
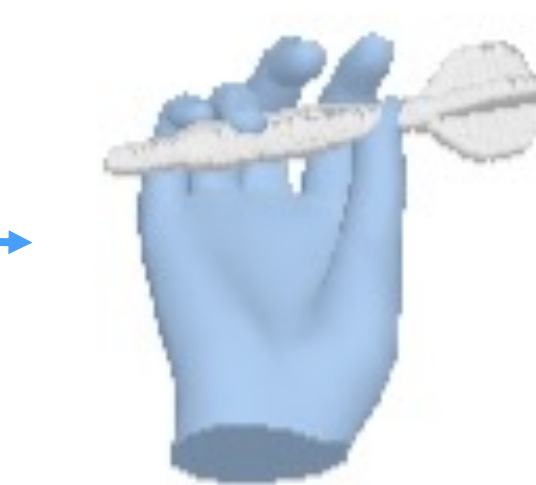





Introduction

Task:  **Model** 

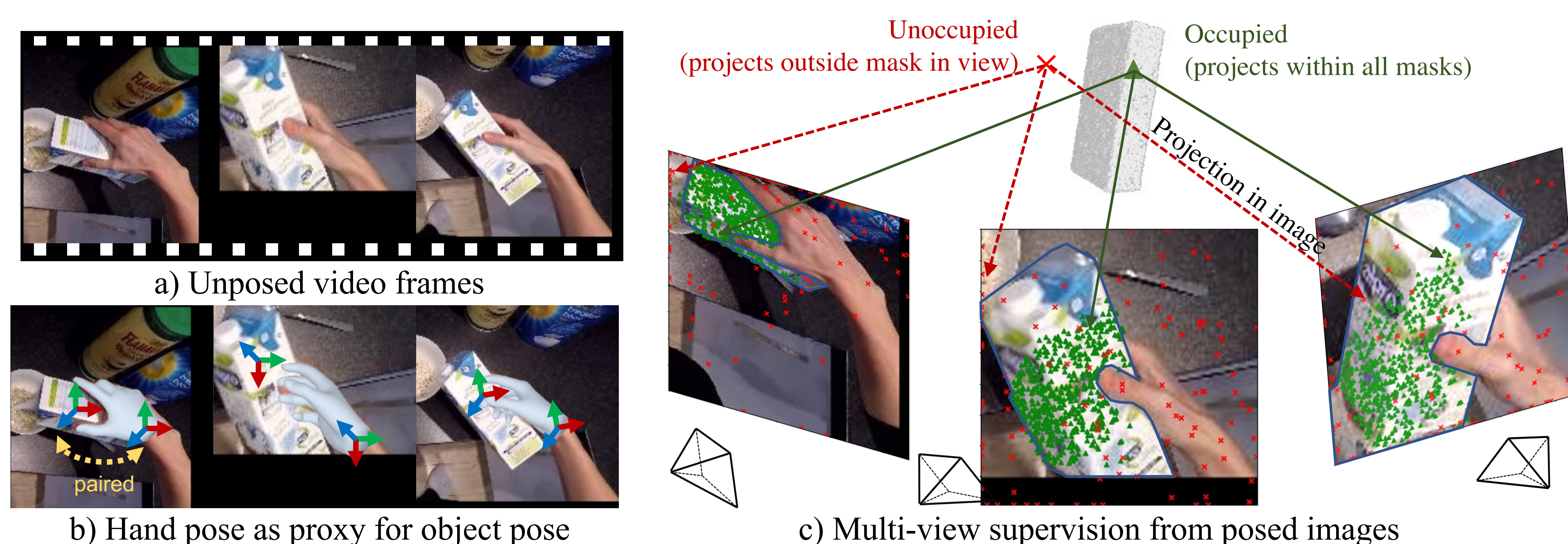
Single image **3D Object Mesh**



Synthetic: ObMan (2.5k objects) Lab setting: HO3D (10 YCB objects) In-the-wild: MOW (121 objects, 512 images)

