A Dataset for Dexterous Bimanual Hand-Object Manipulation

Zicong Fan[1,2], Omid Taheri[2], Dimitrios Tzionas[2], Muhammed Kocabas[1,2], Manuel Kaufmann[1], Michael J. Black[2], Otmar Hilliges[1]

[1]ETH Zürich, Switzerland
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

(5min video for this slides is on our website)

arctic.is.tue.mpg.de

Source: Pexels

Inanimate objects do not move by themselves

Laptop articulation is caused by the left hand

The milk pitcher and the mug are controlled by both hands

Hand-object Datasets

HOI4D

H2O-3D

ARCTIC

(a) Allocentric video

(b) Egocentric video

Our dataset

(c) Bimanual manipulation

(d) Changing contact

(e) Object articulation

ARCTIC

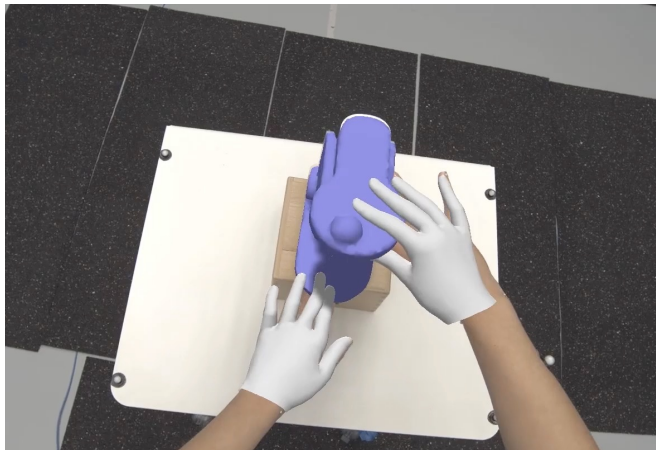
Left Hand

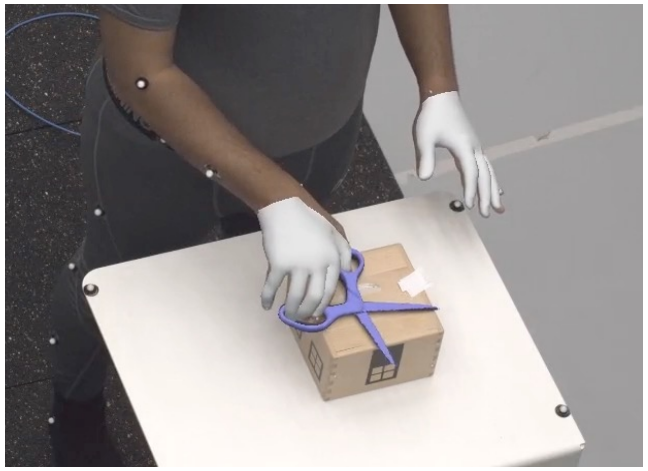
Right Hand

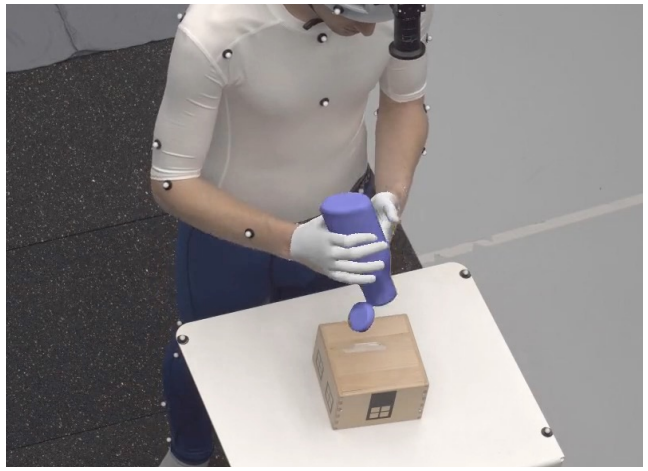
Object

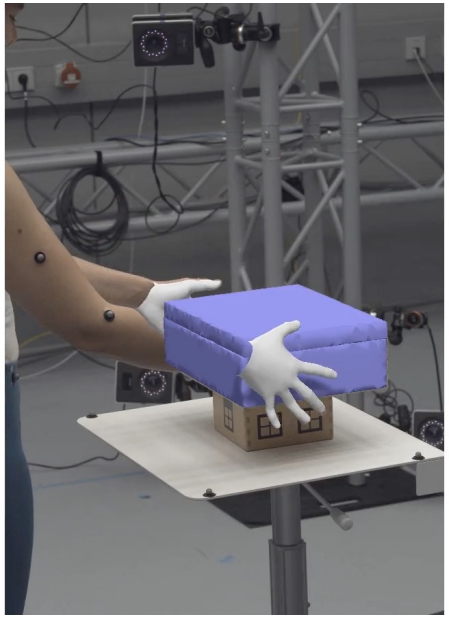
Hand-Object Distance (Brighter: closer)

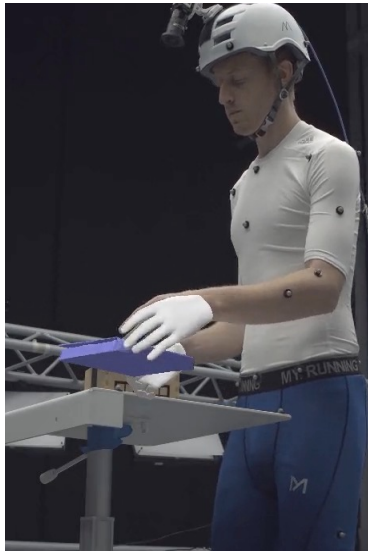
Rendered Sequences of ARCTIC Ground-truth

