Perception and Generation of Physical Interactions

Srinath Sridhar

GAMES Seminar

August 12, 2021

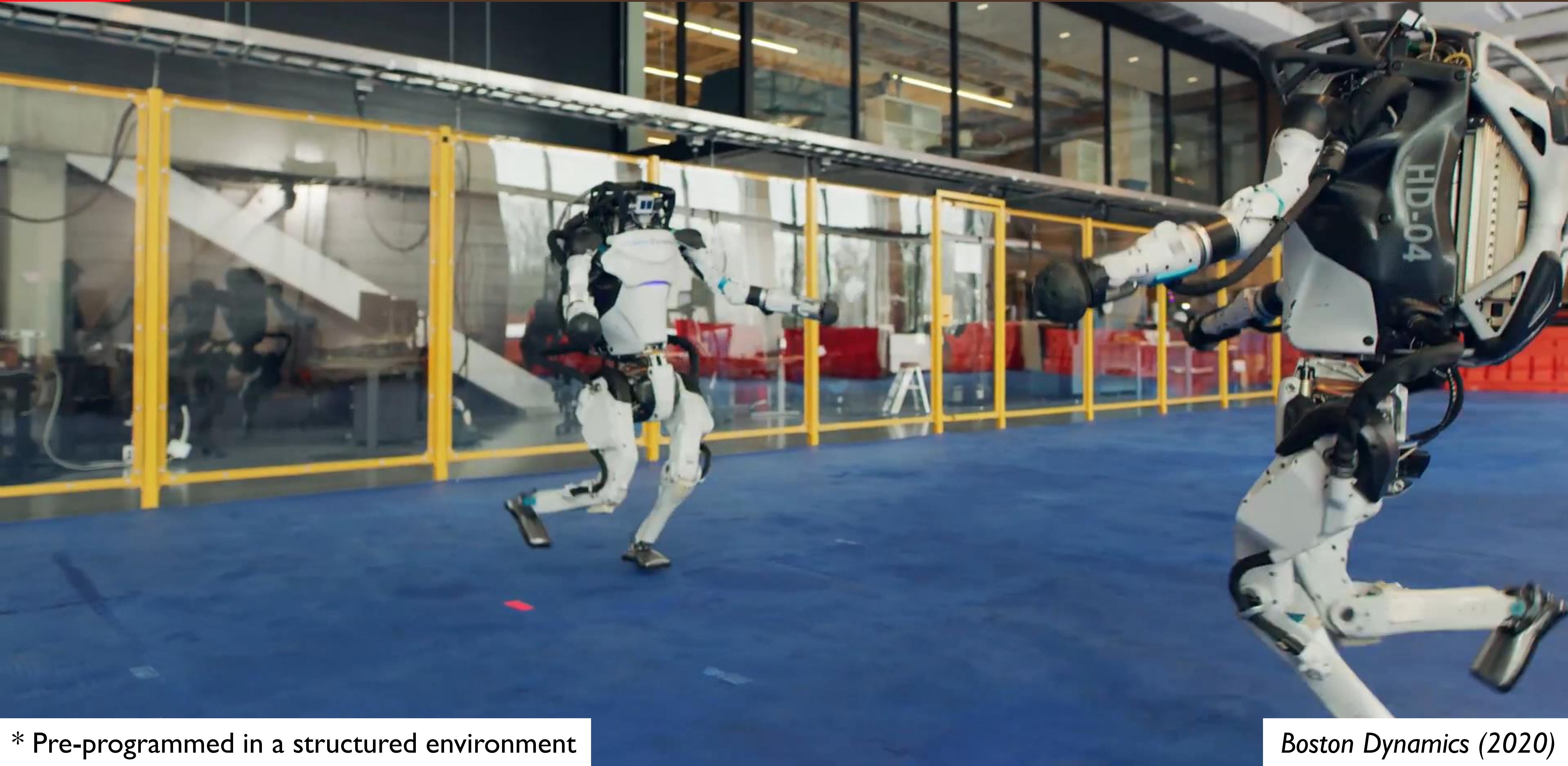
BROWN

AI for Logical Reasoning

Deep Blue versus Garry Kasparov (1997)

AlphaGo versus Lee Sedol (2016)

AI for Physical Motion*



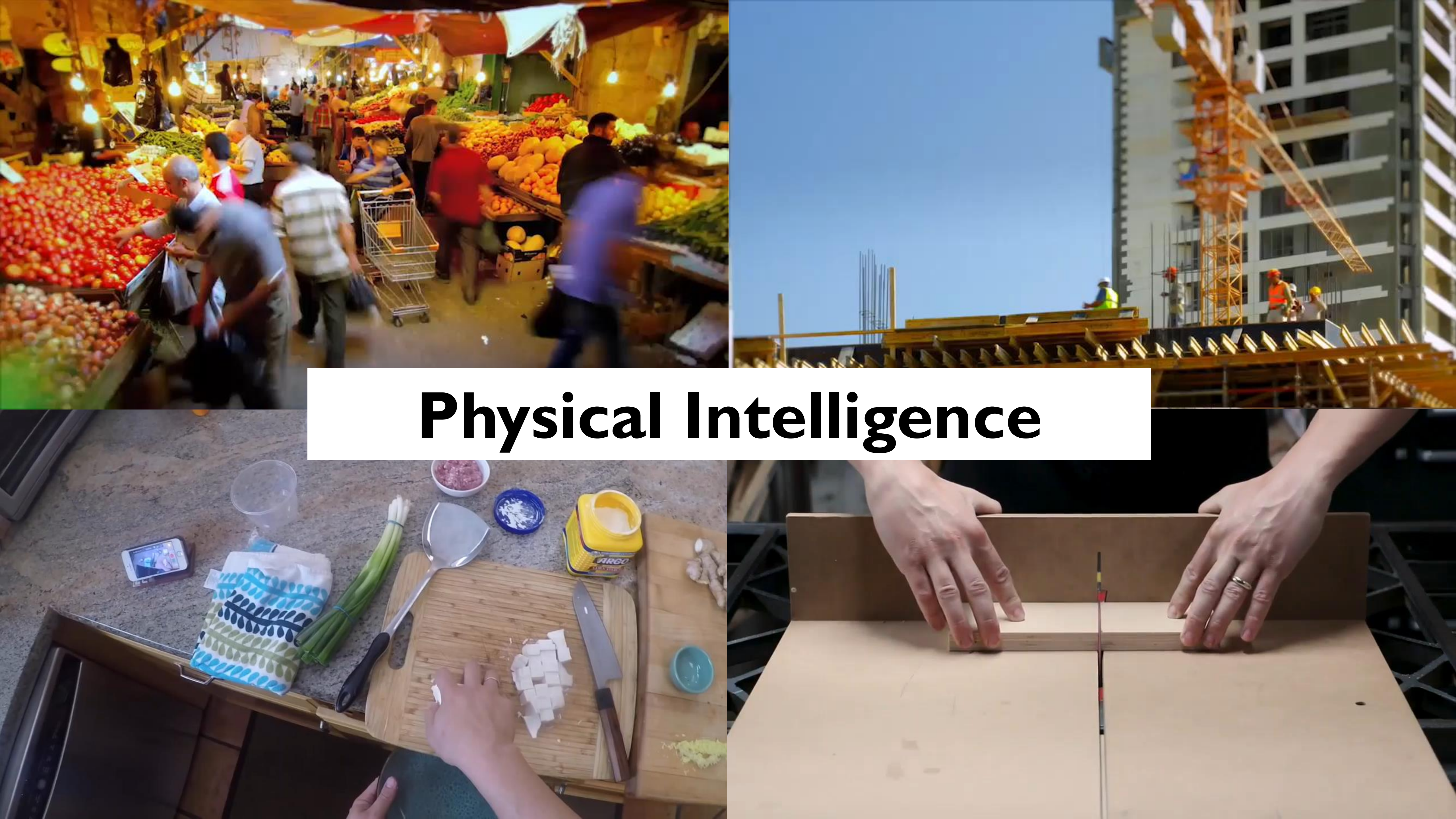
* Pre-programmed in a structured environment

Boston Dynamics (2020)

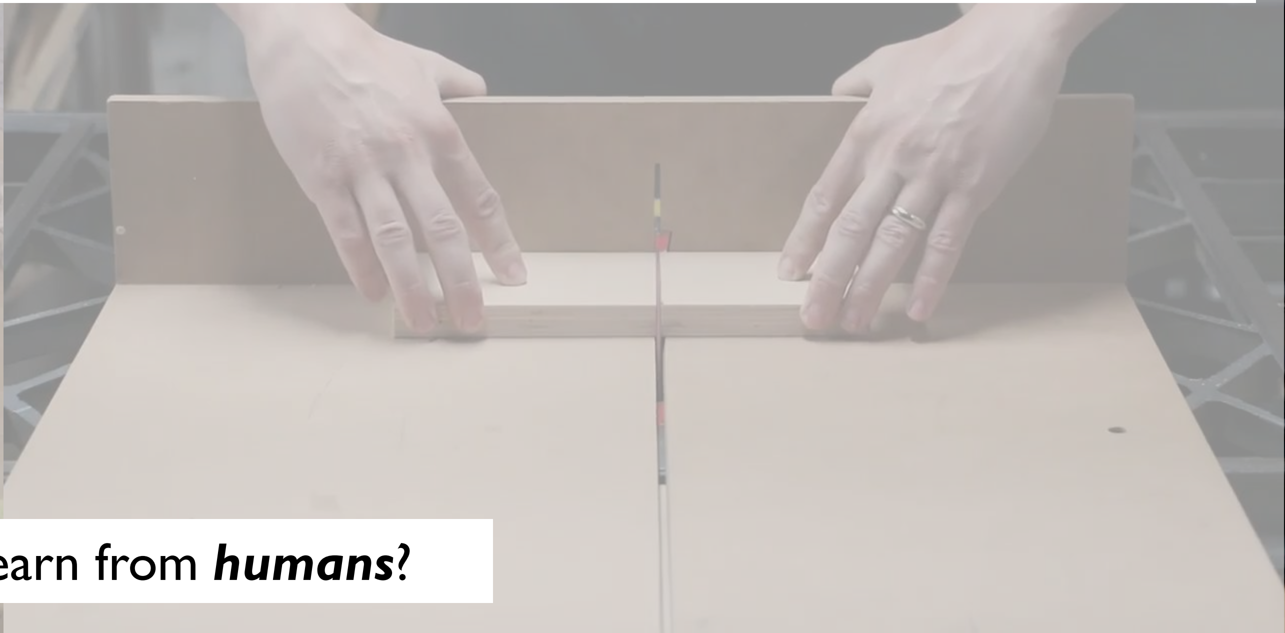
AI for Physical Interaction

Boston Dynamics (2017)

Physical Intelligence



Learn physical intelligence by **observing humans**



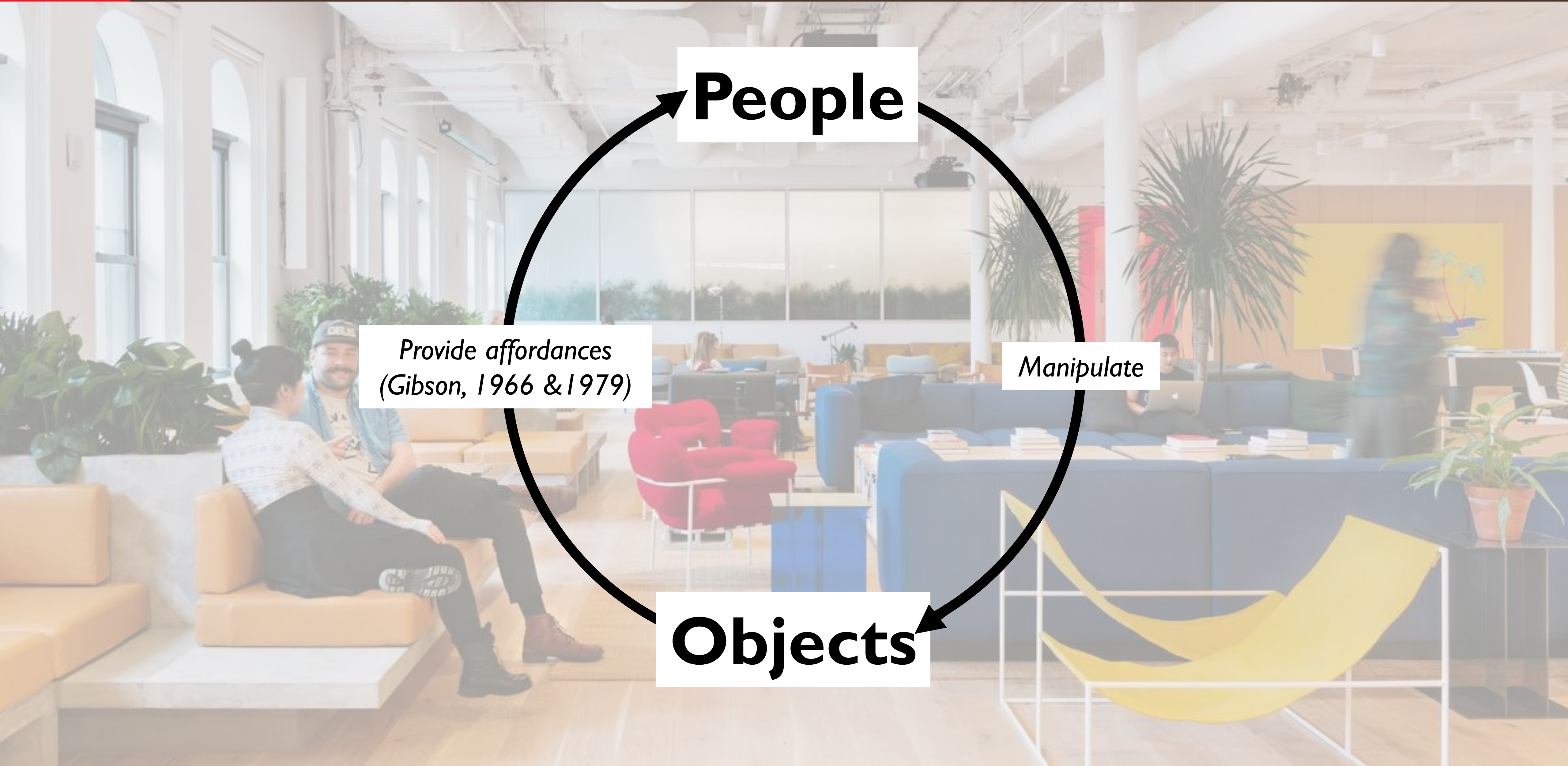
Hmmm, but why learn from ***humans***?

Applications

Interactive Household Robots

Interactive Mixed Reality

Human Physical Intelligence



People

Manipulate

Objects

Provide affordances (Gibson, 1966 &1979)

BROWN

Understanding Physical Intelligence

Understanding = Perception + Generation

What I cannot create, I do not understand.

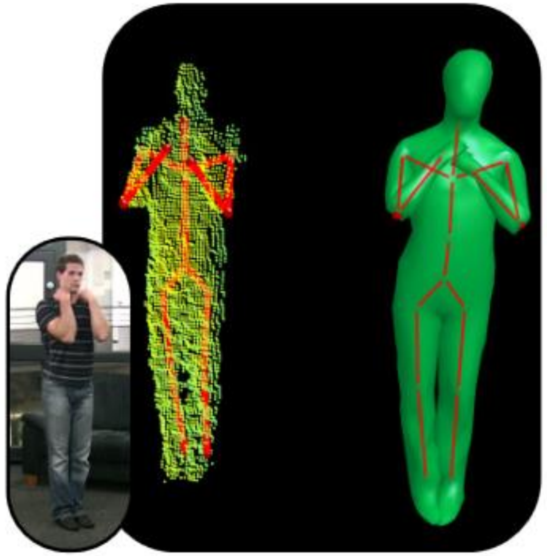
- Richard Feynman

Understanding Humans

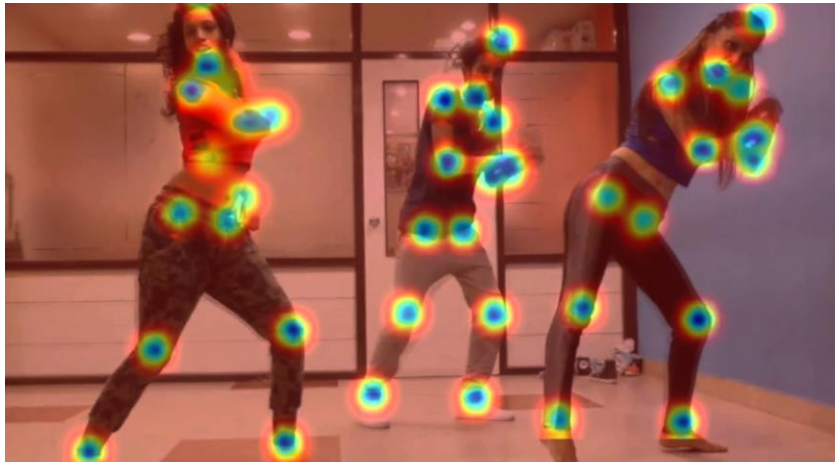
Articulated Human Pose Tracking

Shotton et al. 2013

Baak et al. 2011



Wei et al. 2016



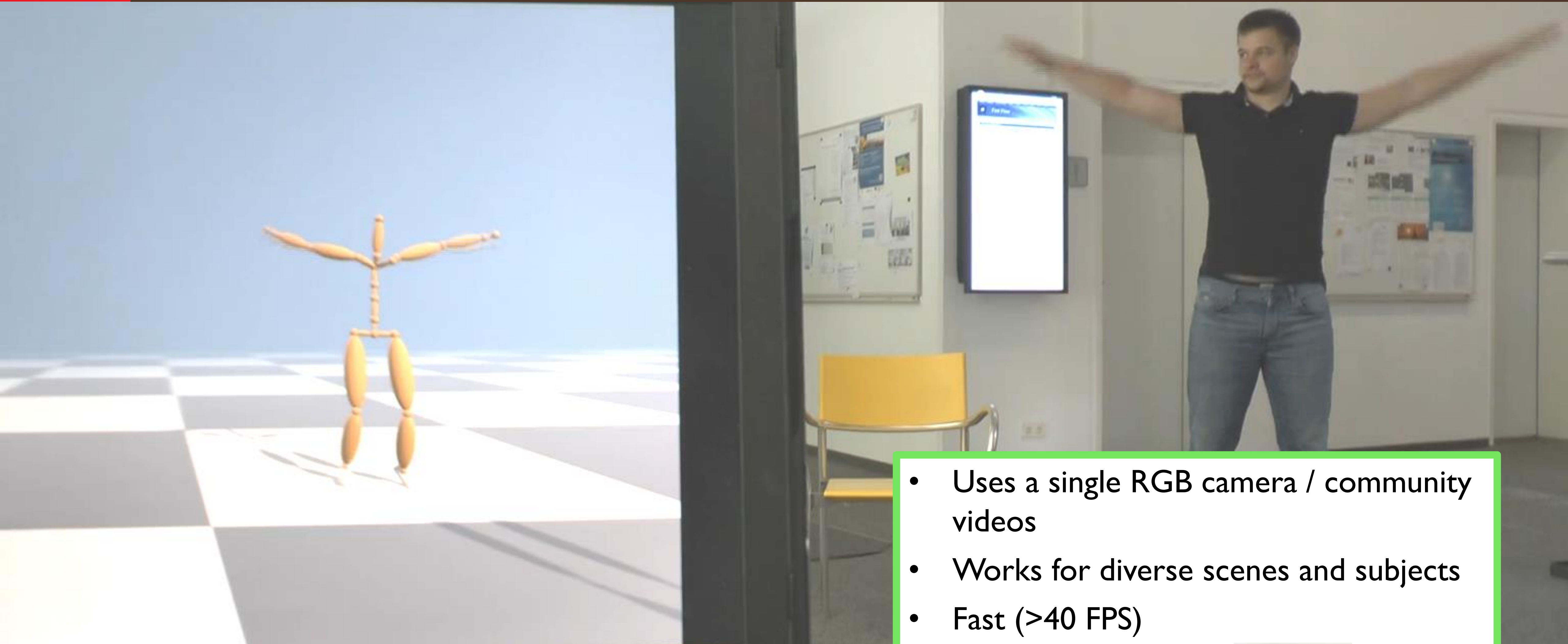
Girshick et al. 2011
Ganapathi et al. 2012
Ma and Wu 2014
Newell et al. 2016
Tompson et al. 2014
Insafutdinov et al. 2016
Cao et al. 2017
…

