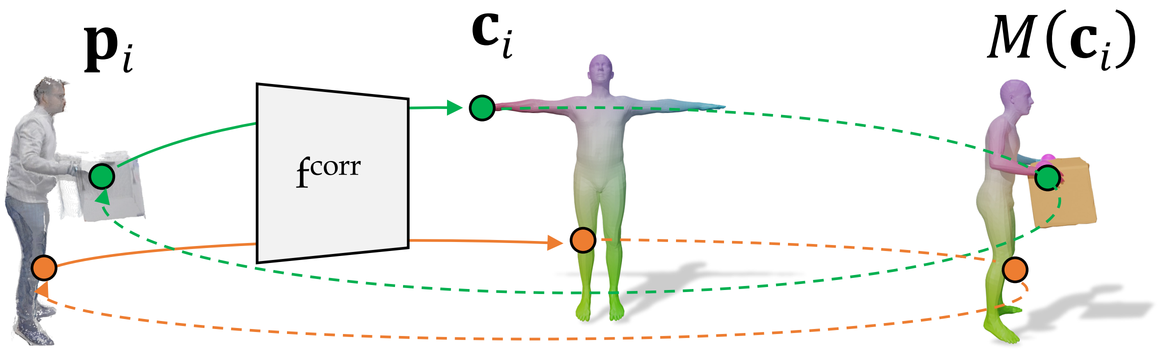
BEHAVE formulation.

1. The SMPL model $M(.)$, should fit the input human.
2. The object mesh should fit the input object.
3. SMPL model and object mesh should satisfy contacts.

$$\boxed{E^{\mathrm{SMPL}}} = \sum_{i=1}^{N} ||\mathbf{p}_i - M(\mathbf{c}_i)|_2 - u_i^h|| + E^{\mathrm{reg}}$$

$\mathbf{c}_i$: Predicted correspondence

u_i^{h} : Predicted distance to human



BEHAVE, CVPR’22

BEHAVE formulation.

1. The SMPL model $M(.)$, should fit the input human.
2. The object mesh should fit the input object.
3. SMPL model and object mesh should satisfy contacts.

$$\boxed{E^{\text{obj}}} = \sum_{\mathbf{v}_j \in O} |u_j^o| + d(O, S^o)$$

$\mathbf{v}_\mathrm{j}$: Vertex on object template

u_j^o : Predicted distance to object

S^O : Input object point cloud

$d(.,.)$: Chamfer distance

BEHAVE, CVPR'22

BEHAVE formulation.

1. The SMPL model $M(.)$, should fit the input human.

2. The object mesh should fit the input object.

3. SMPL model and object mesh should satisfy contacts.

$$\boxed{E^{\text{cont}}} = \sum_{\mathbf{v}_j \in O} \mathbf{1}_j^c |\mathbf{v}_j^o - M(\mathbf{c}_j)|$$

$\mathbf{1}_j^c$ = 1 iff u_j^o, u_j^h < 2cm

$\mathbf{c}_j$: Predicted correspondence

BEHAVE, CVPR'22

Tracking human, object and contacts.

RGB sequence

Tracking with BEHAVE model

BEHAVE, CVPR'22

Tracking human, object and contacts.

RGB sequence

Tracking with BEHAVE model

BEHAVE, CVPR’22

Remaining Problem.

Capturing with external cameras imposes restrictions on the size of the recording volume and the time.

Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors

Vladimir Guzov[1,2], Aymen Mir[1,2], Torsten Sattler[3], Gerard Pons-Moll[1,2]