Approach

2D masks guided 3D sampling – Multiview mask supervision

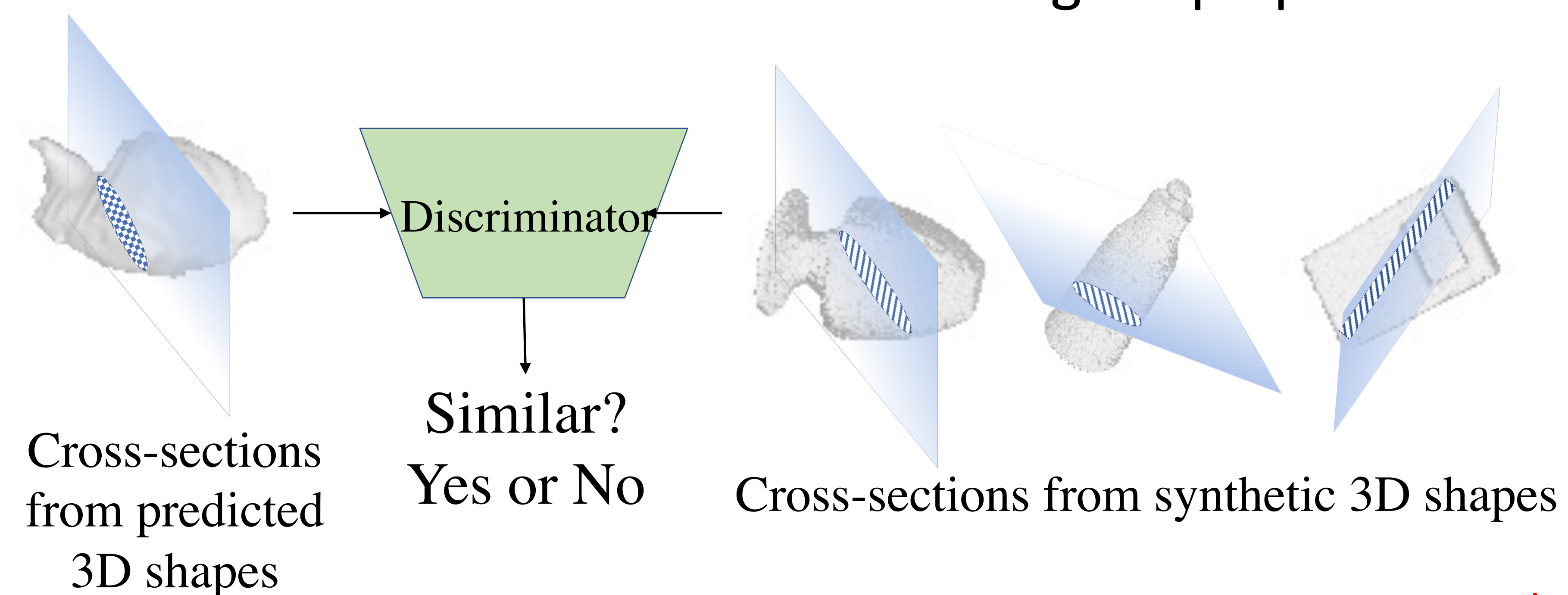


a) Unposed video frames

b) Hand pose as proxy for object pose

c) Multi-view supervision from posed images

2D slide based 3D discriminator – Learning shape priors