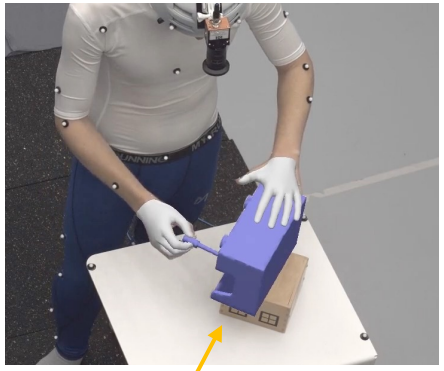






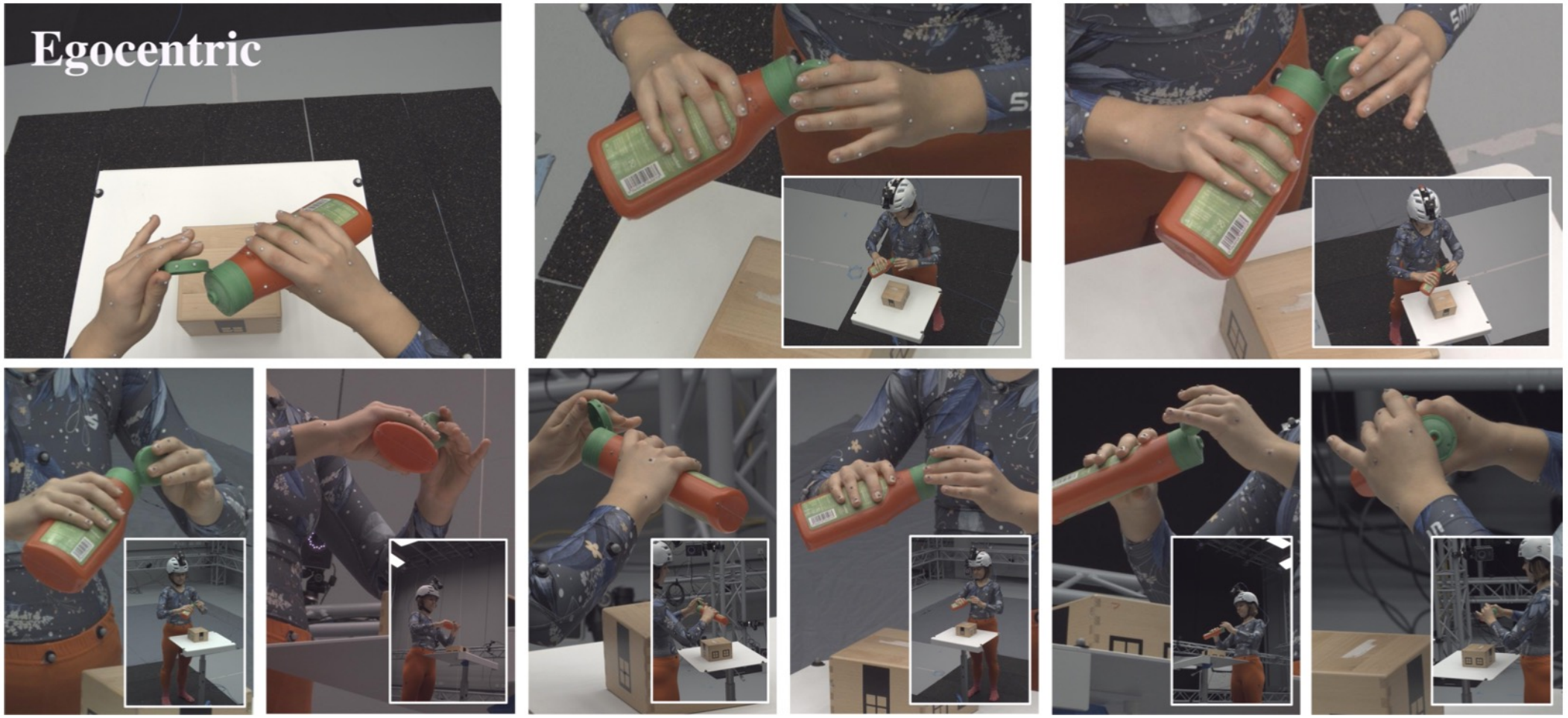