Limitations

- Limited to 2D
- Depth-sensor based
- Lacking generalizability

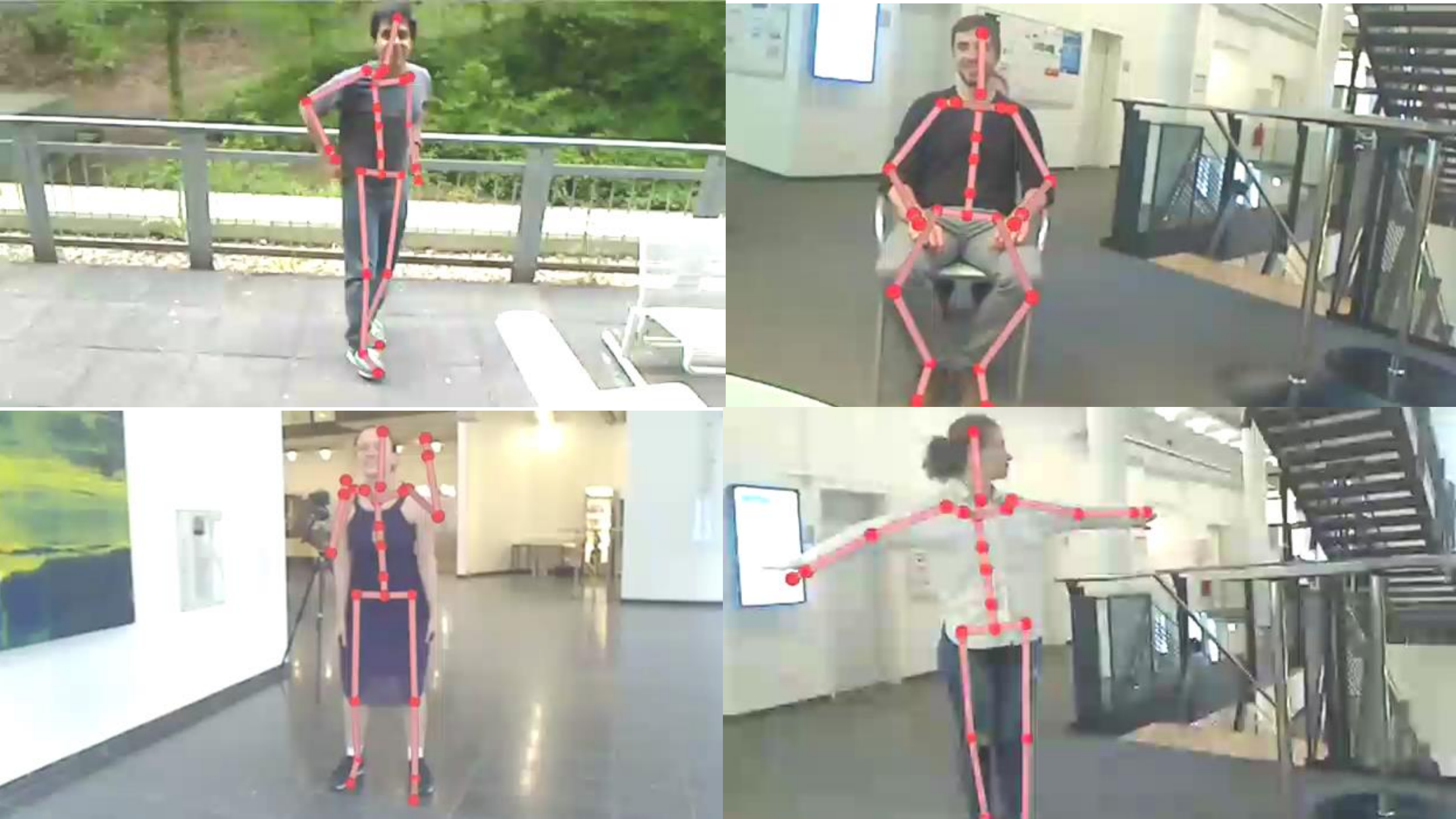
Cao et al. 2019
Zheng et al., 2019
Zhang et al. 2020
Kanazawa et al. 2018
…

Perception: 3D Human Pose

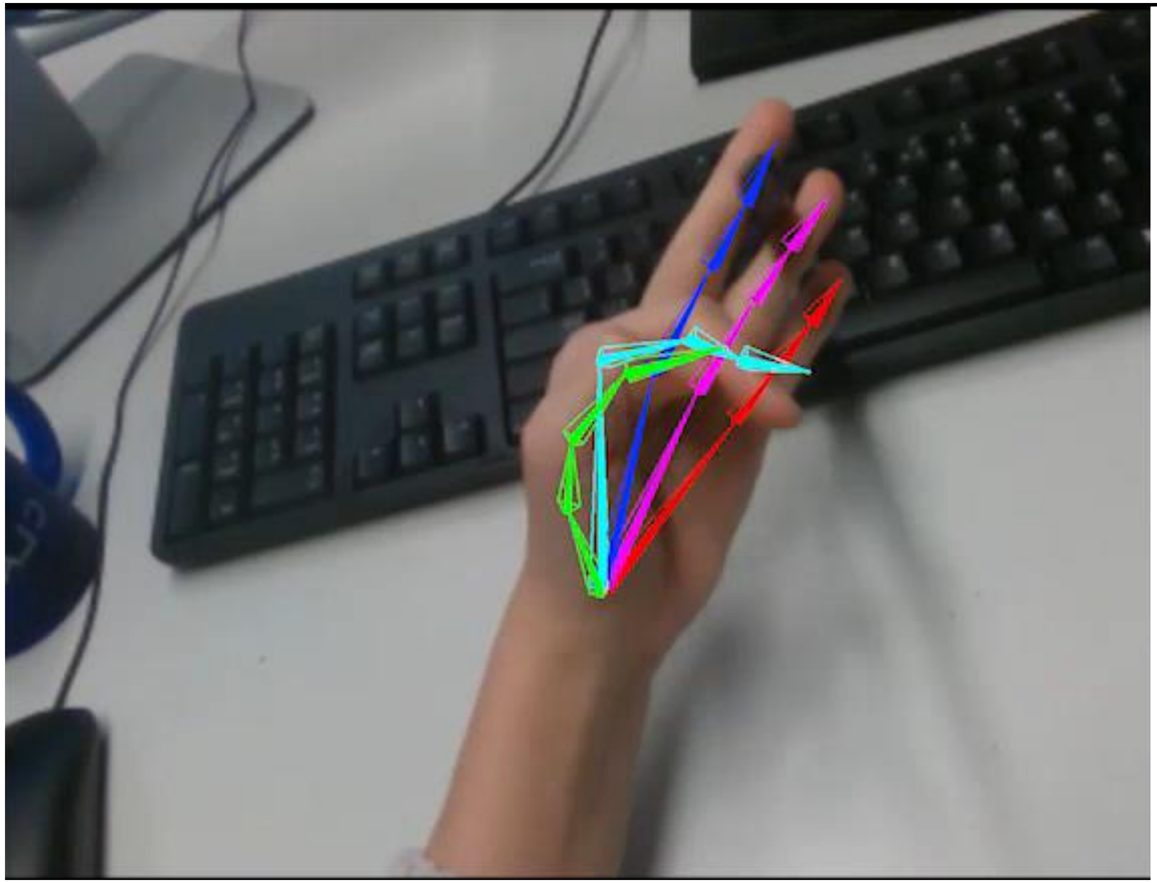


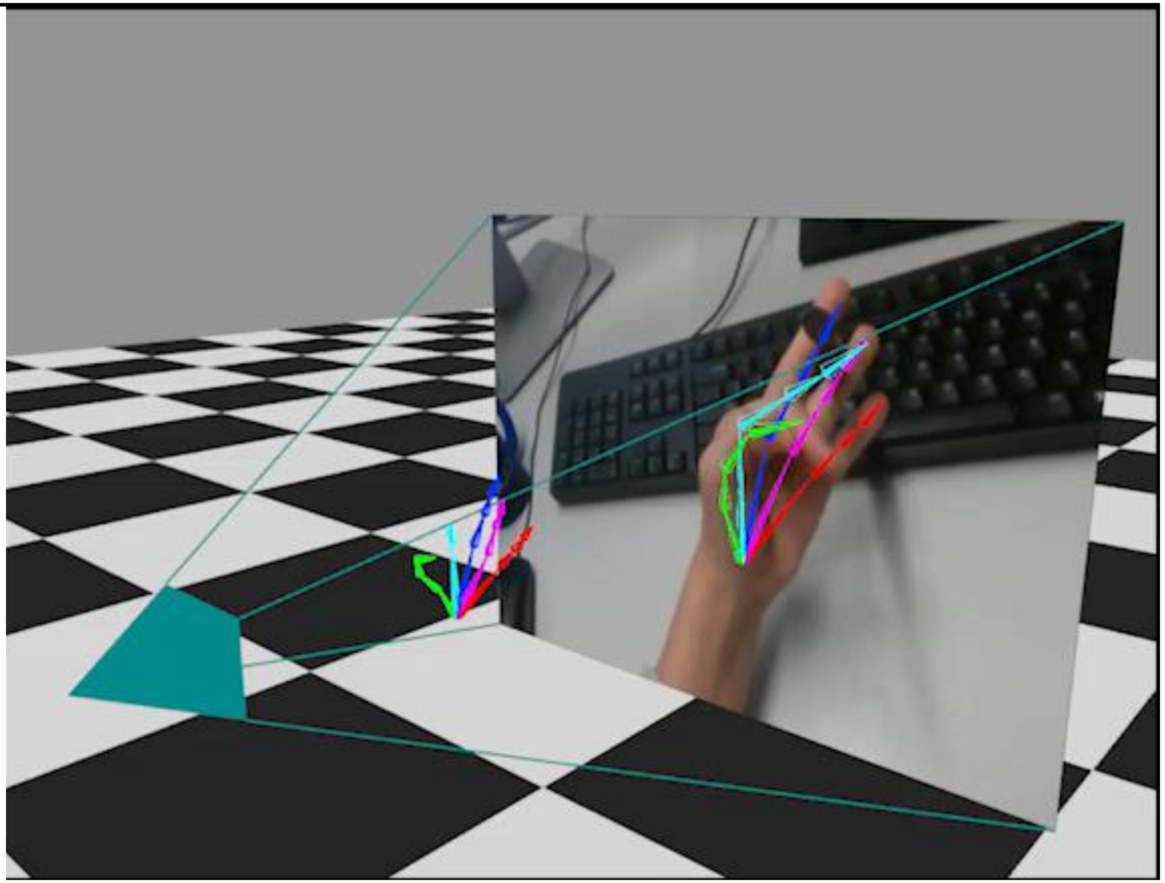
- Uses a single RGB camera / community videos
- Works for diverse scenes and subjects
- Fast (>40 FPS)

VNect. Mehta, Sridhar, Sotnychenko, Rhodin, Shafiei, Seidel, Xu, Casas, Theobalt. **SIGGRAPH 2017**



Perception: 3D Hand Pose



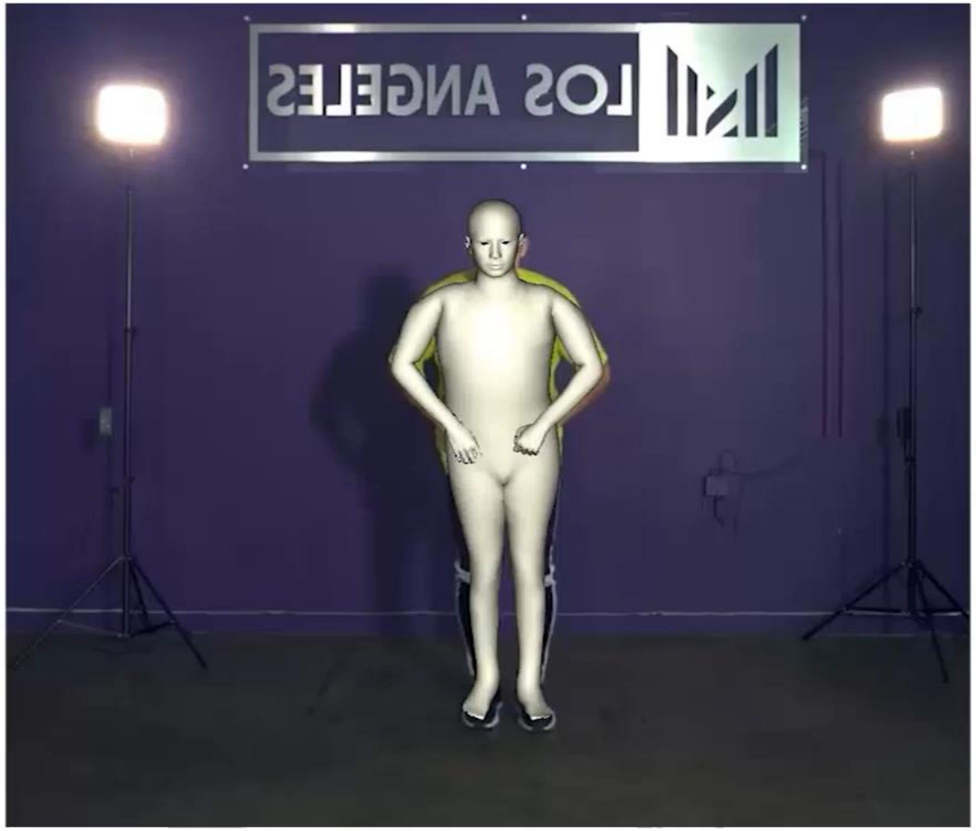


GANerated Hands. Mueller, Bernard, Sotnychenko, Mehta, Sridhar, Casas, Theobalt. **CVPR 2018**

Sufficient for understanding?

Input Video

Monocular Total Capture

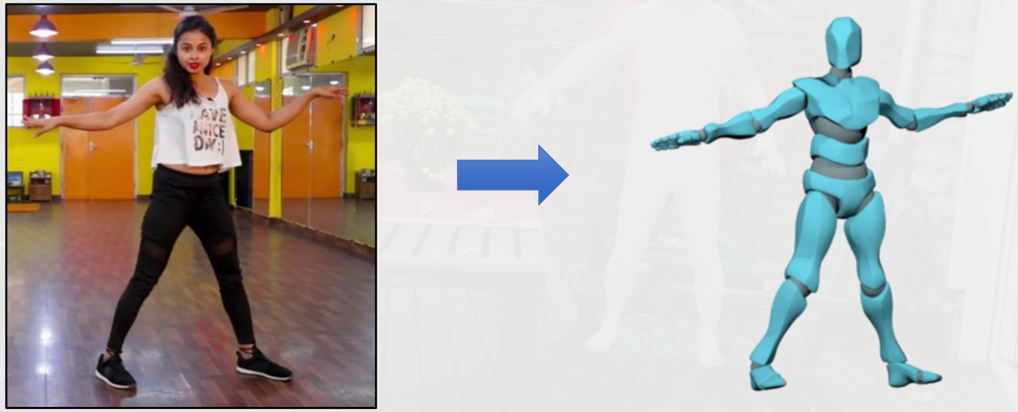


Xiang et al., CVPR 2019

Sufficient for understanding?