[1]University of Tübingen, Germany

[2]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

[3]Czech Technical University in Prague

Our goals

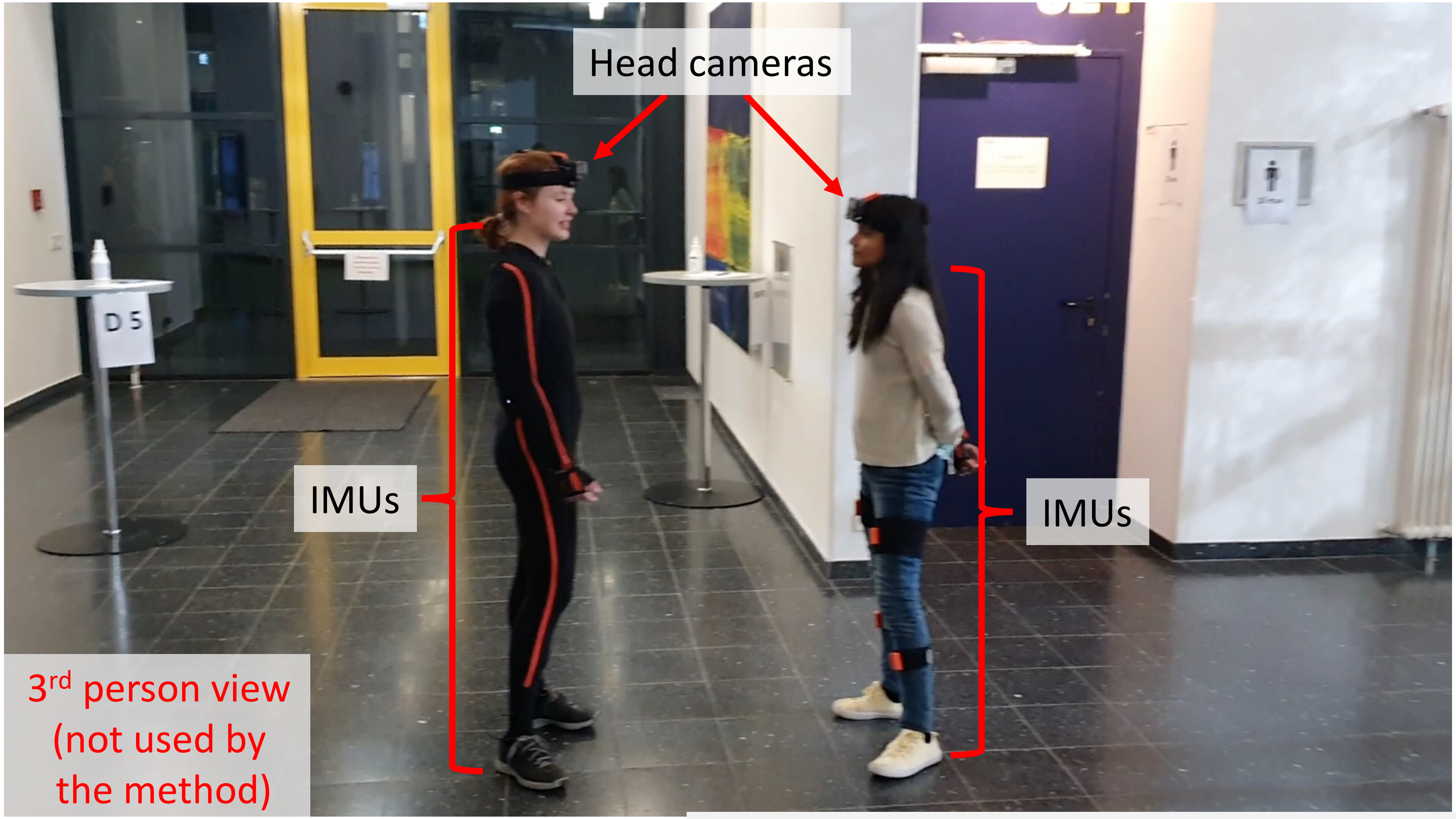
Given a human,

3rd person view
(not used by the method)

Guzov & Mir et al. **HPS**, CVPR'21 [Oral, **Best paper finalist**]

<u>Goal</u>: We want to capture **large** scenes and **long** recordings

→ **Wearable** sensors, **no external** cameras



Guzov & Mir et al. **HPS**, CVPR'21 [Oral, **Best paper finalist**]

- We want to register the digital human within the digital 3D environment.

- Therefore, we **pre-scanned** several **large 3D environments.**

Guzov & Mir et al. **HPS**, CVPR’21 [Oral, **Best paper finalist**]

Guzov & Mir et al. **HPS**, CVPR’21 [Oral, **Best paper finalist**]

How can we capture human motion, and register it with the 3D scenes without external cameras?

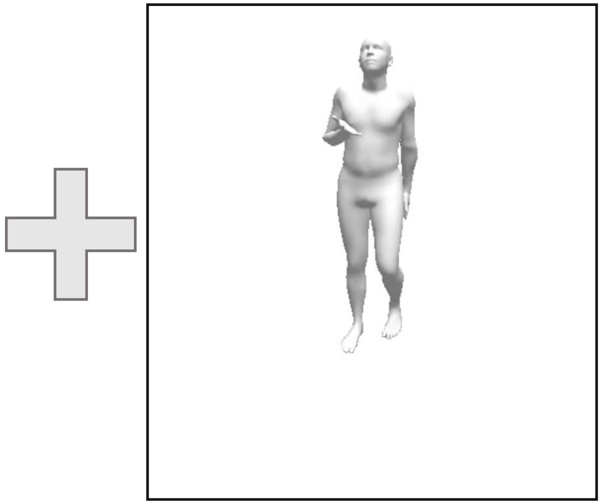
Guzov & Mir et al. **HPS**, CVPR’21 [Oral, **Best paper finalist**]

We can capture human motion with IMUs: but this method is not aware of the scene.