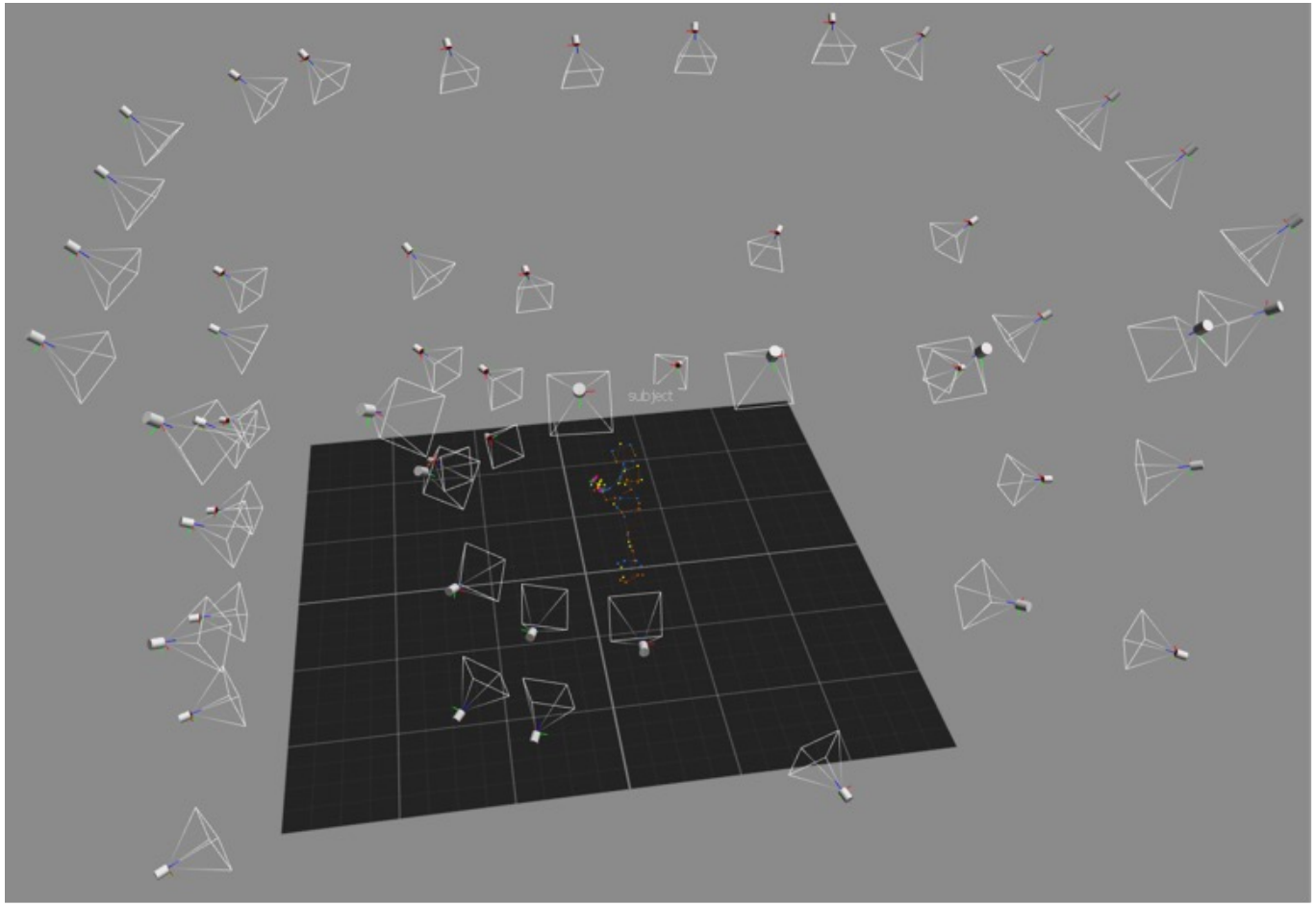


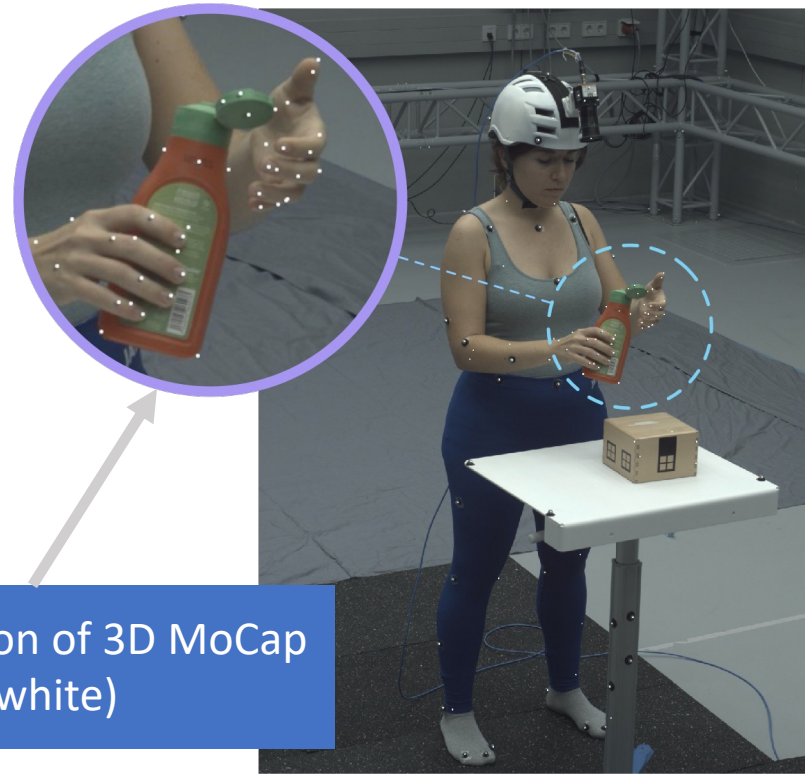
Ground-truth

ARCTIC Views (8x Allocentric + 1x Egocentric)