VIBE

Animation

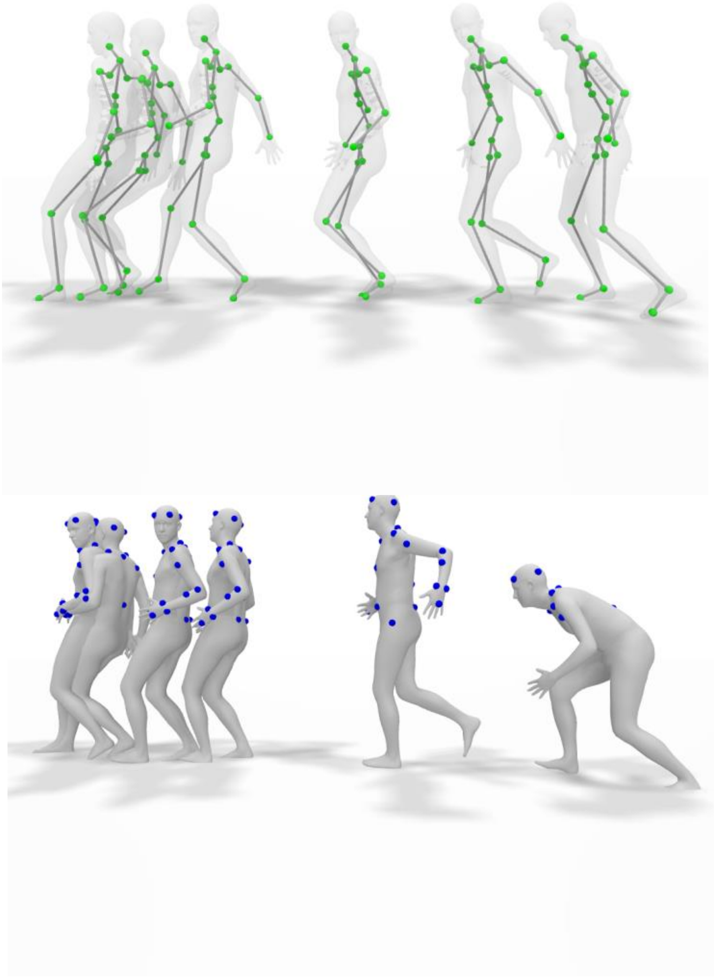


Alternate View

Physical Interactions

Kocabas et al., CVPR 2020

Generation: Human Motion Prior



HuMoR:
3D Human Motion Model for Robust Pose Estimation

Davis **Rempe**, Tolga **Birdal,** Aaron **Hertzmann**,
Jimei **Yang,** Srinath **Sridhar,** Leonidas **Guibas**

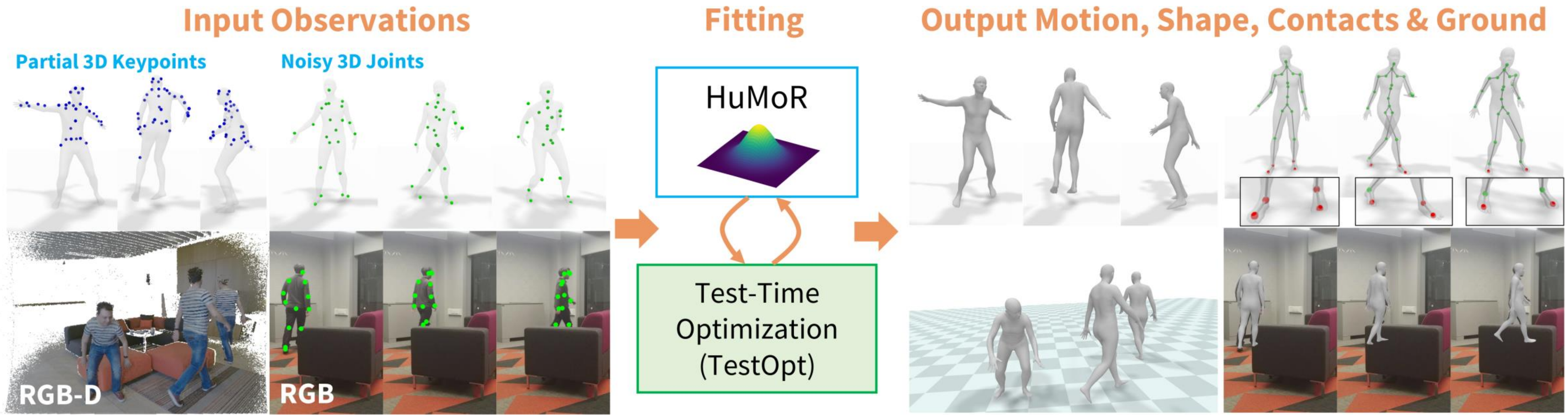
ICCV 2021 (Oral)

geometry.stanford.edu/projects/humor/

HuMoR. Rempe, Birdal, Hertzmann, Yang, Sridhar, Guibas. **ICCV 2021 [oral presentation]**

Key Ideas

1. Learned **generative model** of plausible 3D motion (HuMoR)
2. Time-time optimization (TestOpt) using **HuMoR as a prior**



State Representation

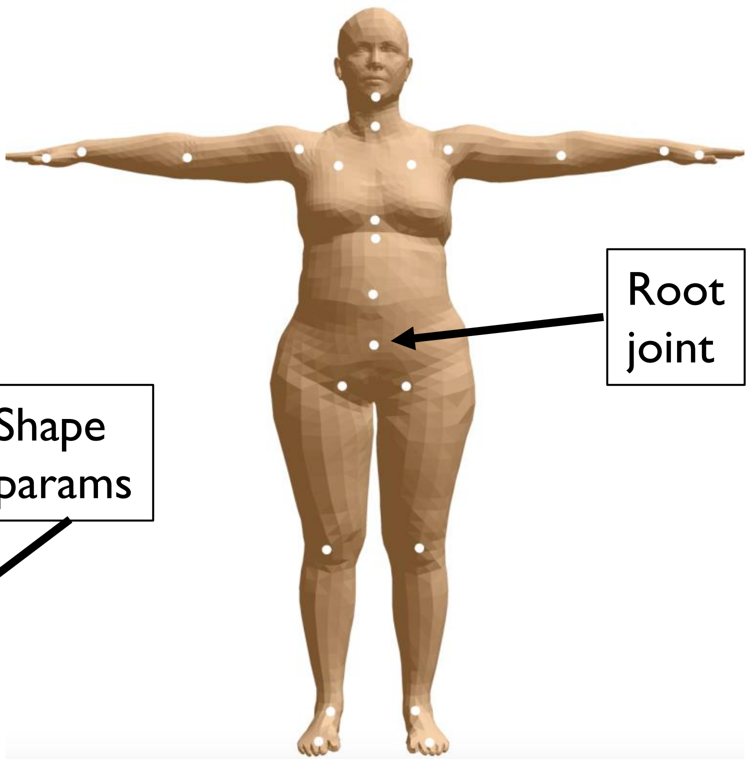
$$\mathbf{x} = \Big[\ \overbrace{\underbrace{\mathbf{r}\quad \dot{\mathbf{r}}}_{\substack{\text{Position/Vel}\\ \in \mathbb{R}^3}}\quad \underbrace{\boldsymbol{\Phi}\quad \dot{\boldsymbol{\Phi}}}_{\substack{\text{Rot/Vel}\\ \in\mathbb{R}^3}}}^{\substack{\text{SMPL}\\ \text{Root}}}\quad \overbrace{\underbrace{\boldsymbol{\Theta}}_{\substack{\text{Joint}\\ \text{Angles}\\ \in\mathbb{R}^{3\times 21}}}}^{\substack{\text{SMPL}\\ \text{Body}}}\quad \overbrace{\underbrace{\mathbf{J}\quad \dot{\mathbf{J}}}_{\substack{\text{Joint}\\ \text{Pos/Vel}\\ \in\mathbb{R}^{3\times 22}}}}^{\text{Joints}}\ \Big]$$

Over-parameterization of joint positions:

(i) Implicit through SMPL $\mathbf{J}^{\text{SMPL}} = M(\mathbf{r}, \boldsymbol{\Phi}, \boldsymbol{\Theta}, \beta)$

(ii) Explicit from state $\mathbf{J}$

SMPL



Root joint

Shape params

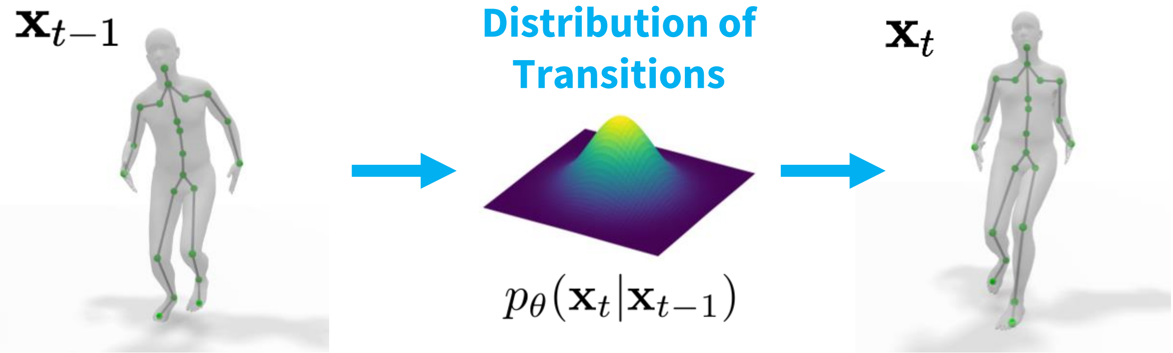
Loper et al., SIGGRAPH Asia 2015

Modeling Human Dynamics

- For **state** $\mathbf{x}_t$ at time t

$$p_\theta(\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T) = p_\theta(\mathbf{x}_0) \prod_{t=1}^{T} \underbrace{p_\theta(\mathbf{x}_t \mid \mathbf{x}_{t-1})}_{\text{HuMoR}}$$

Learn the *plausibility* of a transition, *i.e.*, **distribution of dynamics**

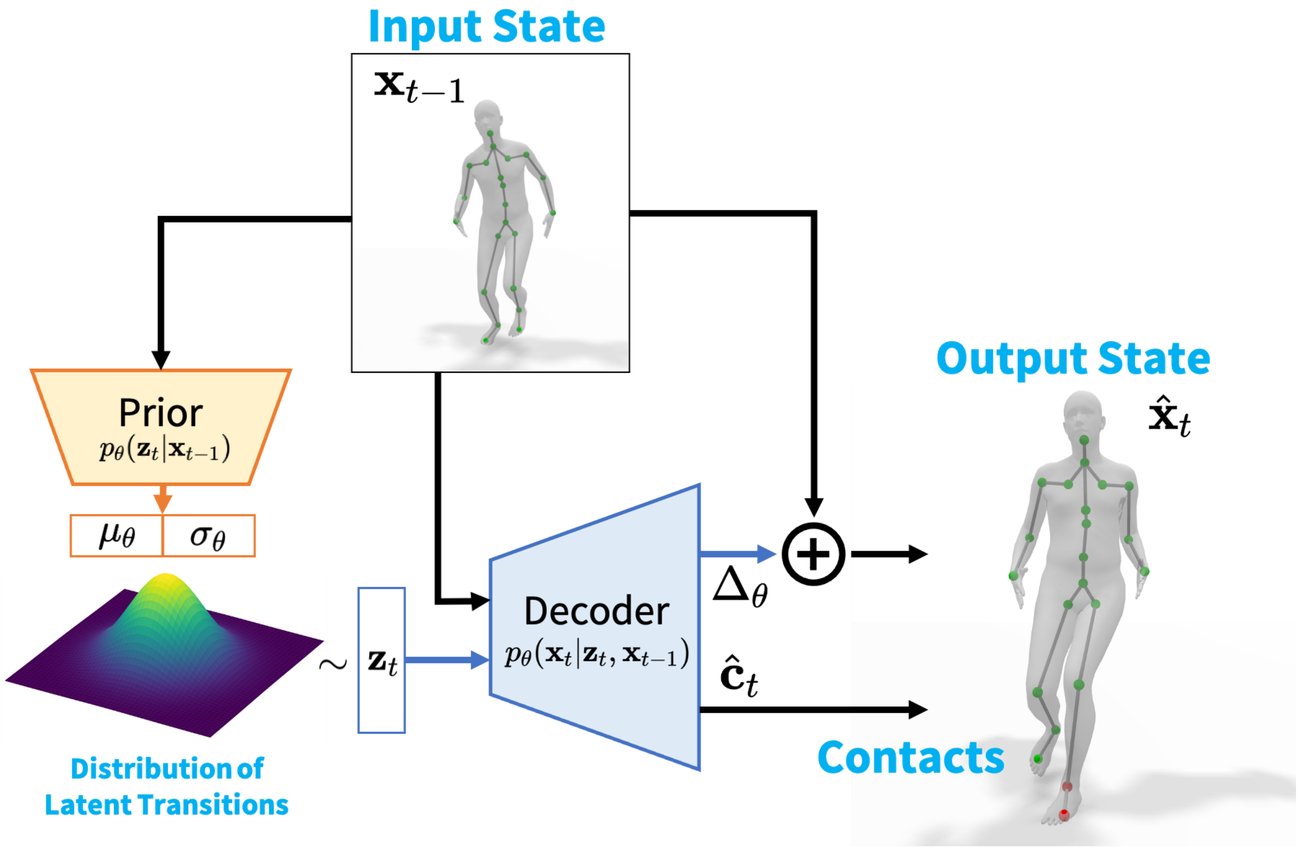

$\mathbf{x}_{t-1}$
Distribution of Transitions
$p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$
$\mathbf{x}_t$

Latent Variable Dynamics Model

- Use **latent variable model**
- Conditional VAE

Generation:

- Conditional Prior
- Decoder



Motion Model Decoder

Outputs:

Change in state Δ_θ

Ground *contact* classification $\mathbf{c}_t$

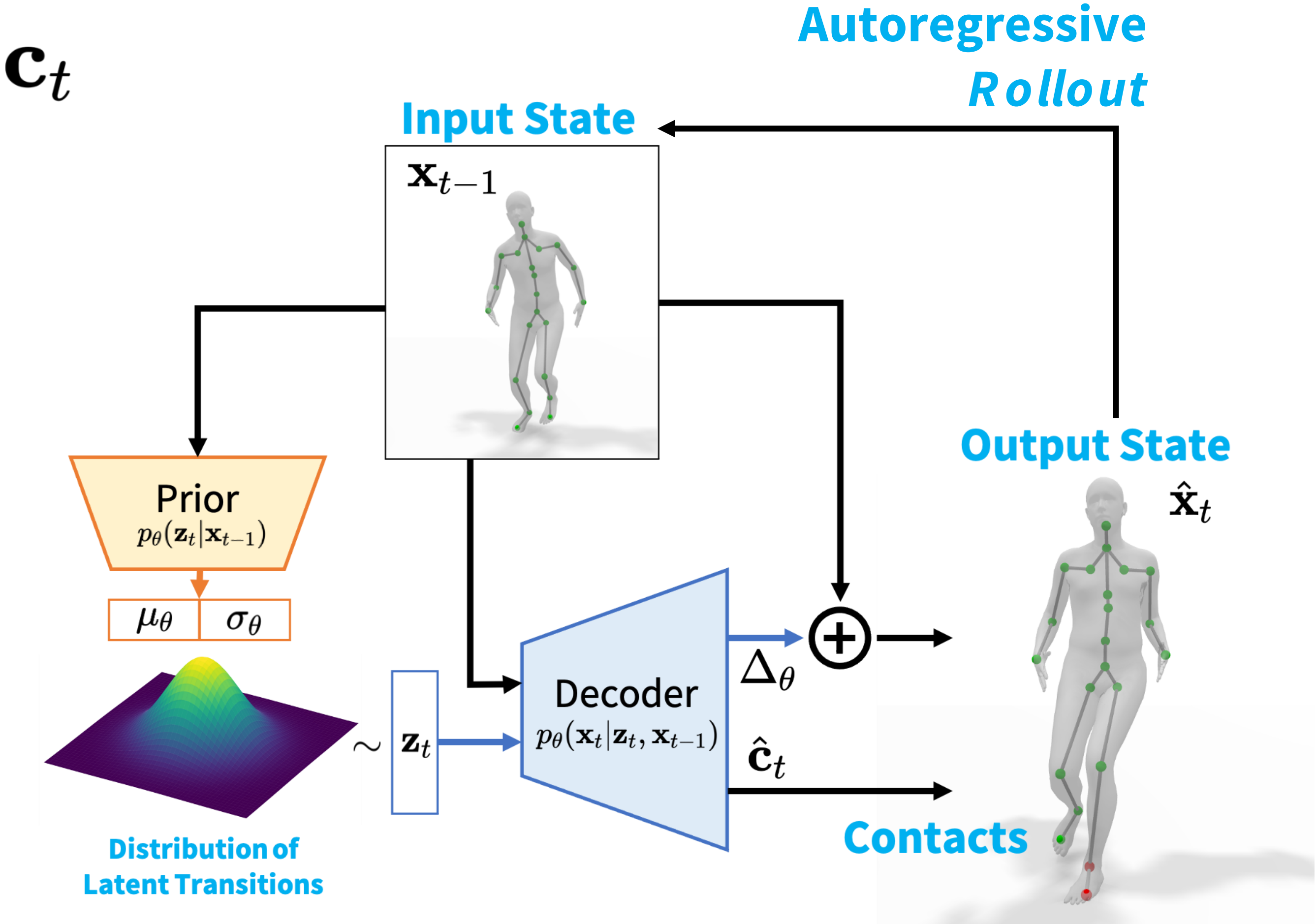
Autoregressive **sampling**

Deterministic **rollout**

Motion “parameters” $\mathbf{x}_0, \mathbf{z}_{1:T}$

give $\mathbf{x}_t = \mathbf{x}_{t-1} + \Delta_\theta(\mathbf{z}_t, \mathbf{x}_{t-1})$