Guzov & Mir et al. **HPS**, CVPR’21 [Oral, **Best paper finalist**]

Our initial solution: combine self-localization with IMU.

IMU pose estimation

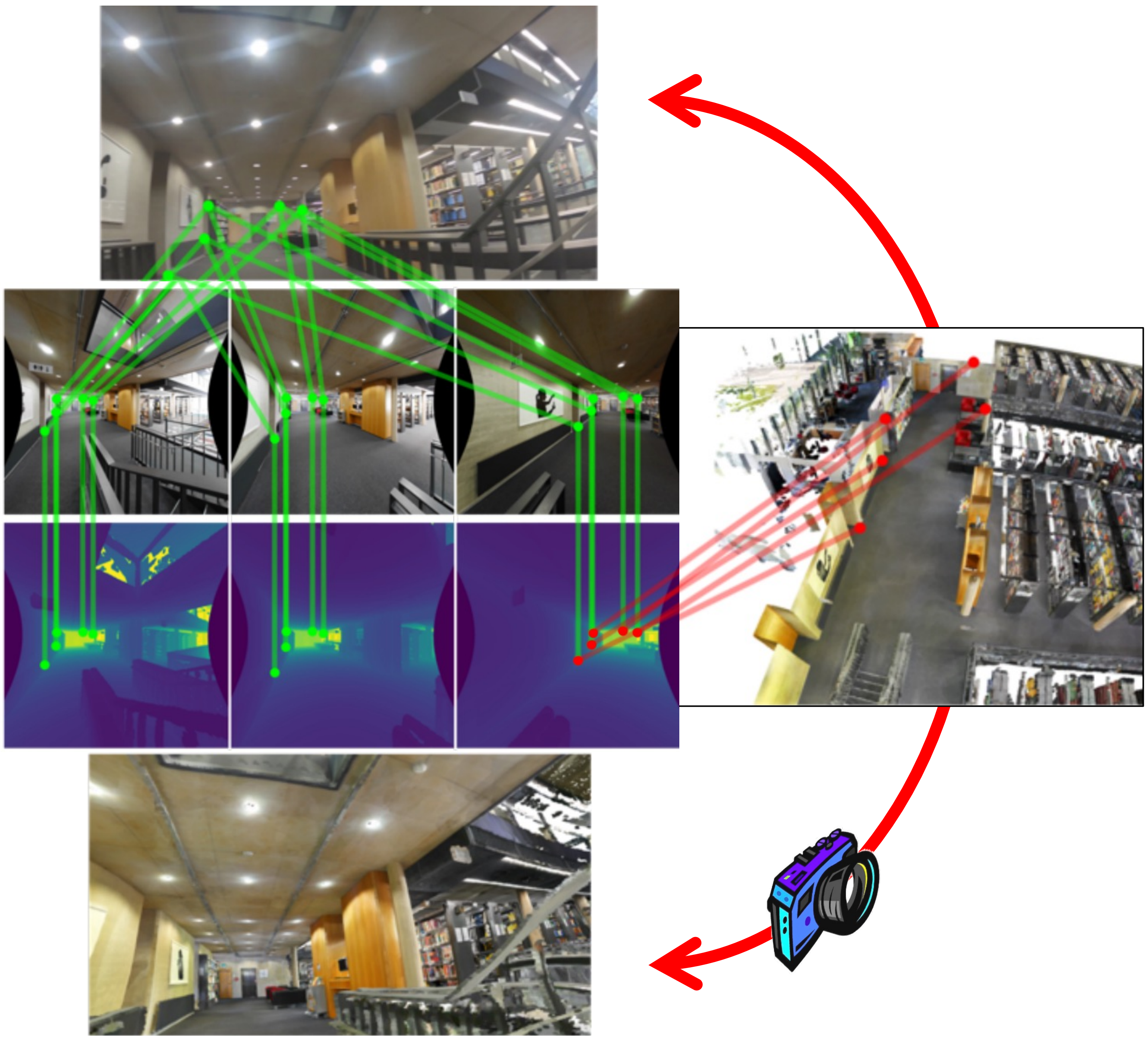


Guzov & Mir et al. **HPS**, CVPR’21 [Oral, **Best paper finalist**]

What is self localization?

Given a 3D scene…

Find the 6D pose of the camera that took the test image.