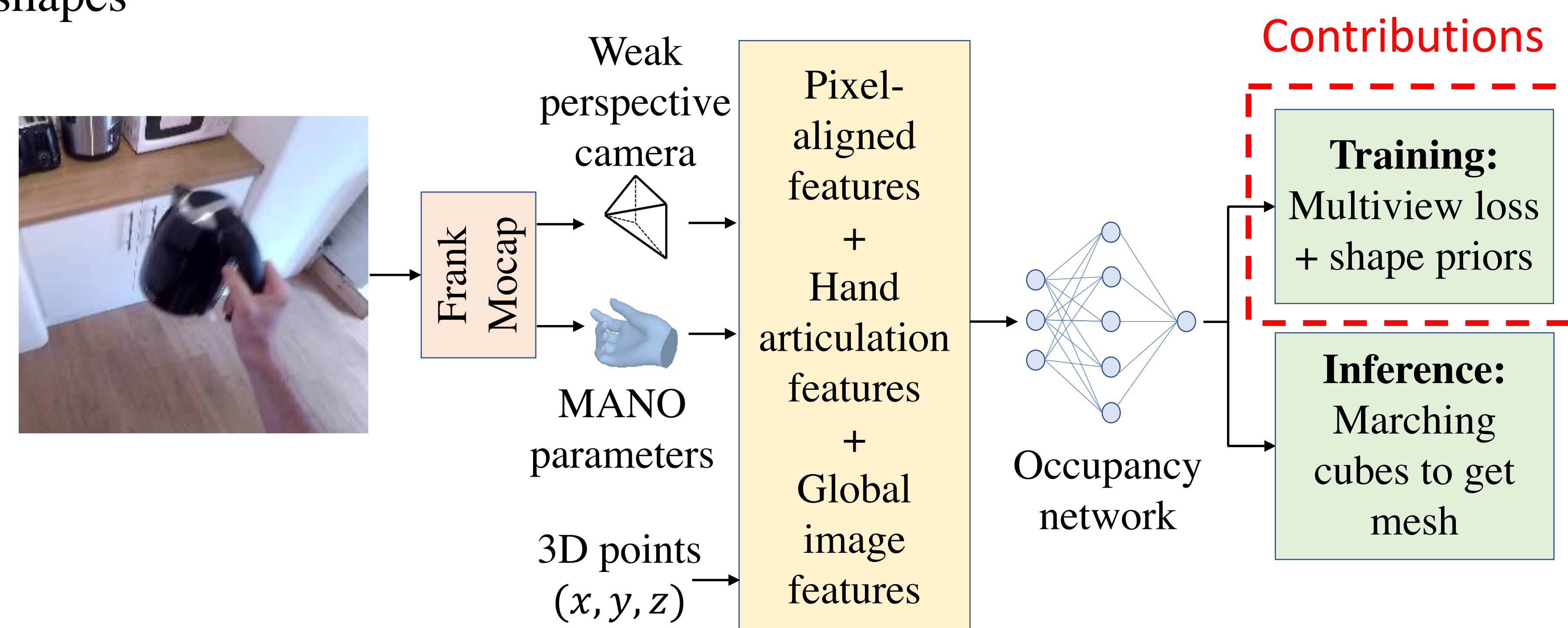


Discriminator

Similar? Yes or No

Cross-sections from predicted 3D shapes Cross-sections from synthetic 3D shapes

Pipeline:



Weak perspective camera

FrankMocap

MANO parameters

3D points (x, y, z)