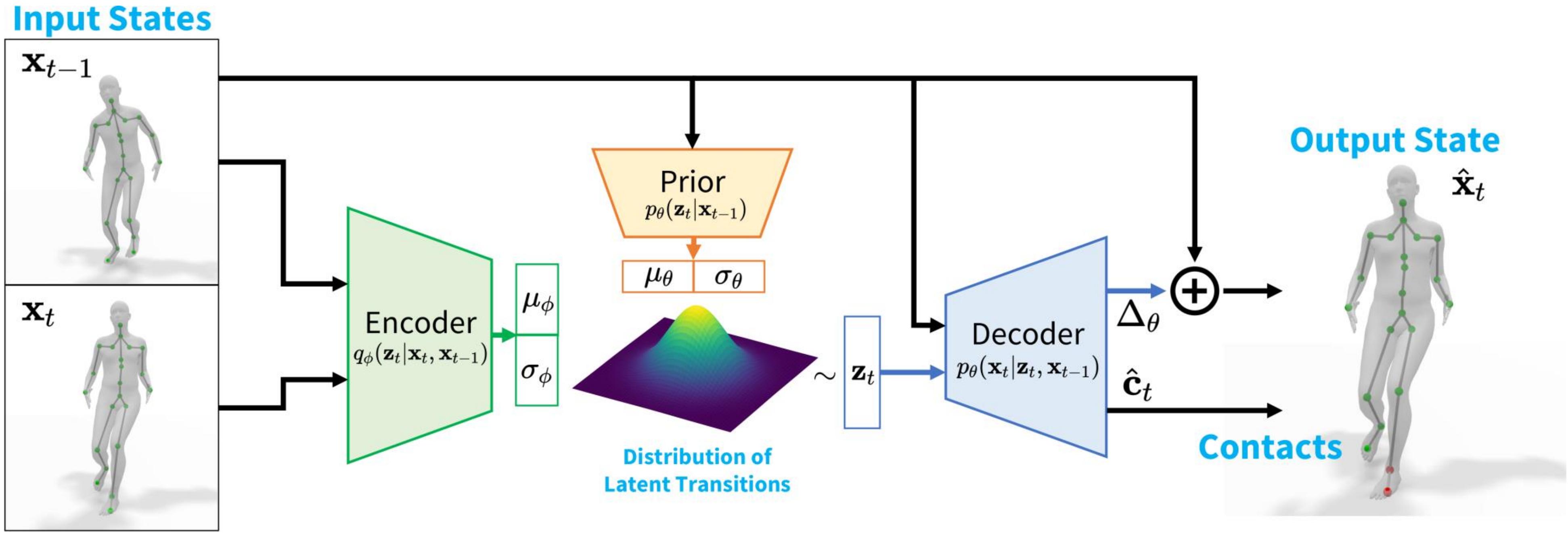
Have *prior* on $\mathbf{z}_{1:T}$



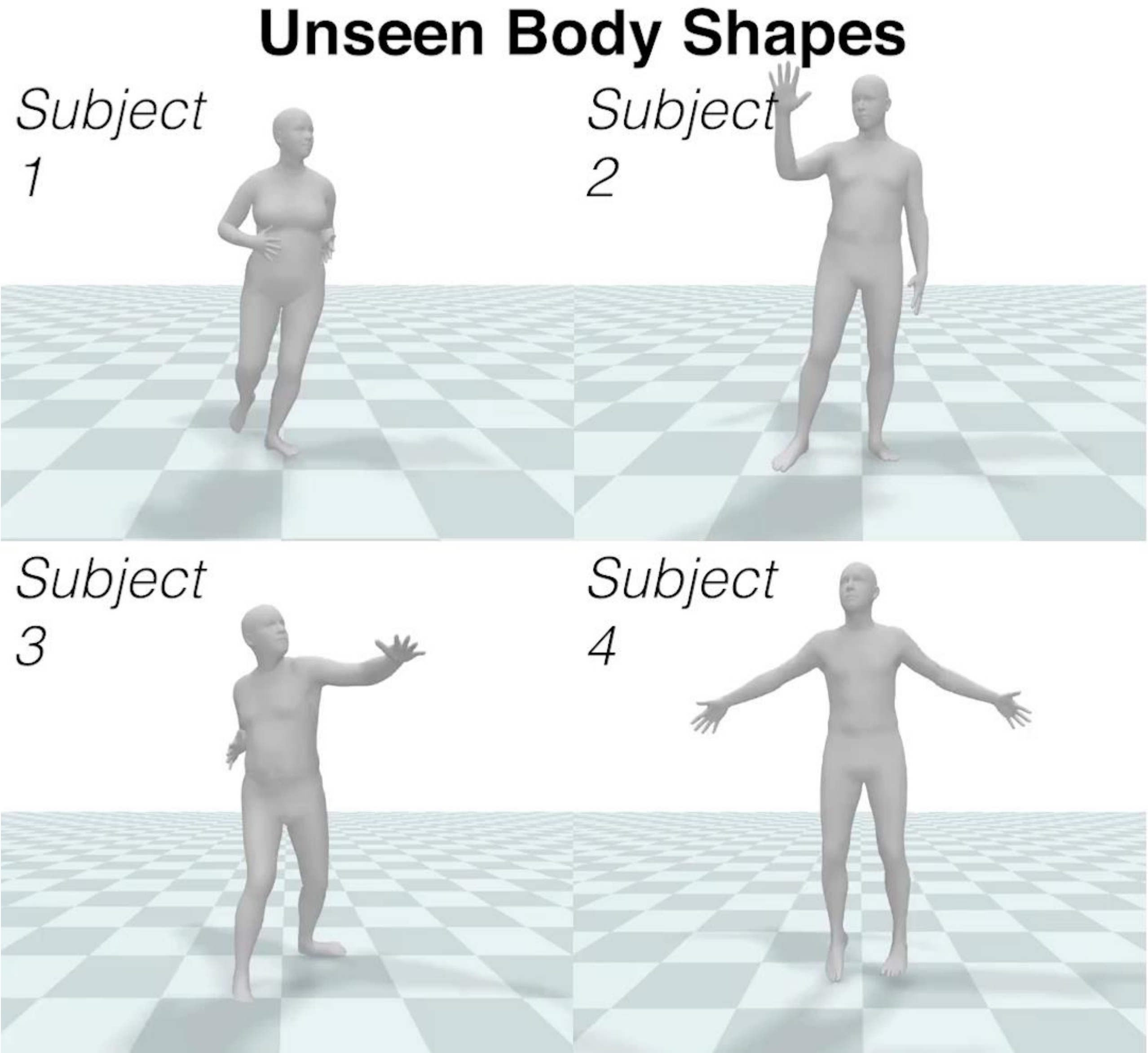
Training

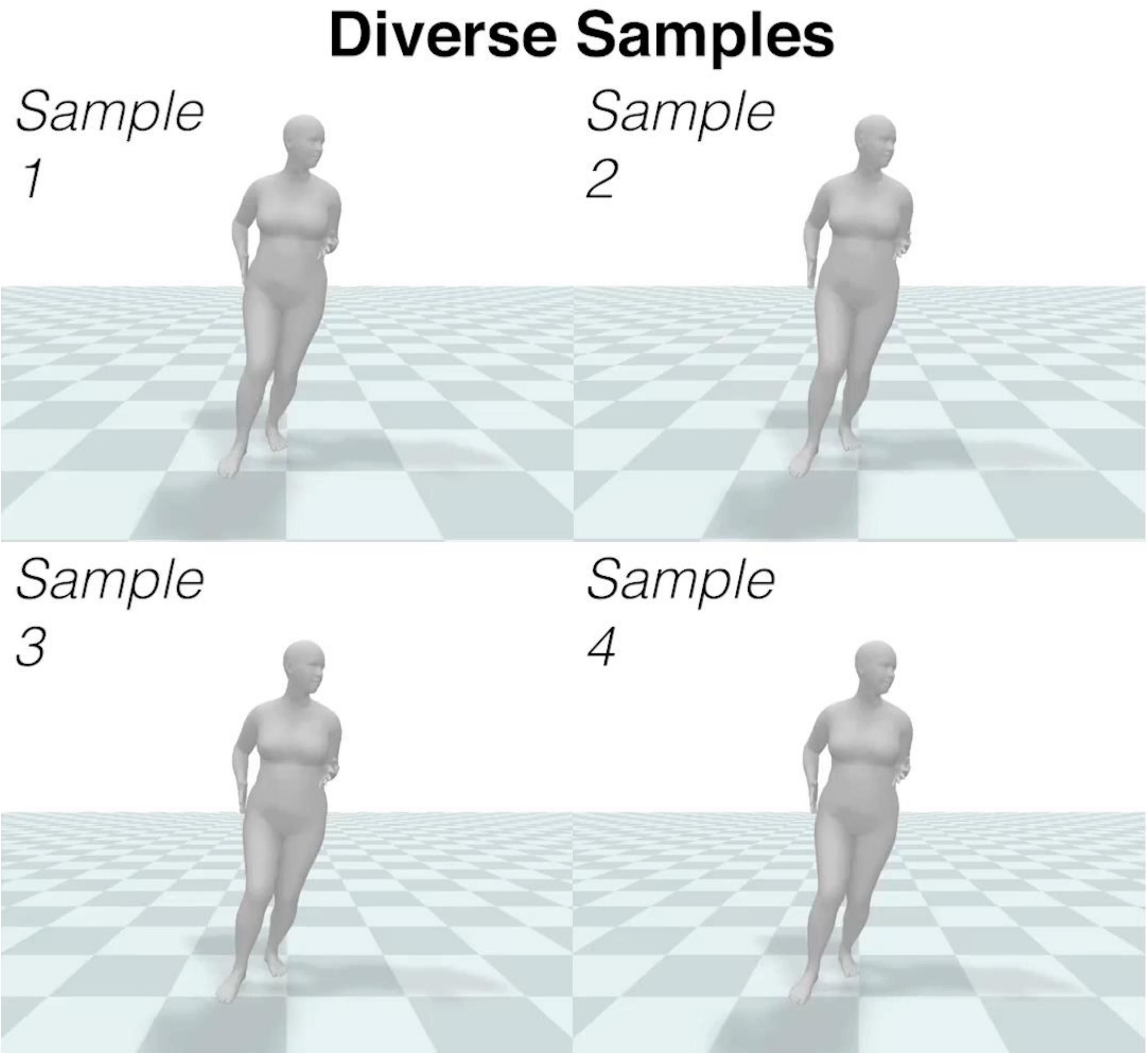
- Encoder for training on AMASS [Mahmood et al., ICCV 2019]
- Loss based on lower bound
  - Reconstruction, KL, consistency

$$\log p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}) \geq \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}_t|\mathbf{z}_t, \mathbf{x}_{t-1})] - D_{\mathrm{KL}}(q_\phi(\mathbf{z}_t|\mathbf{x}_t, \mathbf{x}_{t-1}) \parallel p_\theta(\mathbf{z}_t|\mathbf{x}_{t-1}))$$

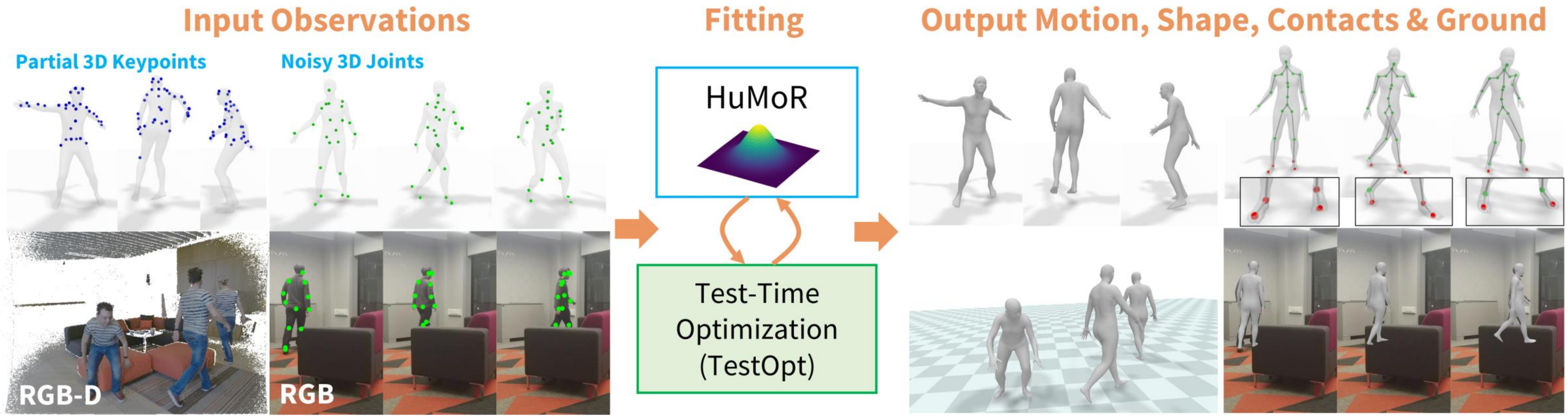


Generation Results

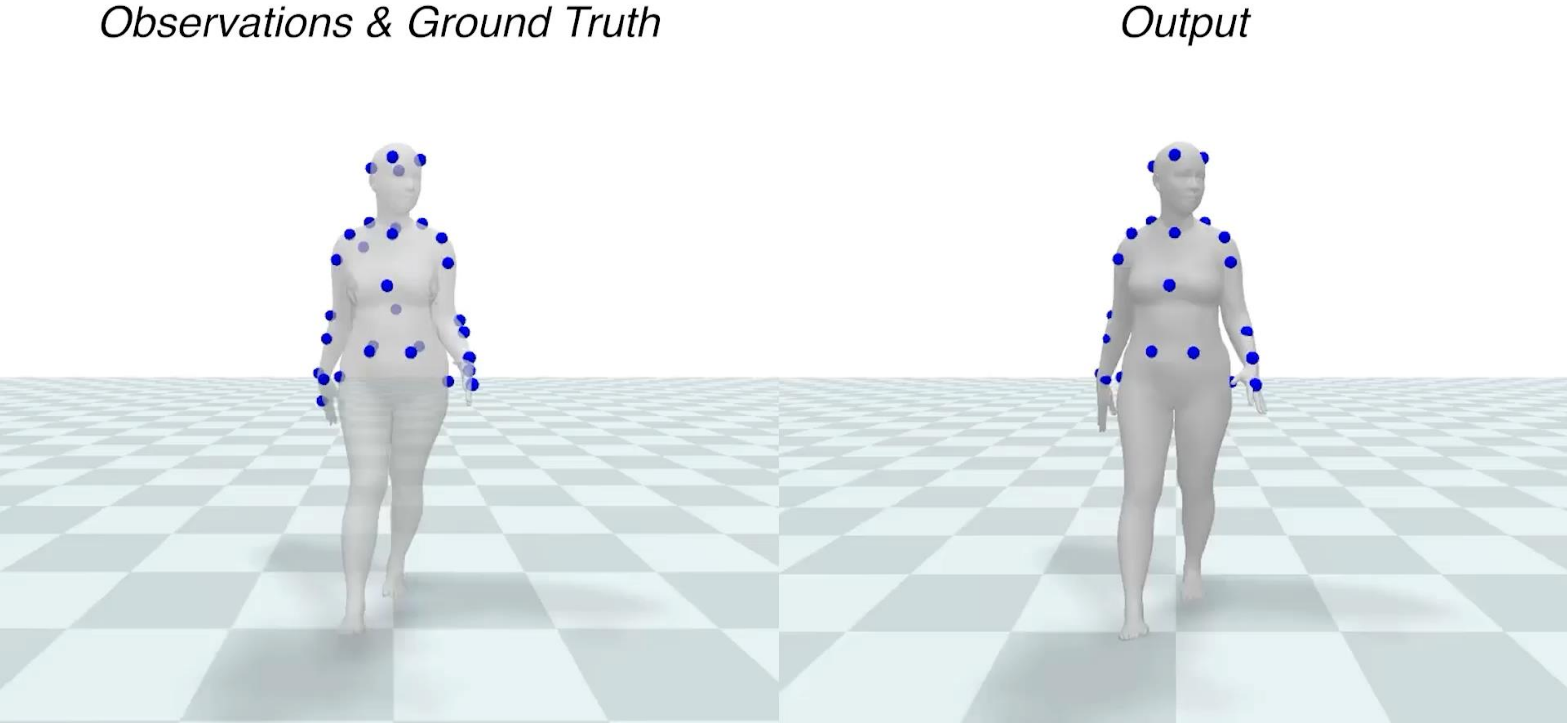

Unseen Body Shapes


Diverse Samples

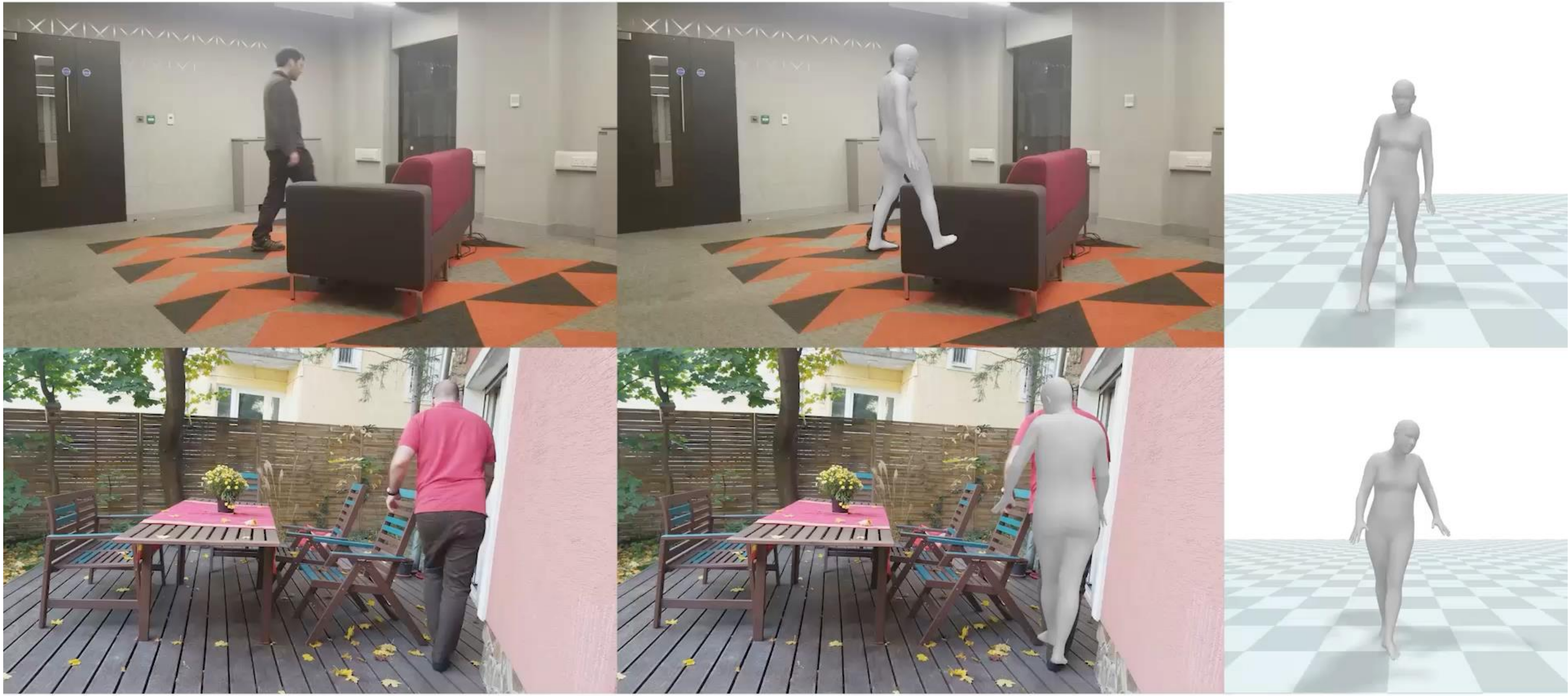
Test-time Optimization



Partial 3D Keypoints: **Sequence 1**



Fitting to 2D Joints (RGB)