1. Extract features (keypoints/ edges etc.) of the test image.

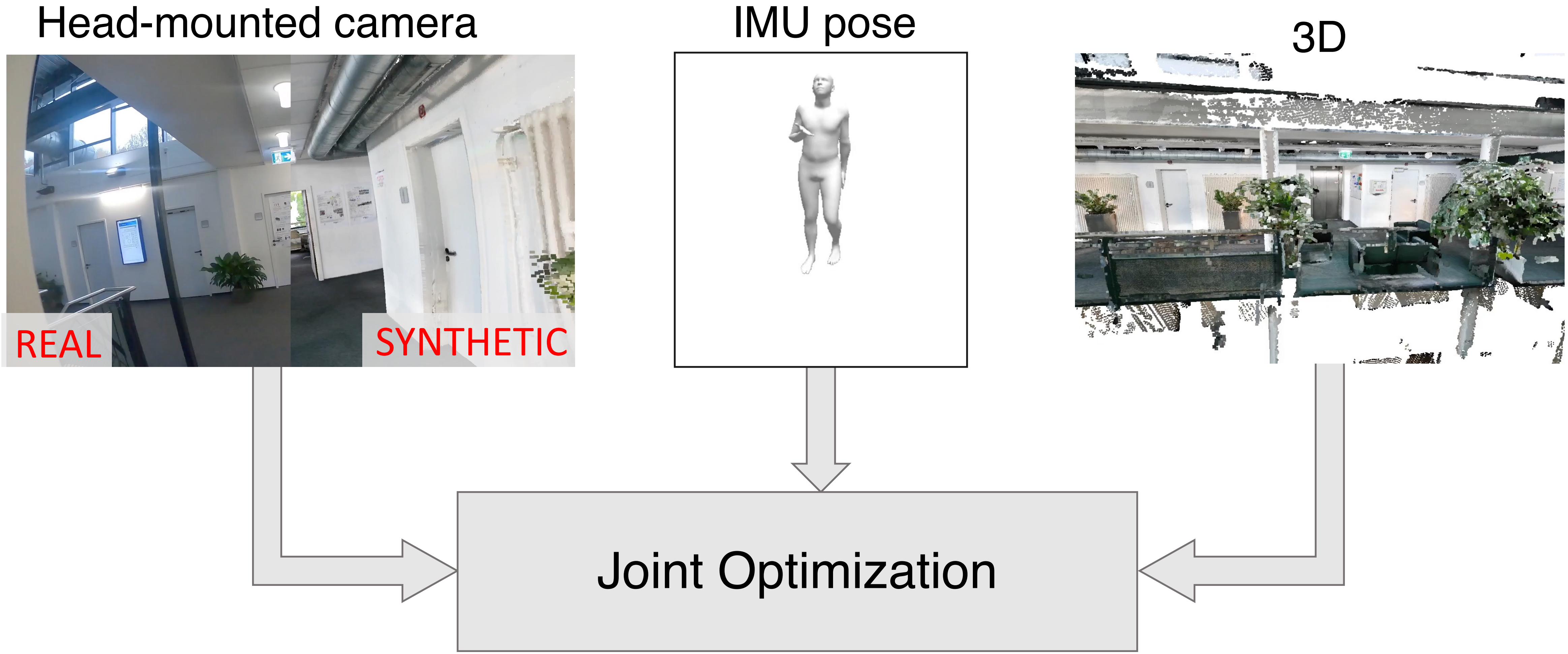
2. Retrieve similar looking parts of 3D scene using the features.

3. Optimize the camera until features match.

But camera localization is noisy. Hence, the resulting motion is unstable and unrealistic.

Guzov & Mir et al. **HPS**, CVPR’21 [Oral, **Best paper finalist**]

Our **solution**: HPS to jointly optimize self-localization with IMU **and scene.**



Guzov & Mir et al. **HPS**, CVPR'21 [Oral, **Best paper finalist**]

HPS Objective.

$$E(\theta_{1:T}, \mathbf{t}_{1:T}) = w_1 E_{self} + w_2 E_{\mathrm{scene}} + w_3 E_{\mathrm{sm}} + w_4 E_{\mathrm{IMU}}$$

Optimize SMPL pose and translation over T frames

Scene contact term.
When IMU detects a contact b/w foot and scene, the foot should not slide.

Pose term.
SMPL pose should match IMU.

Self-localization term.
Match 3D scene against images from head mounted camera.

Motion smoothness term.

HPS Results

Guzov & Mir et al. **HPS**, CVPR'21 [Oral, **Best paper finalist**]

We captured a large variety of motions

This includes exercising, dancing, interacting with a scene objects and more

Key takeaways

- Capturing HOI is important and challenging.
- BEHAVE proposes a simple capture solution.
- We can uses data and neural fields to fit SMPL and object mesh.
- External cameras not suitable for long range/ long time recordings.
- Joint localization and IMU based optimization can track person using just body mounted IMUs and camera.