Pixel-aligned features + Hand articulation features + Global image features

Occupancy network

Contributions

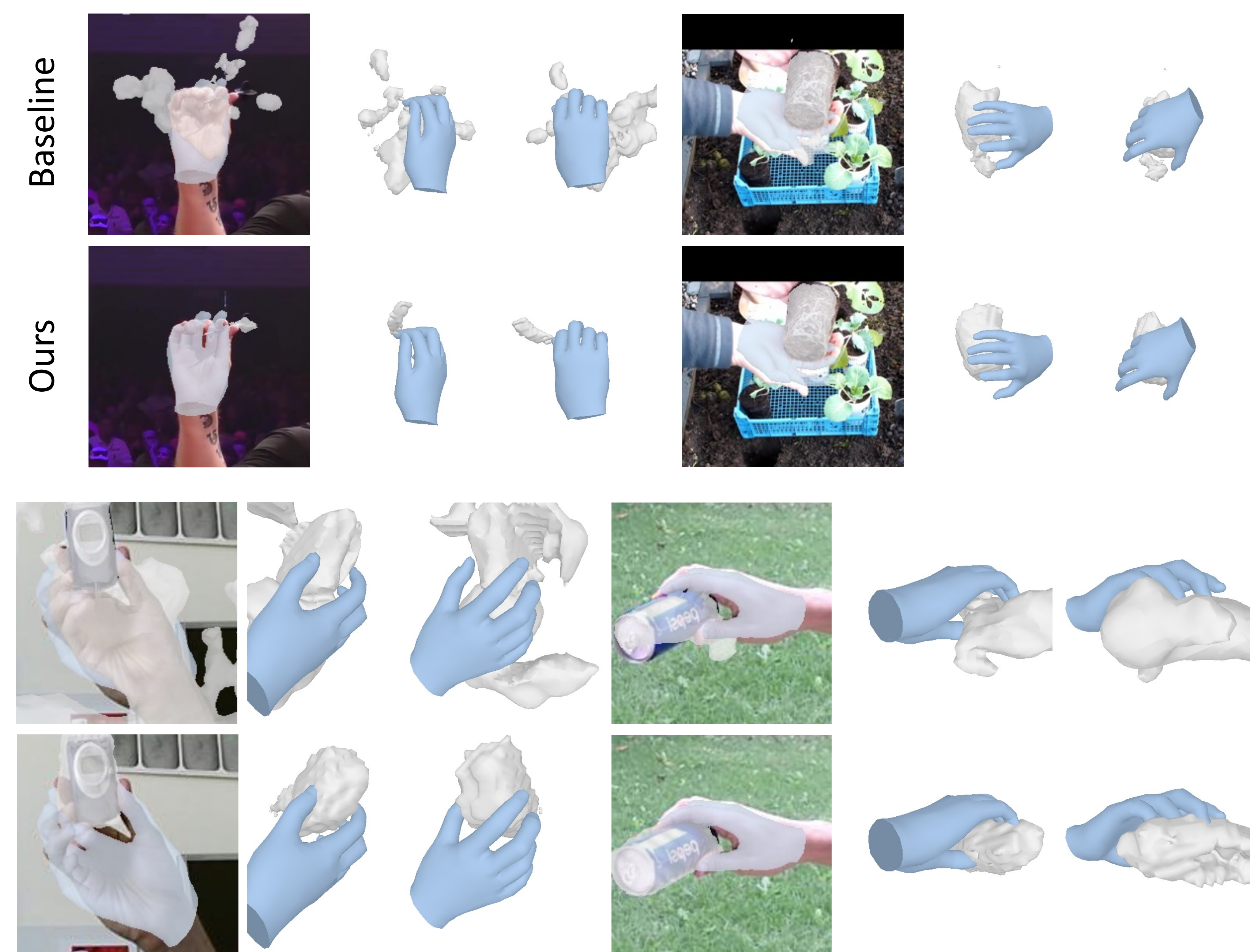
Training: Multiview loss + shape priors

Inference: Marching cubes to get mesh

Results

Method	Dataset & supervision: ObMan (Synthetic) +	F@5 ↑	F@10 ↑
AC-SDF	-	0.10	0.19
AC-SDF	+ HO3D (3D)	0.08	0.15
AC-SDF	+ HO3D (3D) + HOI4D (3D)	0.09	0.19
Ours	+ VISOR (2D masks) + Synthetic Shape priors	0.12 (+11%)	0.22 (+11%)