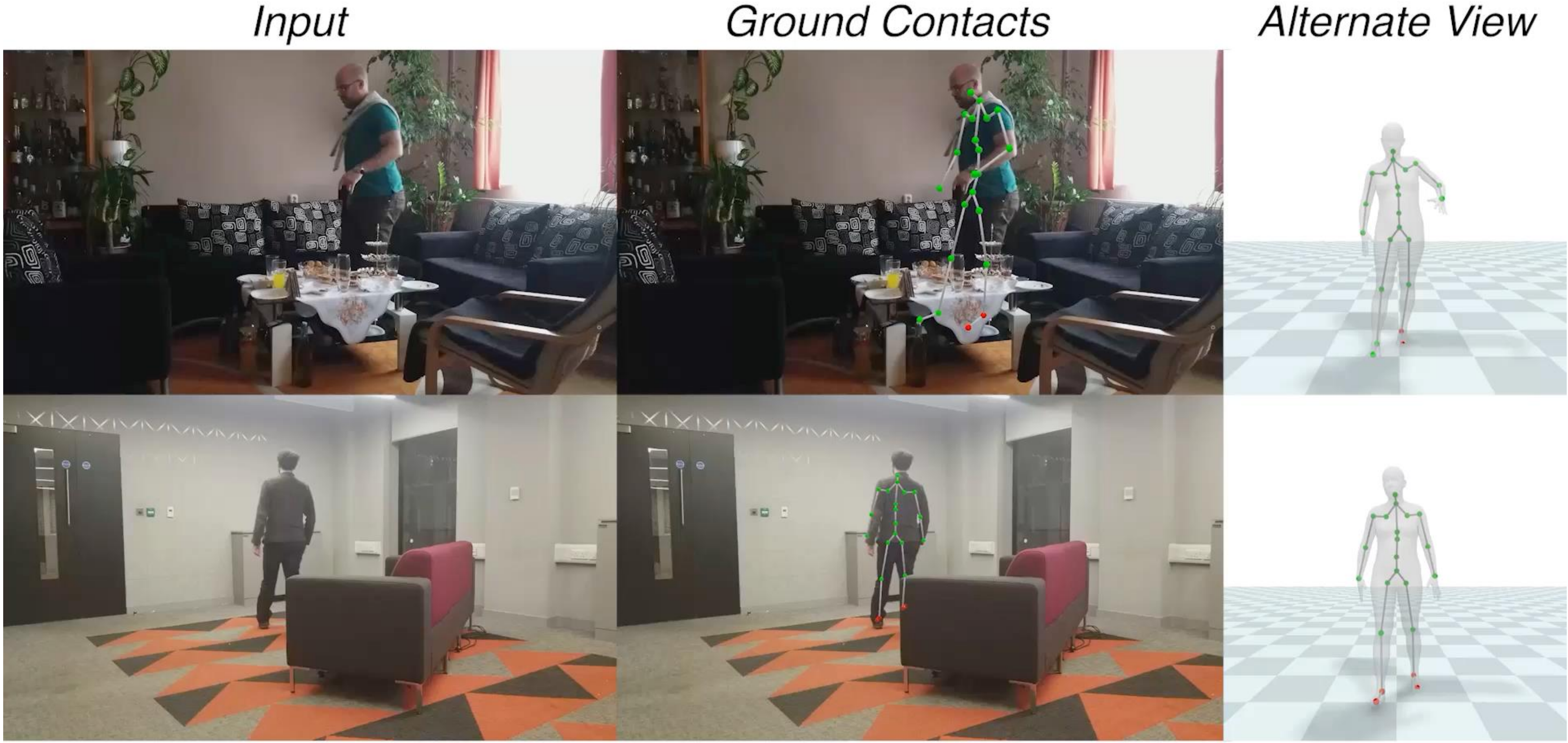


Input

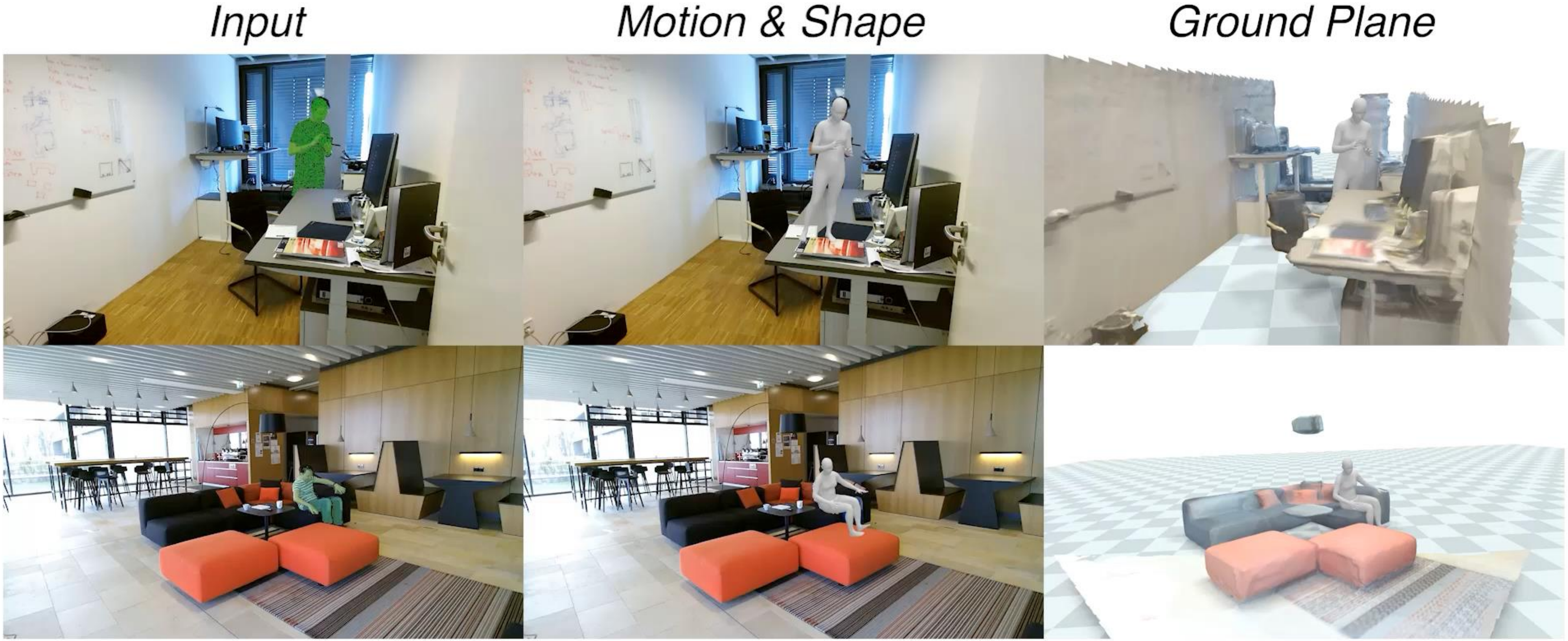
Motion & Shape

Alternate View

Fitting to 2D Joints (RGB)



Fitting to 2D Joints + 3D (RGB-D)

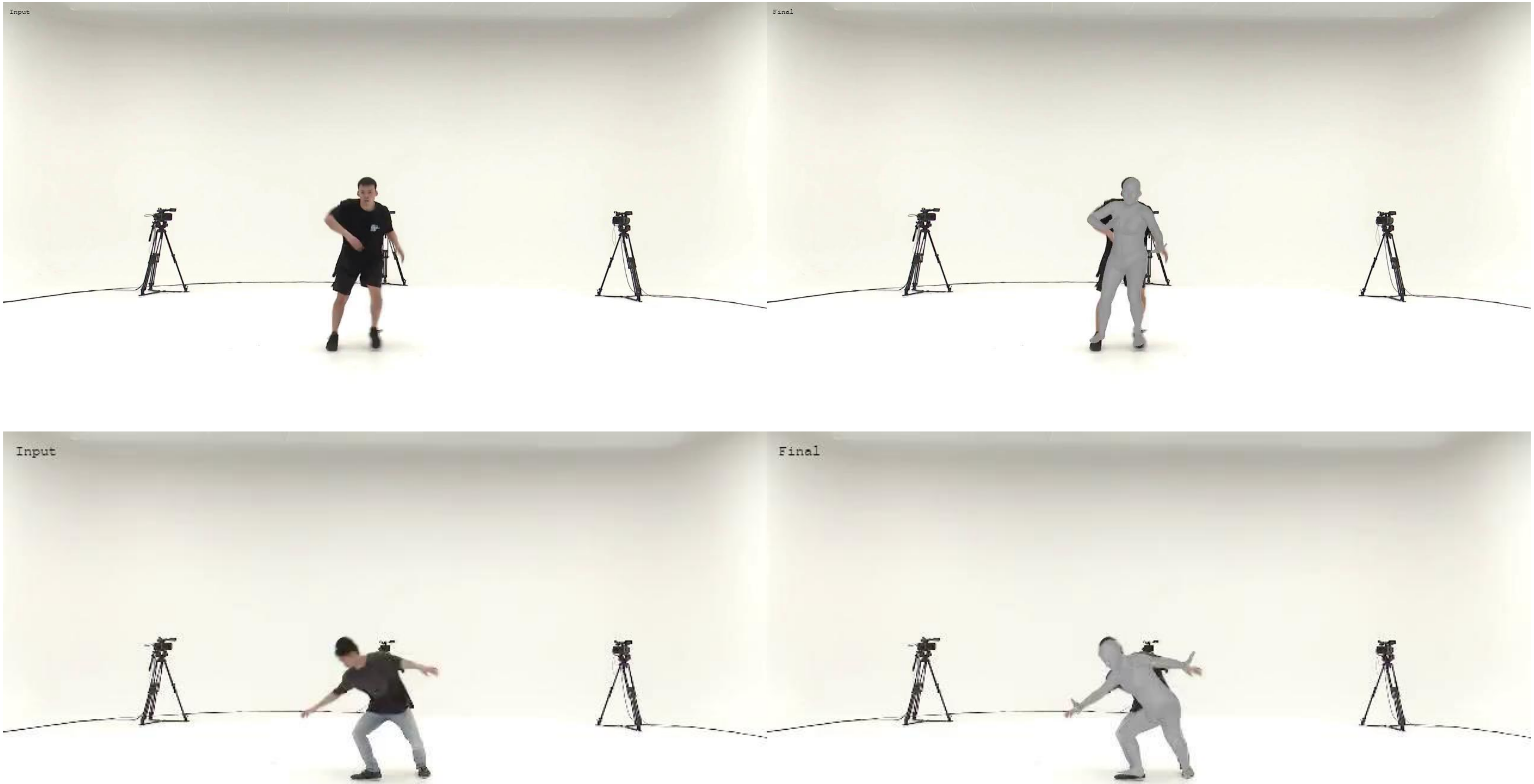


Input

Motion & Shape

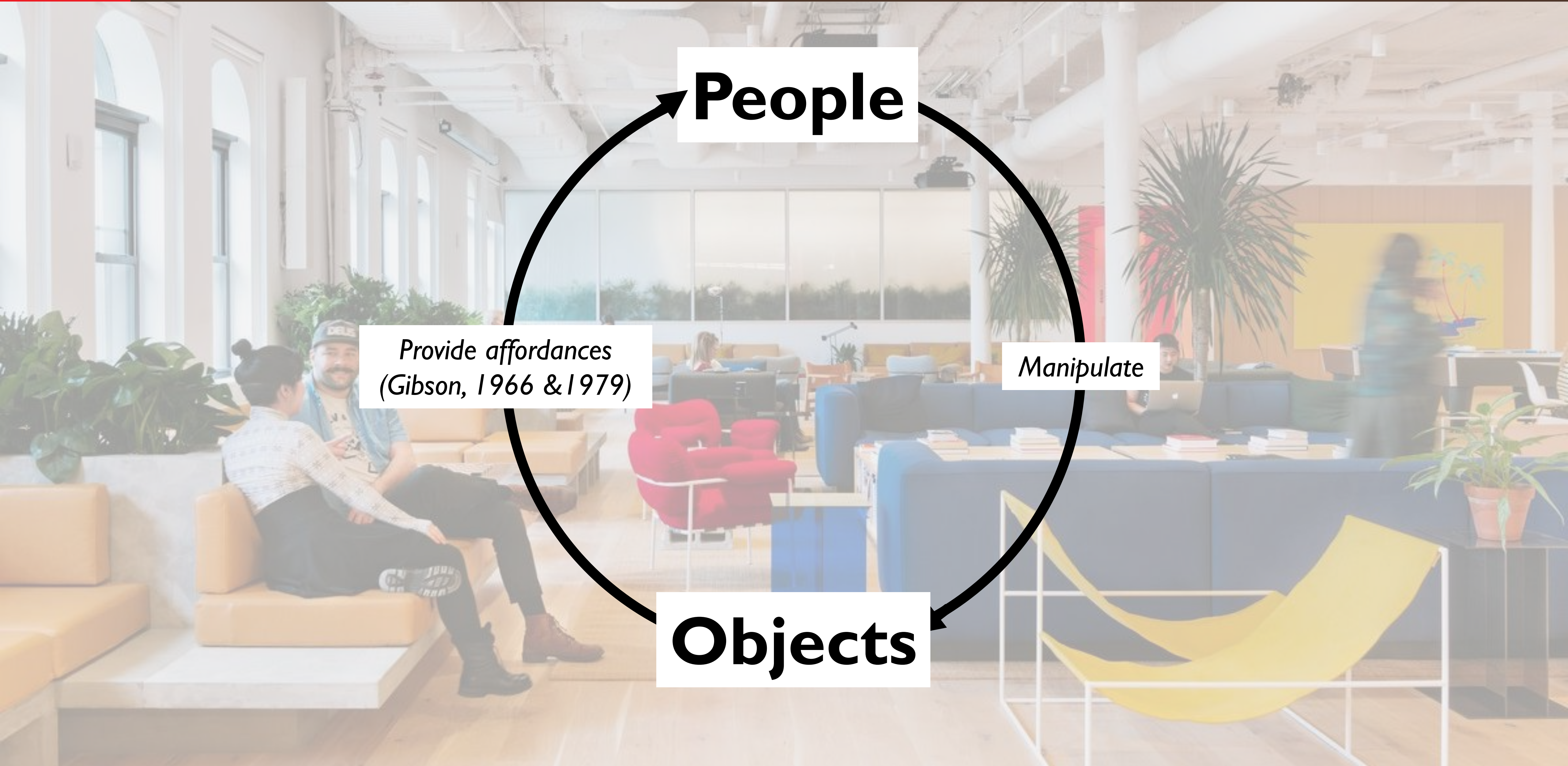
Ground Plane

Dynamic RGB Videos



Tsuchida et al., ISMIR 2019

Human Physical Intelligence



People

Manipulate

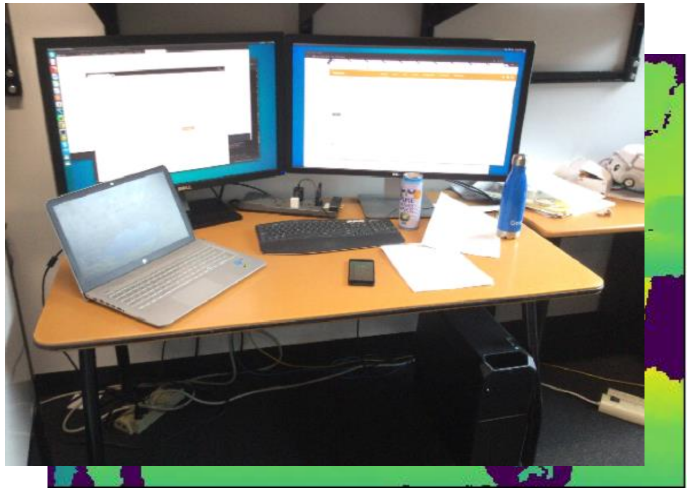
Objects

Provide affordances (Gibson, 1966 &1979)

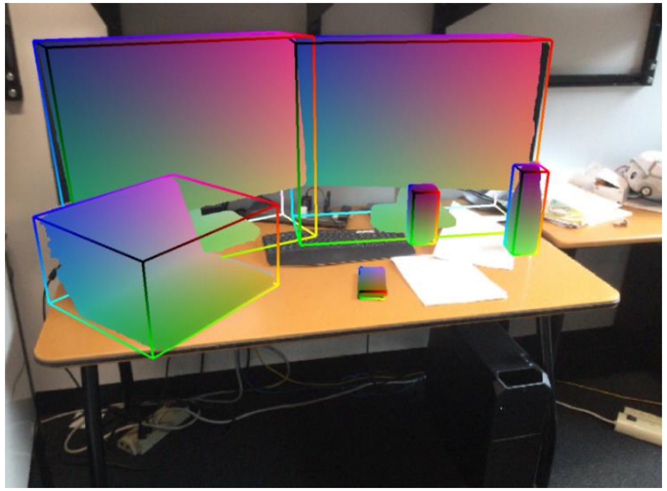
Understanding Objects

Perceive and generate object properties such as class, pose, shape, … for **novel object instances**

Related Work



RGB(-D)