High Quality MoCap + RGB Setup





2D projection of 3D MoCap (**plotted** in white)

ARCTIC Objects

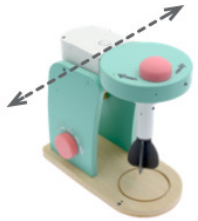

Laptop


Ketchup bottle


Microwave


Espresso machine


Notebook

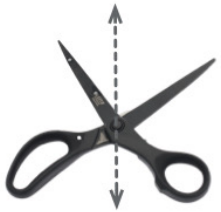

Mixer


Waffle iron


Phone

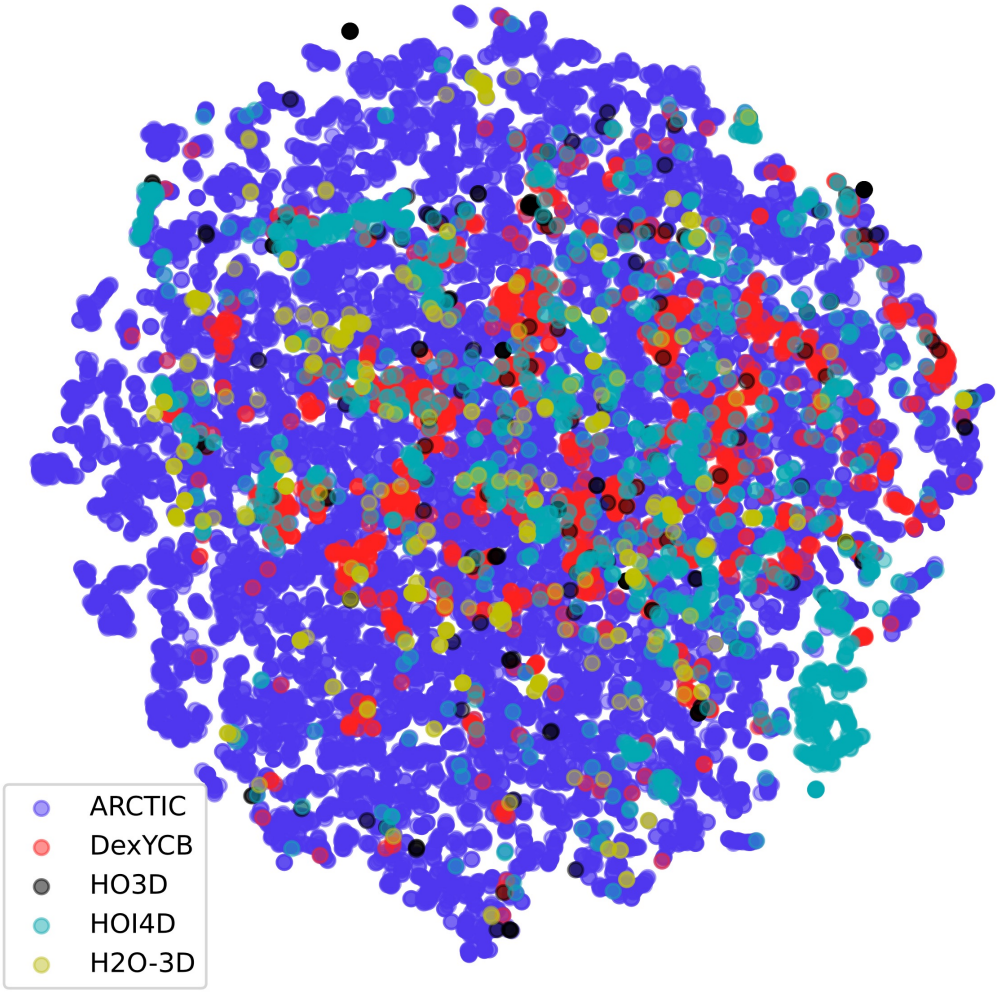