Existing models overfit to few object categories

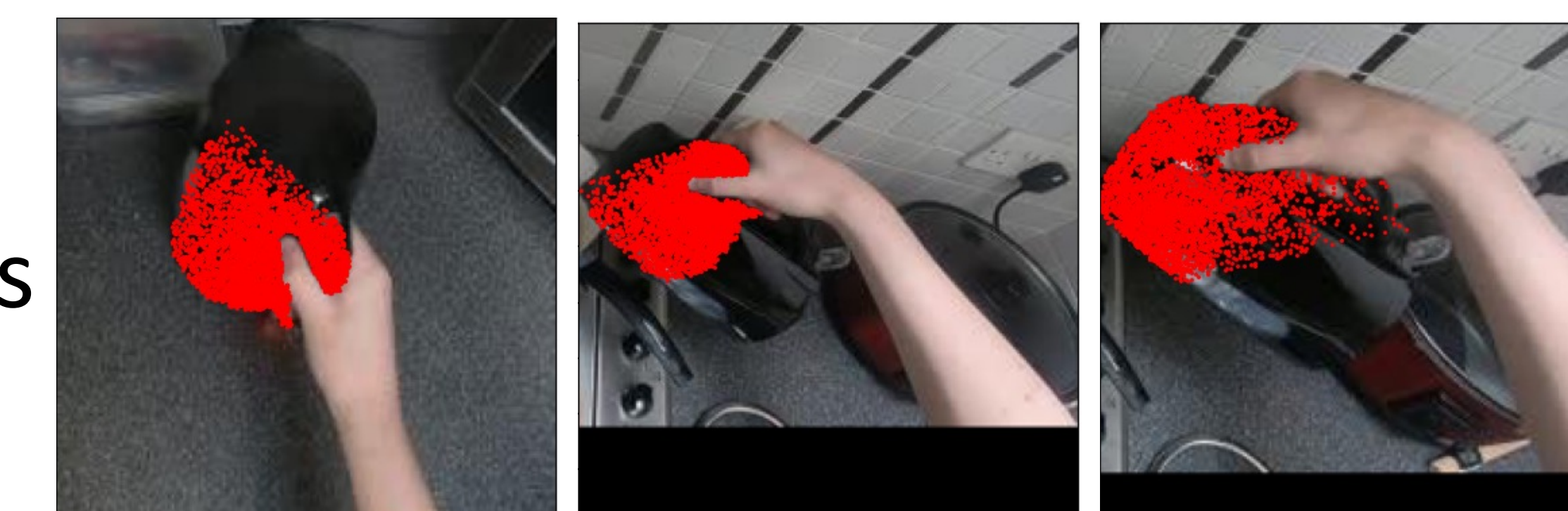


Baseline

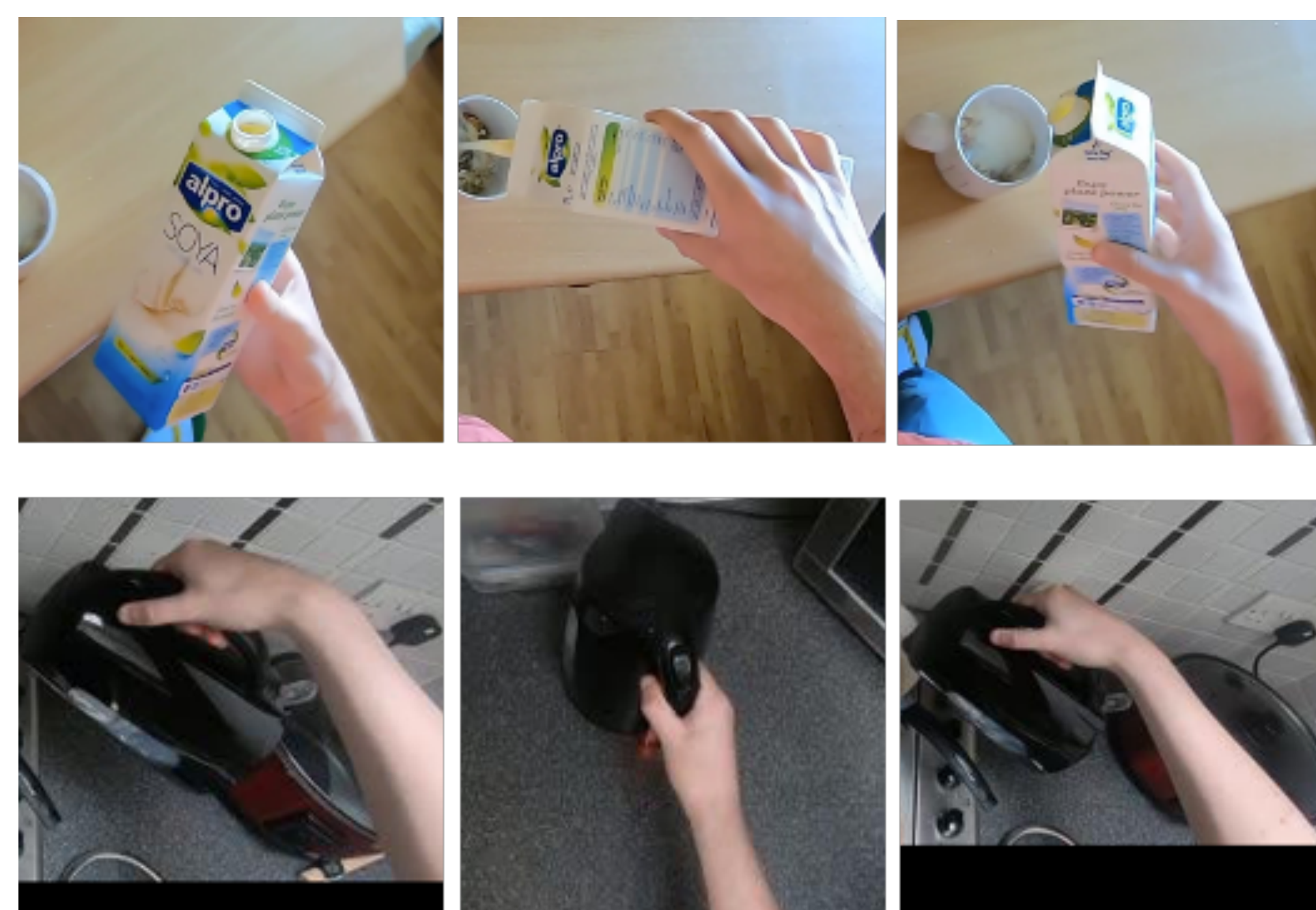
Ours

Limitations:

- Inaccurate hands
- Limited views



Diverse incidental multi-view sequences from EPIC



VISOR masks, FrankMocap for hand pose