- 6 DoF Pose
  - 3D position
  - 3D orientation
- Object dimensions

Limitations

- Limited category-level reasoning
- Datasets
- Generalizability

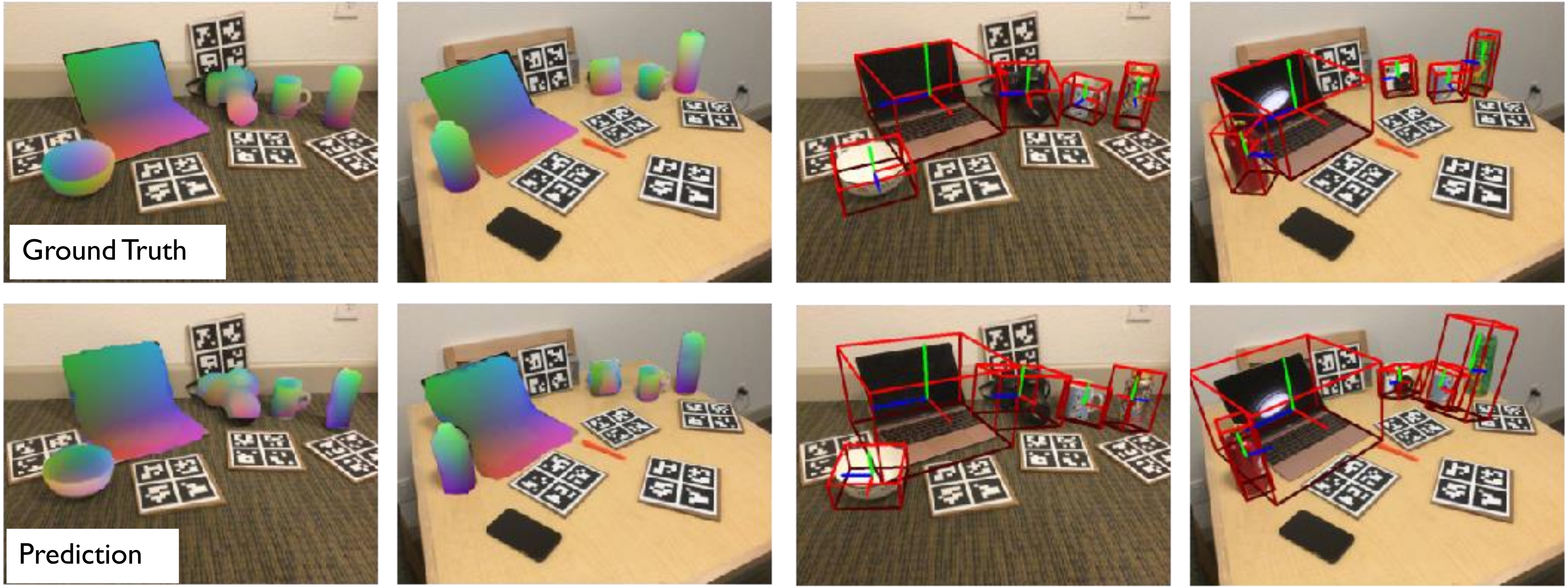
6 DoF Pose

- Brachmann et al. 2014
- Kehl et al. 2017
- Xiang et al. 2017
- Krull et al. 2016
- Doumanoglou et al. 2016 …

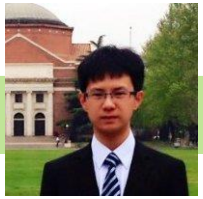
3D Object Detection

- Gupta et al. 2013, 2014
- Engelcke et al. 2017
- Song et al. 2016
- Qi et al. 2018
- Zhou et al. 2017 …

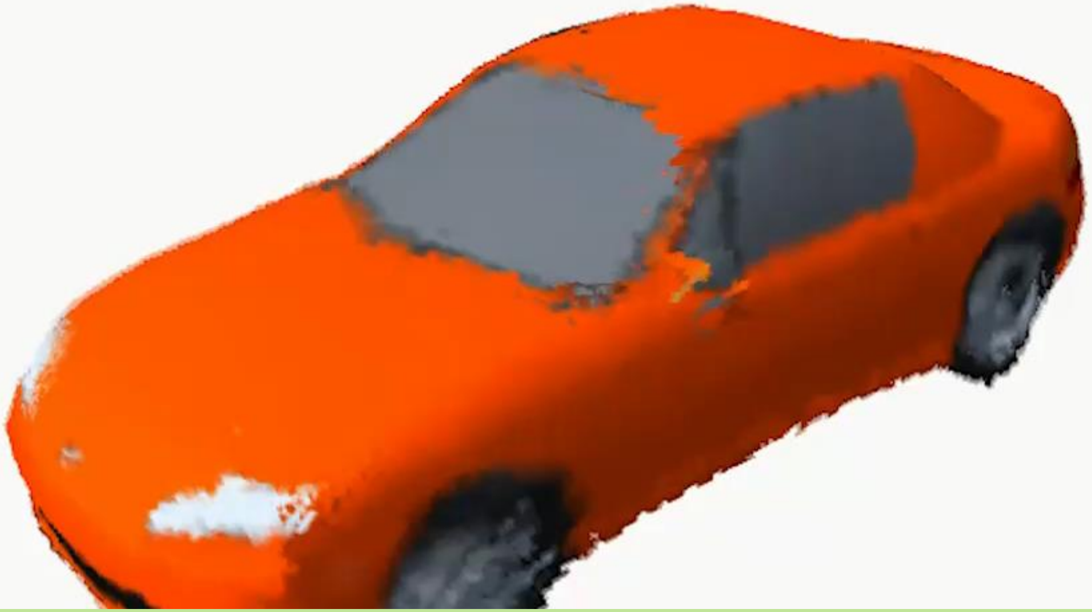
Perception: 6 DoF Pose



Wang, Sridhar, Huang, Valentin, Song, Guibas. **CVPR 2019 [Oral Presentation]**

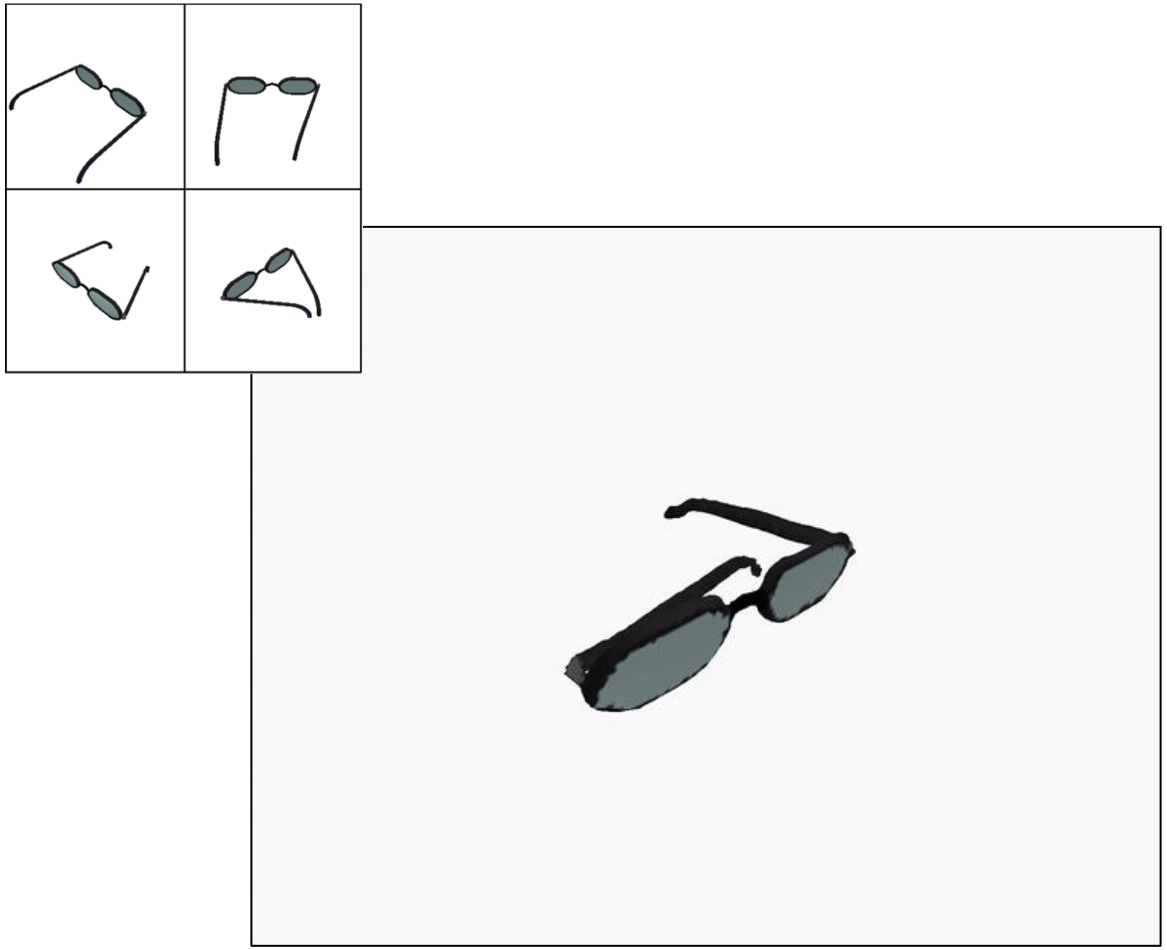


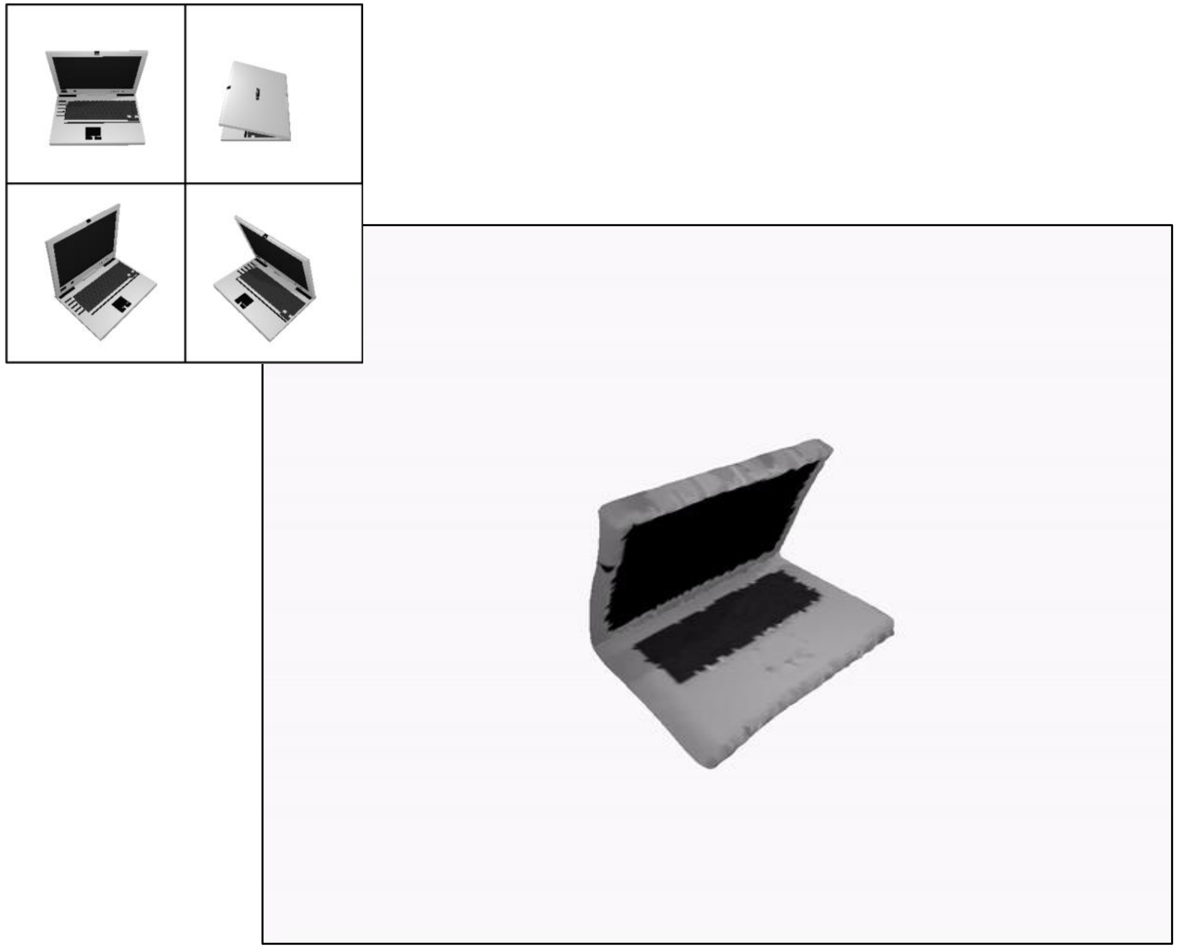
Perception: 3D Reconstruction



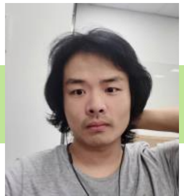
Pix2Surf. Lei, Sridhar, Guerrero, Sung, Mitra, Guibas. **ECCV 2020**

Perception: 3D Articulation





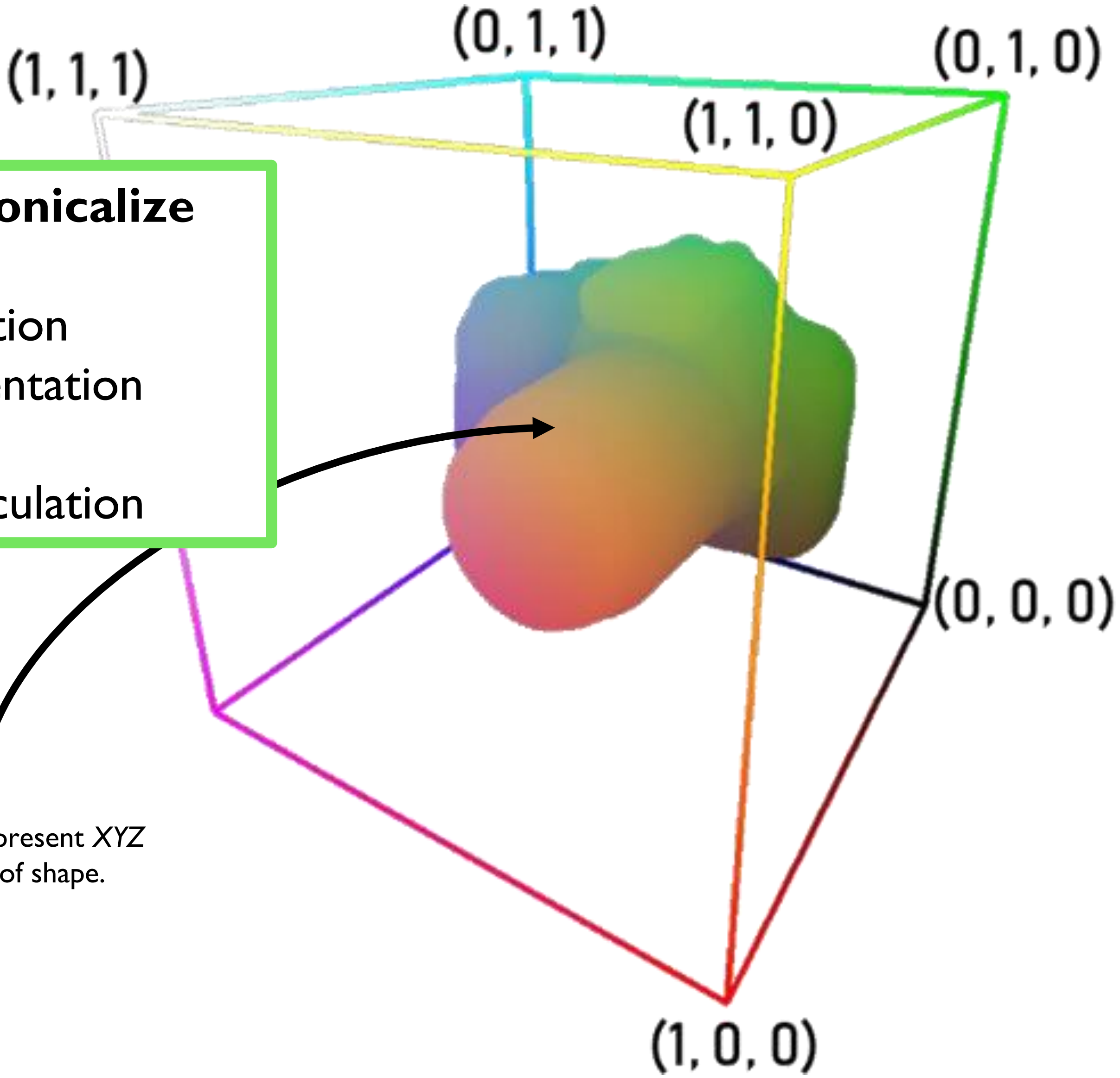
StrobeNet. Zhang, Litany, Sridhar, Guibas. **arXiv 2021**