Box


Capsule machine


Scissors

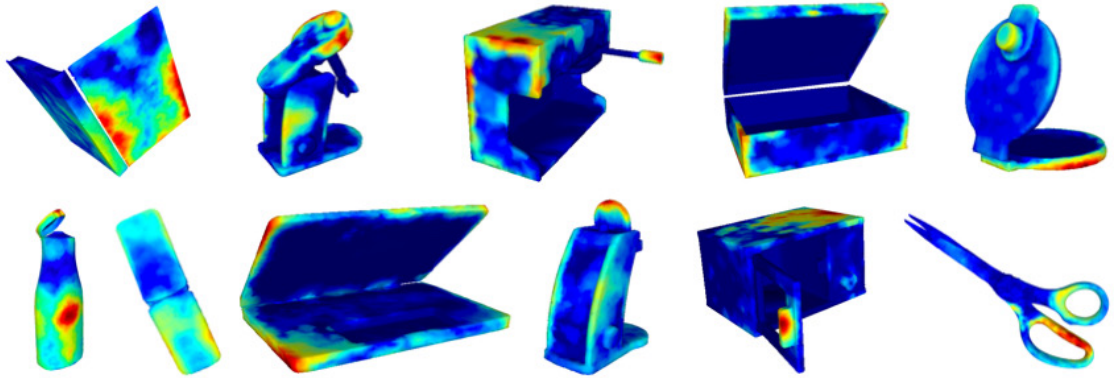
Compare ARCTIC with Existing Datasets

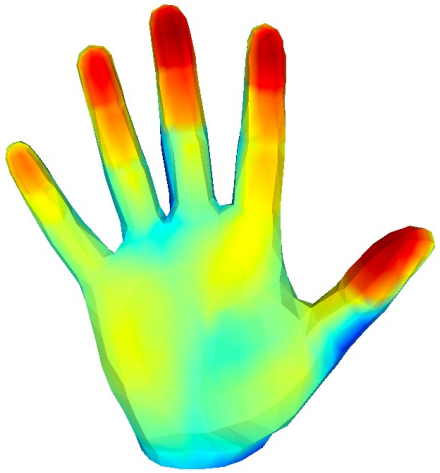
T-SNE Clustering



Aggregated Contact Heatmaps

Ours:

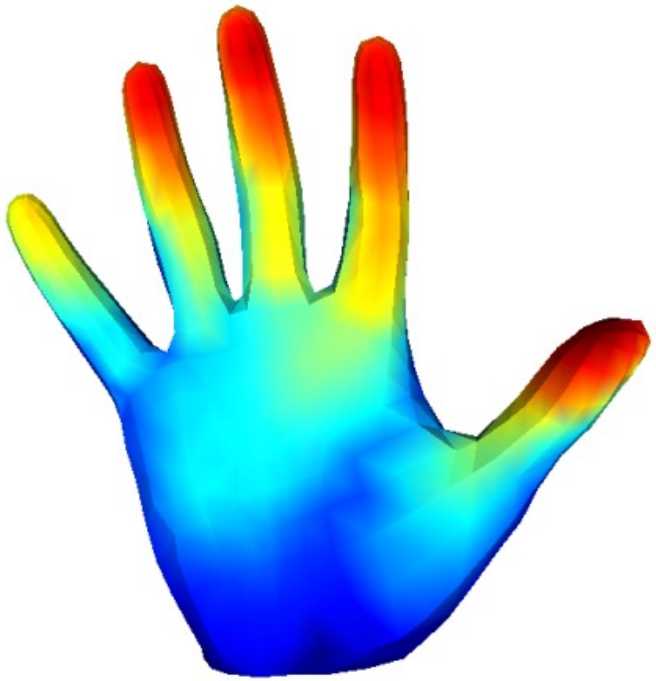




ARCTIC (Ours)

Aggregated Contact Heatmaps

HO3D



GRAB

ARCTIC (Ours)