NOCS: Normalized Object Coordinate Space

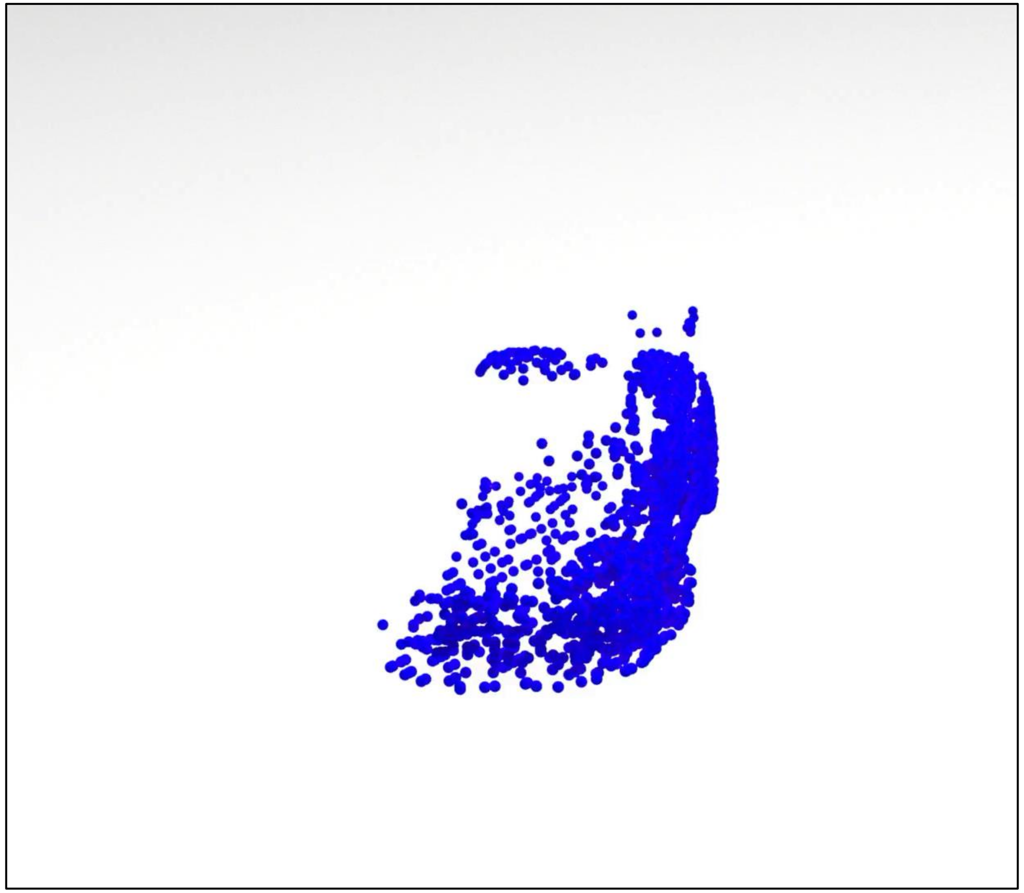
Canonicalize

- Position
- Orientation
- Size
- Articulation



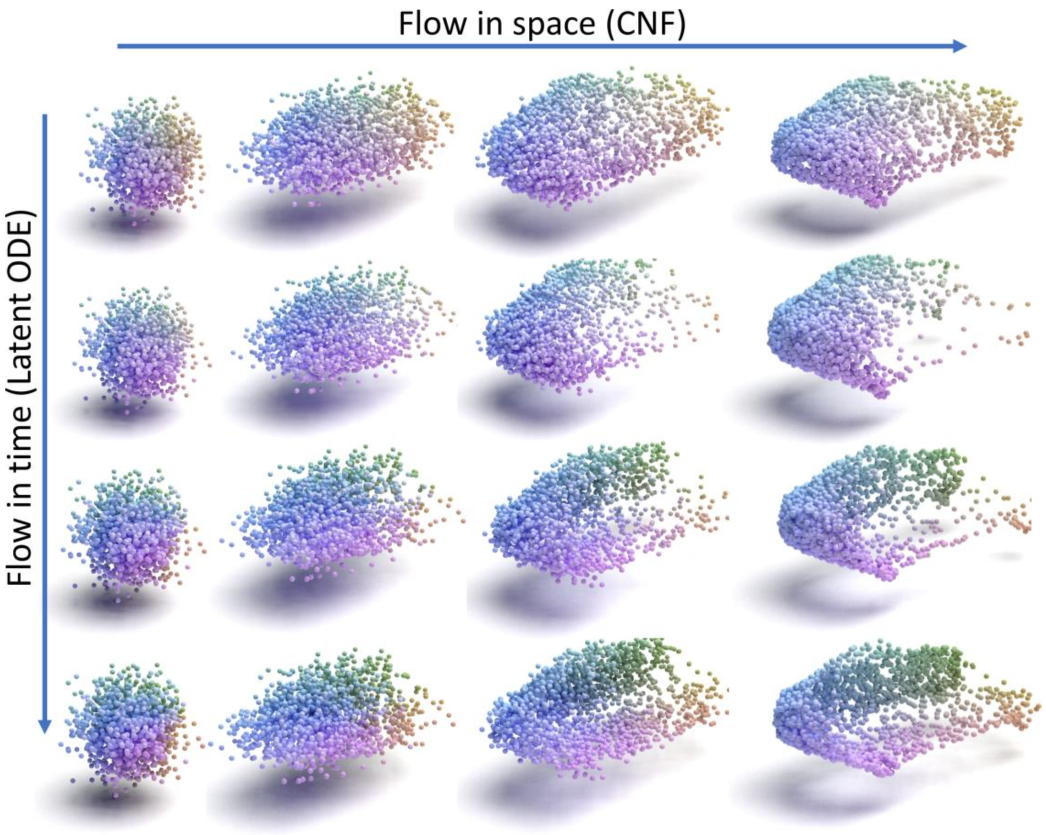
RGB colors represent *XYZ* coordinates of shape.

Sufficient for understanding?



Motion over time. How do we reconstruct?

Generation: Spatiotemporal Change



CaSPR: Learning Canonical Spatiotemporal Point Cloud Representations

Davis Rempe, Tolga Birdal, Yongheng Zhao, Zan Gojcic, Srinath Sridhar, Leonidas Guibas

NeurIPS 2020 (Spotlight)

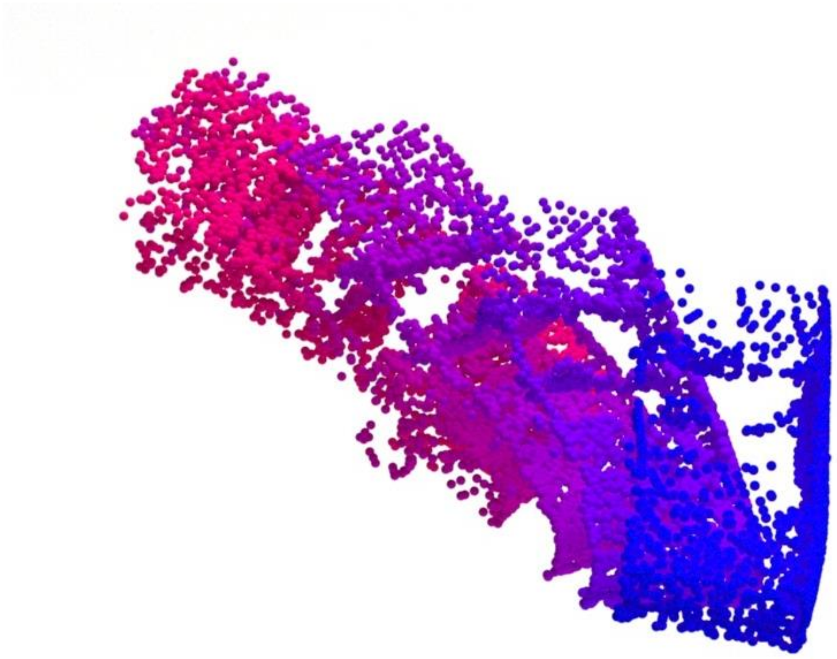
geometry.stanford.edu/projects/caspr/

CaSPR. Rempe, Birdal, Zhao, Gojcic, Sridhar, Guibas. **NeurIPS 2020 [spotlight presentation]**

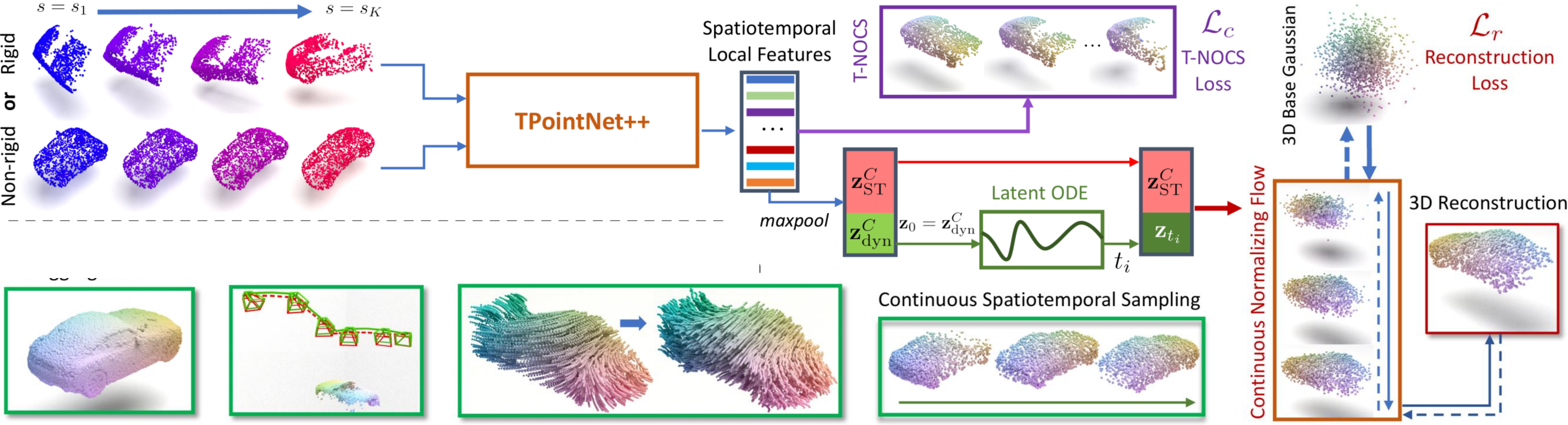
Representing Dynamic Point Clouds

<u>Goal</u>:

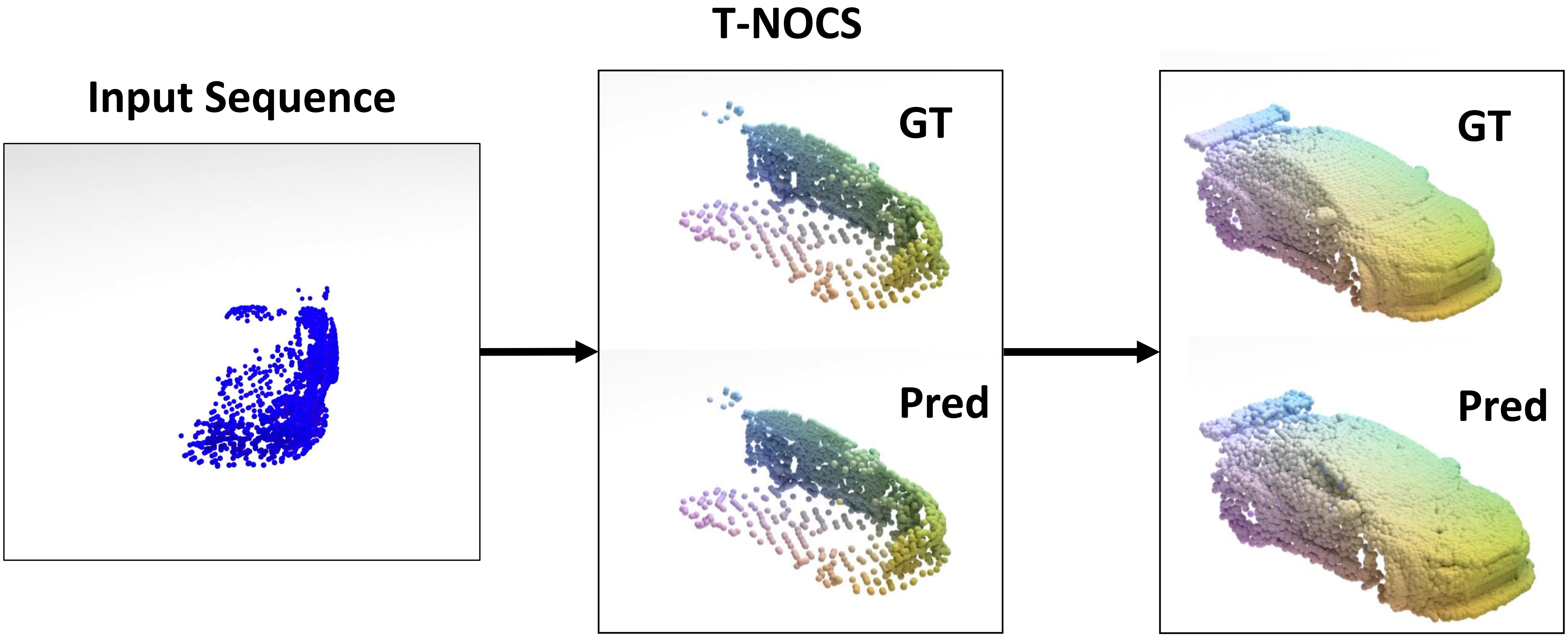
Learn a **representation** from point cloud inputs that can capture and generate **spatio-temporal** changes in **object properties**.



CaSPR Architecture



Canonicalization Results



Input Sequence

T-NOCS

GT

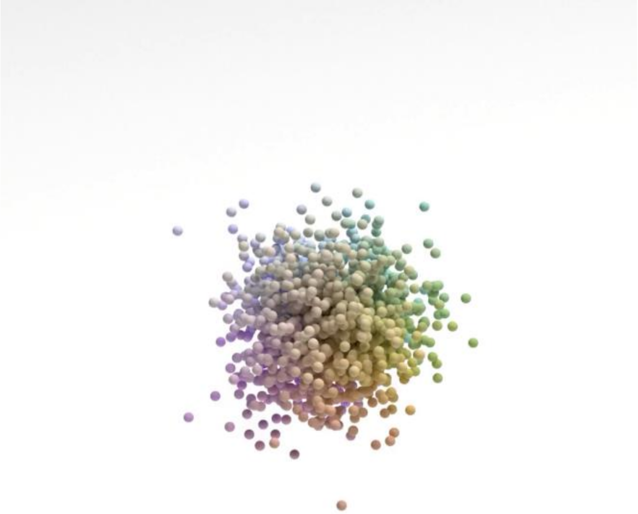
Pred

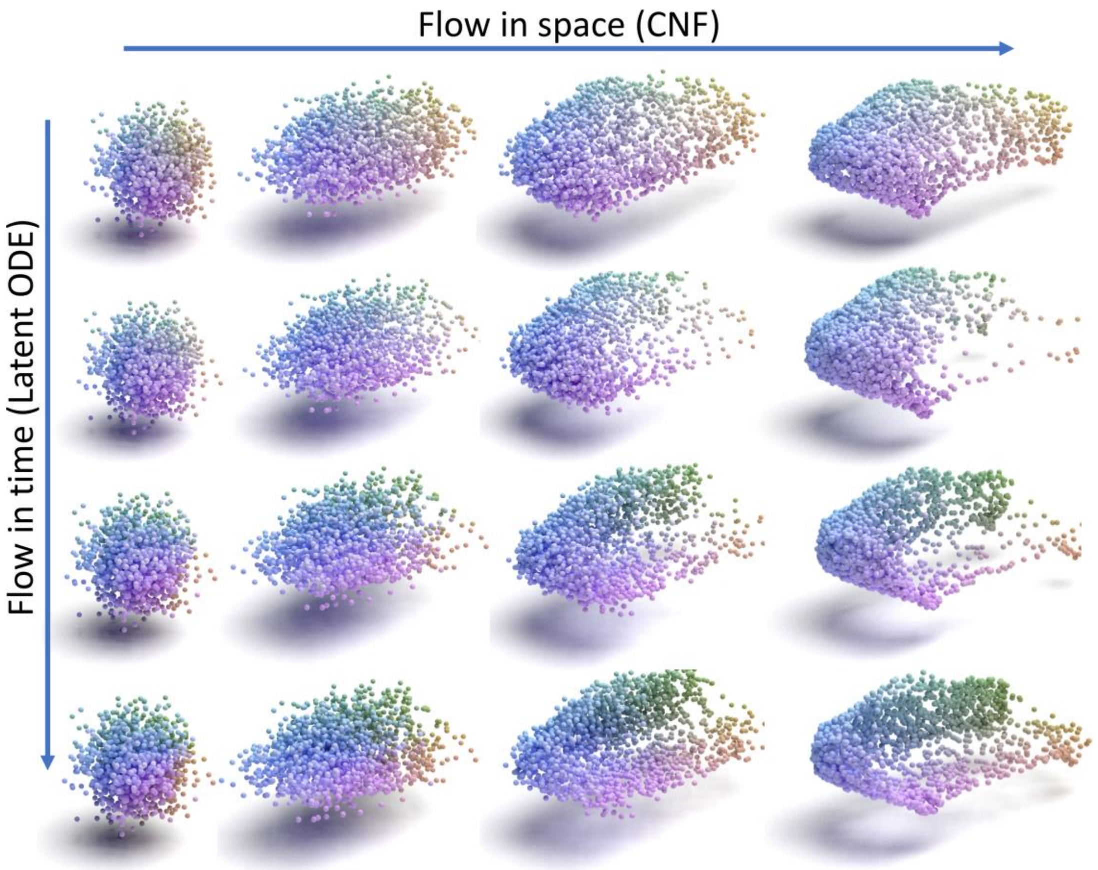
GT

Pred

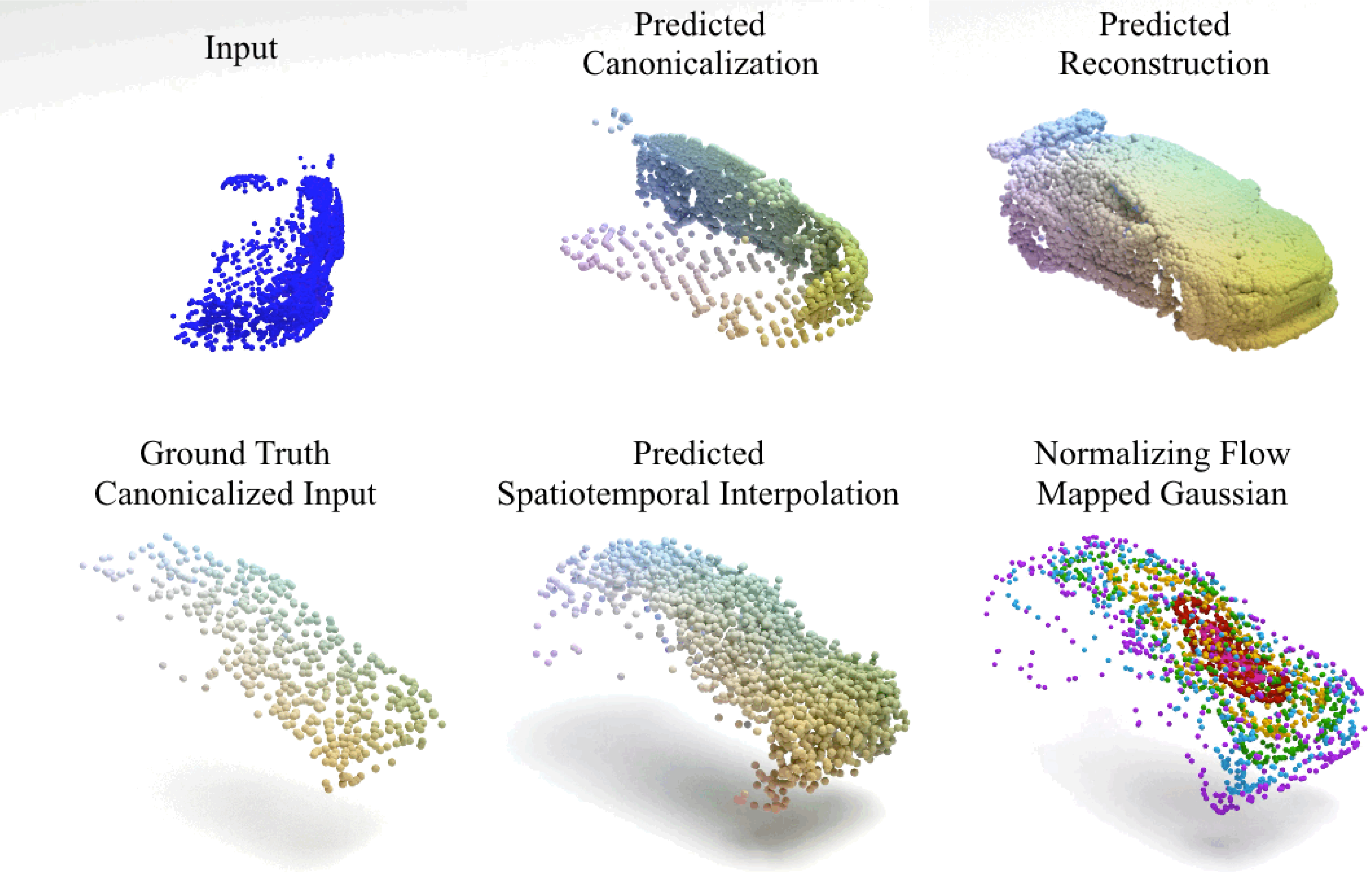
Generation

- Gives **spatiotemporal continuity**
- **Latent ODE** allows "querying" any intermediate timestamp
- Continuous Normalizing Flow (CNF) allows dense spatial sampling





Canonicalization & Generation



Input

Predicted Canonicalization

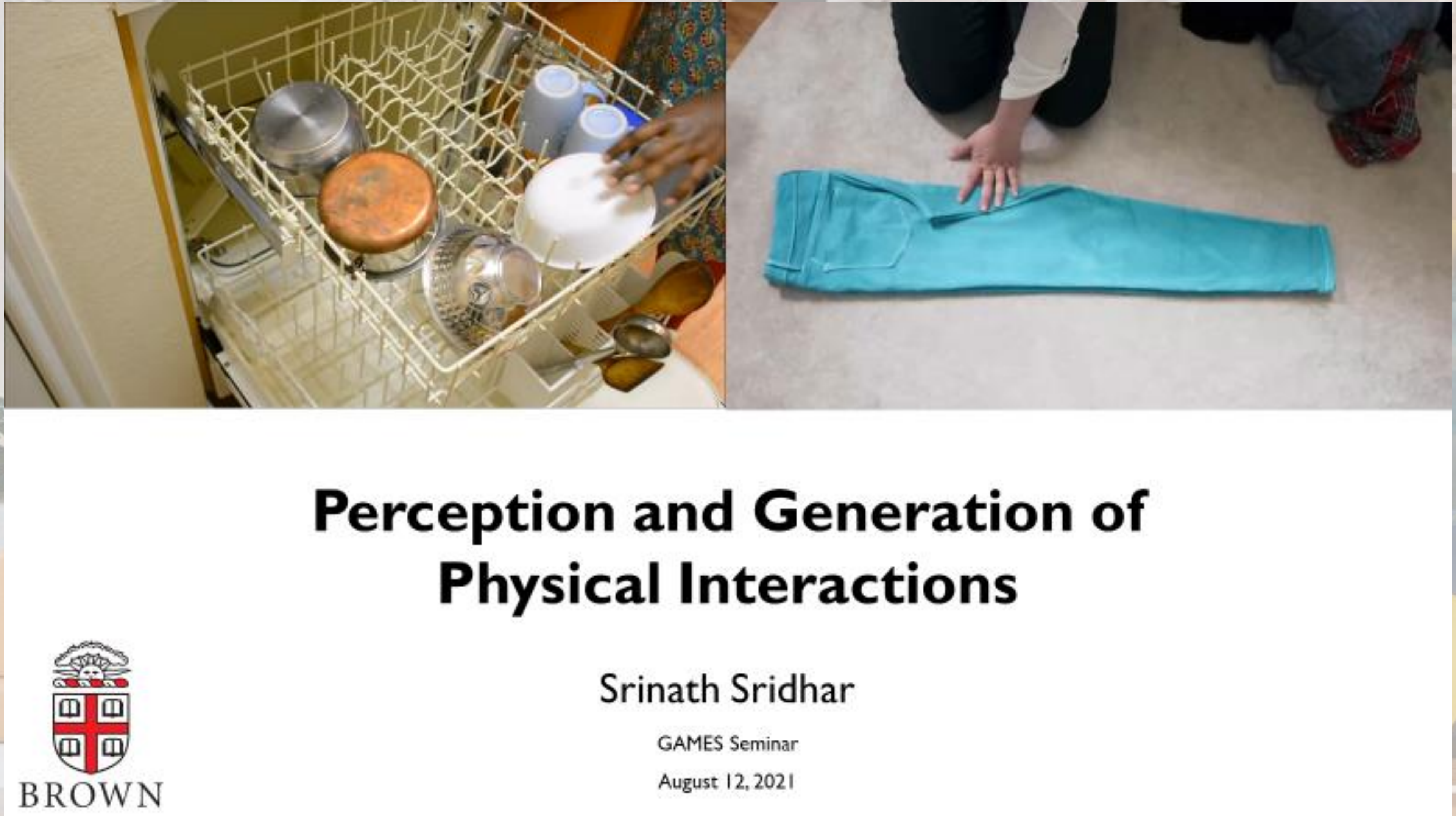
Predicted Reconstruction

Ground Truth Canonicalized Input

Predicted Spatiotemporal Interpolation

Normalizing Flow Mapped Gaussian

Human Physical Intelligence



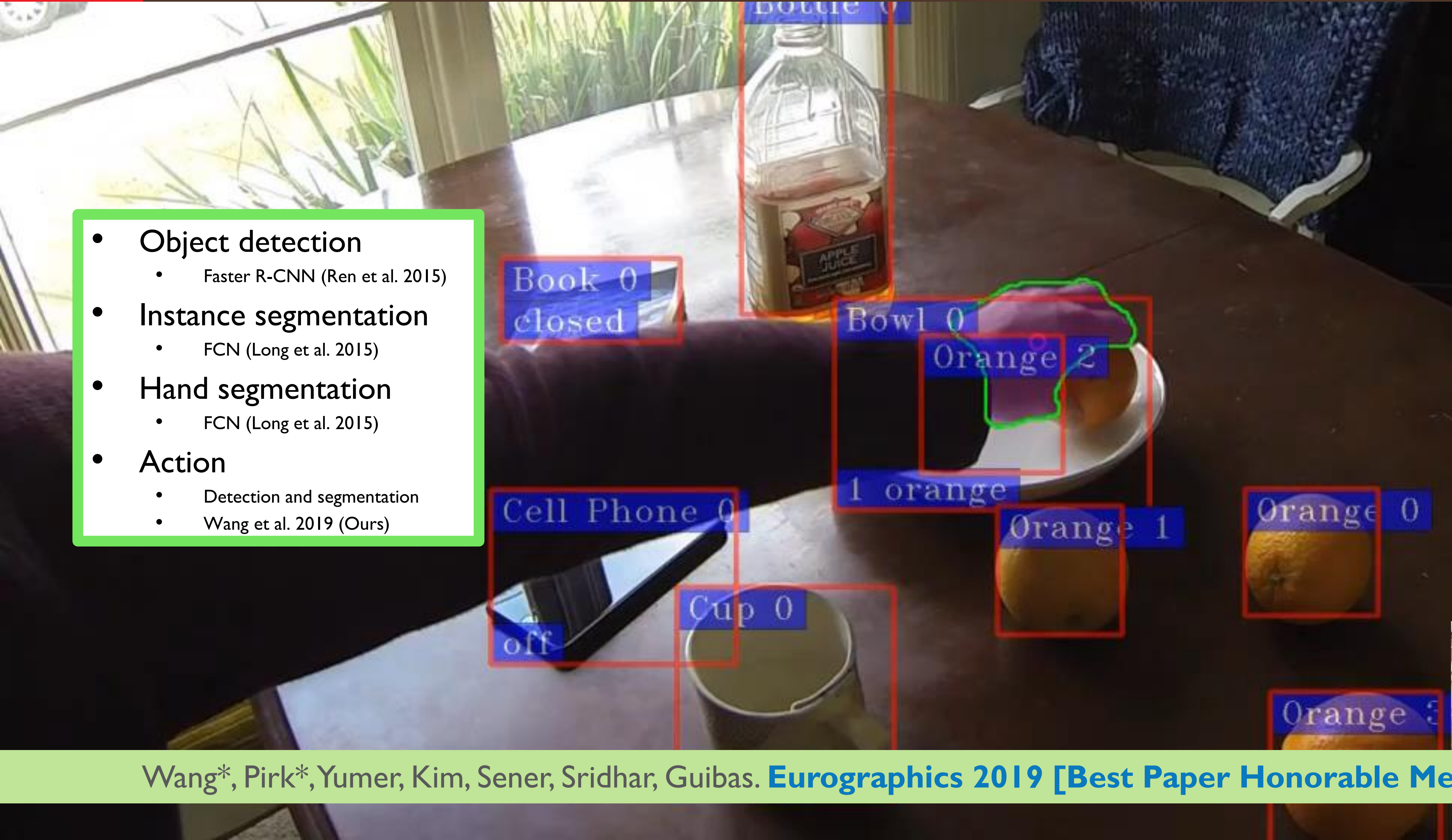
Perception and Generation of Physical Interactions

Srinath Sridhar

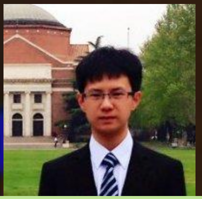
GAMES Seminar

August 12, 2021

Perception: 2D Human-Object Interactions

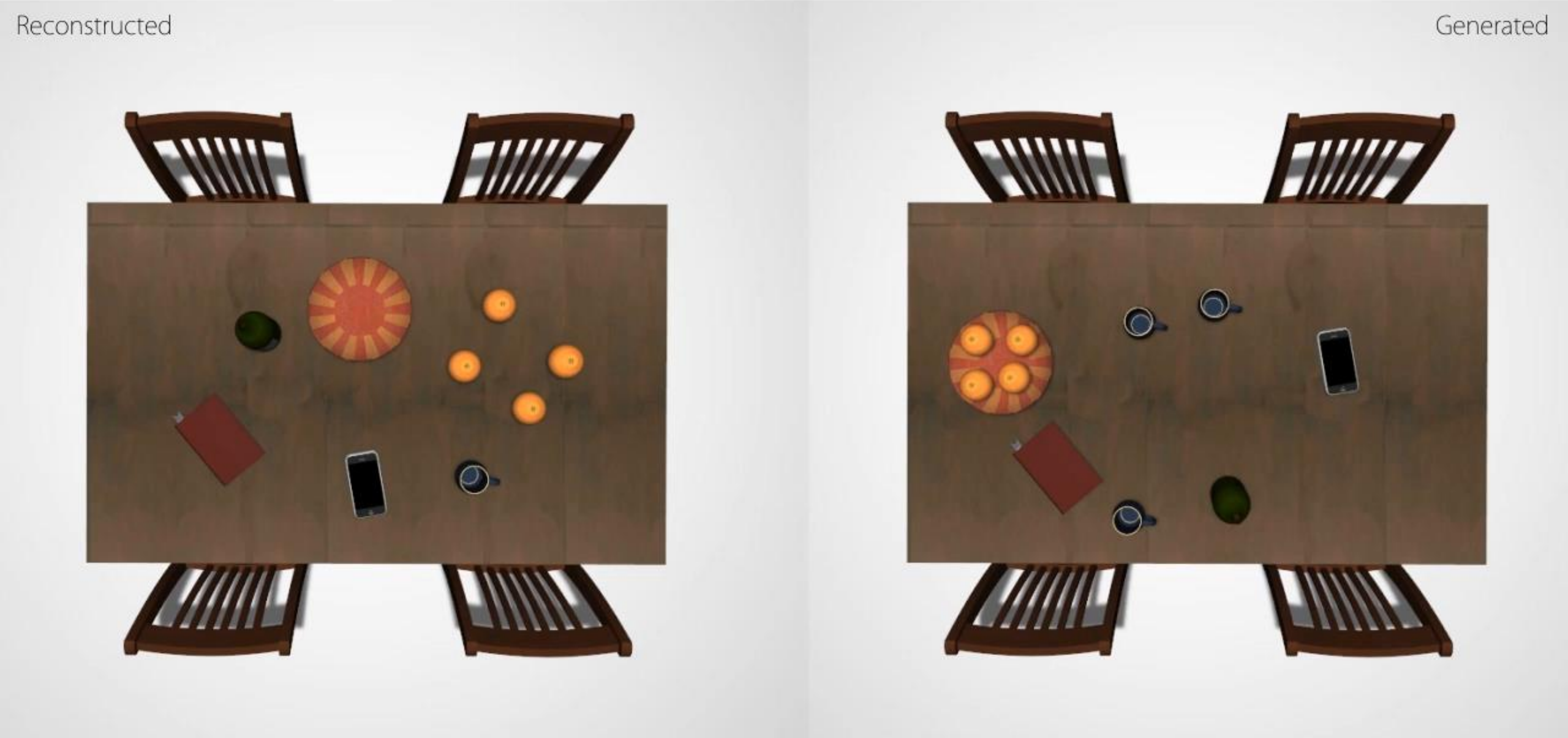


- Object detection
  - Faster R-CNN (Ren et al. 2015)
- Instance segmentation
  - FCN (Long et al. 2015)
- Hand segmentation
  - FCN (Long et al. 2015)
- Action
  - Detection and segmentation
  - Wang et al. 2019 (Ours)



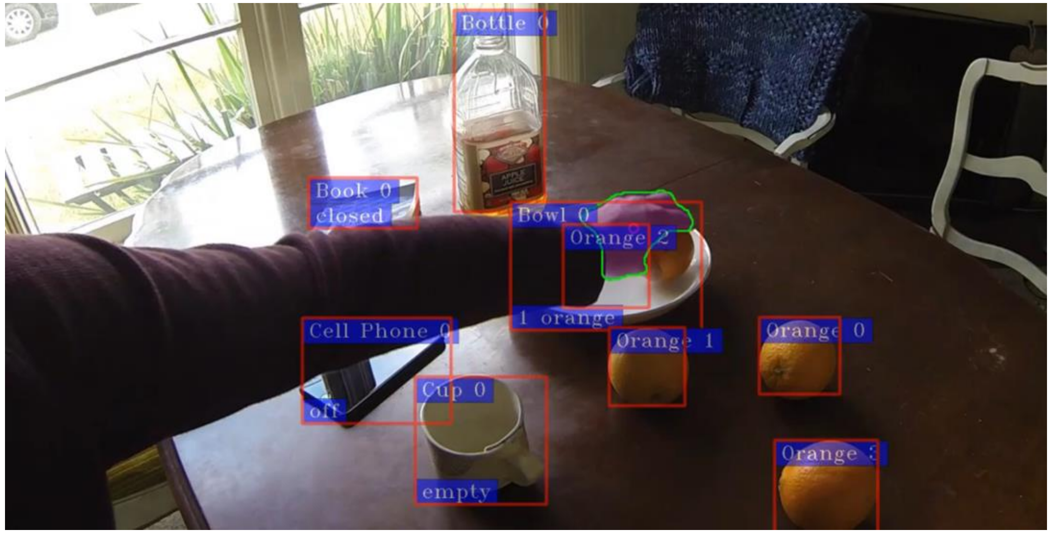
Wang*, Pirk*, Yumer, Kim, Sener, Sridhar, Guibas. **Eurographics 2019 [Best Paper Honorable Mention]**

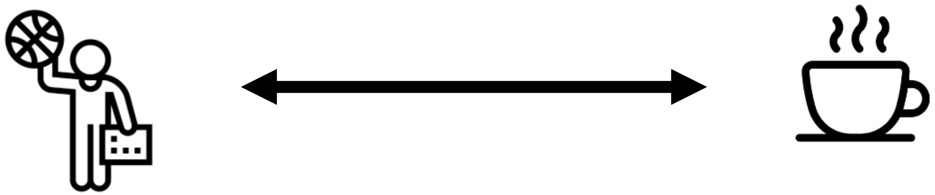
Generation: Human-Object Interactions



Future Work

Perception: “what is”





learn to digitize

Requires

- 3D Understanding
- Deeper knowledge of human skills
- Expressive models of objects
- Properties other than shape & appearance

Leo Guibas

Christian Theobalt

Antti Oulasvirta

Hans-Peter Seidel

Niloy Mitra

Karlin Bark / Lee Beckwith / Florian Bernard / Rishabh Bhandari / Sebastian Boring / Sofien Bouaziz / Dan Casas / Anna Maria Feit / Paul Guerrero / Aaron Hertzmann / Judy Hoffman / Jingwei Huang / Krishna Murthy Jatavallabhula / Vladimir Kim / K. Madhava Krishna / Jiahui Lei / Or Litany / Anders Markussen / Dushyant Mehta / Ari Morcos / Franziska Mueller / Victor Ng-Thow-Hing / Soeren Pirk / Gerard Pons-Moll / Davis Rempe / Helge Rhodin / Ozan Sener / Mohammad Shafiei / Rahul Sajnani / AadilMehdi Sanchawala / Shuran Song / Oleksandr Sotnychenko / Minhyuk Sung / Cuong Tran / Julien Valentin / He Wang / Weipeng Xu / Jimei Yang / Zhangsihao Yang / Ersin Yumer / Michael Zollhöfer

BROWN

Stanford

mpi max planck institut informatik

NSF

AIR FORCE RESEARCH LABORATORY

TOYOTA RESEARCH INSTITUTE

aws

srinathsridhar.com

Brown IVL Group

Brown
Visual Computing

PhD Students

Kefan Chen
Rao Fu

MS Students

Sijie Ding
Trevor Houchens
Aparna Natarajan

Undergrad

Qiuhong (Anna) Wei
Yiheng Xie

Visitors

Radhika Dua
Shivam Duggal

Jivitesh Jain
Rahul Sajnani
Apoorve Singhal

Always looking for motivated students!

Topics:

3D computer vision, deep learning, robotics, graphics, …

srinath@brown.edu

@*drsrinathsridha*

ivl.cs.brown.edu