Comparison to Existing Datasets

dataset	real images	# number of: img	# number of: view	ego-centric	image resol.	articulated objects	both hands	human body	dexterous manipulation	annot. type
FreiHand [69]	✓	37k	8	✗	224×224	✗	✗	✗	✗	semi-auto
ObMan [18]	✗	154k	1	✗	256×256	✗	✗	✗	✗	synthetic
FHPA [12]	✓	105k	1	✓	1920×1080	✗	✗	✗	✗	magnetic
HO3D [15]	✓	78k	1-5	✗	640×480	✗	✗	✗	✗	multi-kinect
ContactPose [5]	✓	2.9M	3	✗	960×540	✗	✗	✗	✗	multi-kinect
GRAB [52]	-	-	-	-	-	✗	✓	✓	✗	mocap
DexYCB [7]	✓	582k	8	✗	640×480	✗	✗	✗	✗	multi-manual
H2O [27]	✓	571k	5	✓	1280×720	✗	✓	✗	✗	multi-kinect
H2O-3D [16]	✓	76k	5	✗	640×480	✗	✓	✗	✗	multi-kinect
HOI4D [30]	✓	2.4M	1	✓	1280×800	✓	✗	✗	✗	single-manual
ARCTIC (Ours)	✓	2.1M	9	✓	2800×2000	✓	✓	✓	✓	mocap

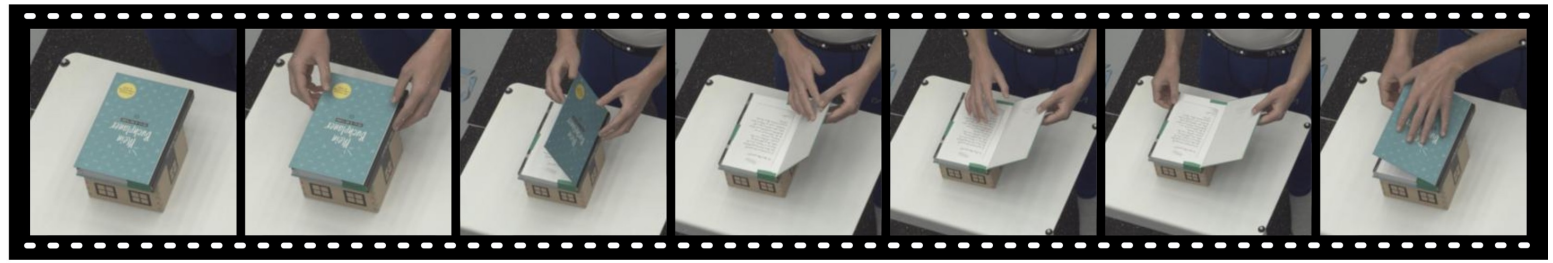
Table 1. **Comparison of our ARCTIC dataset with existing datasets**. The keyword "single/multi-manual" denotes whether single or multiple views being used to annotate manually.

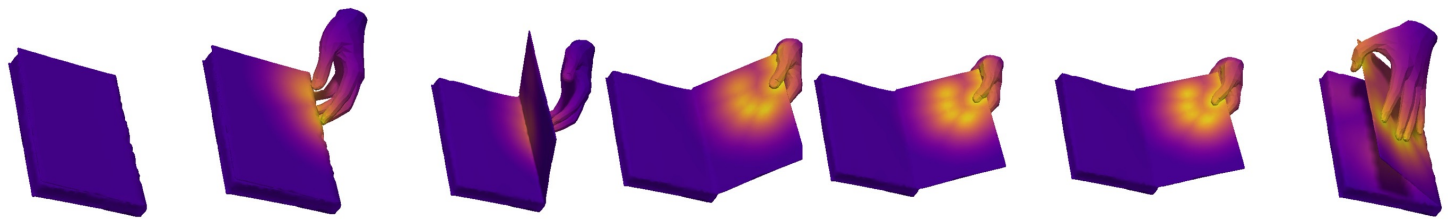
Consistent Motion Reconstruction

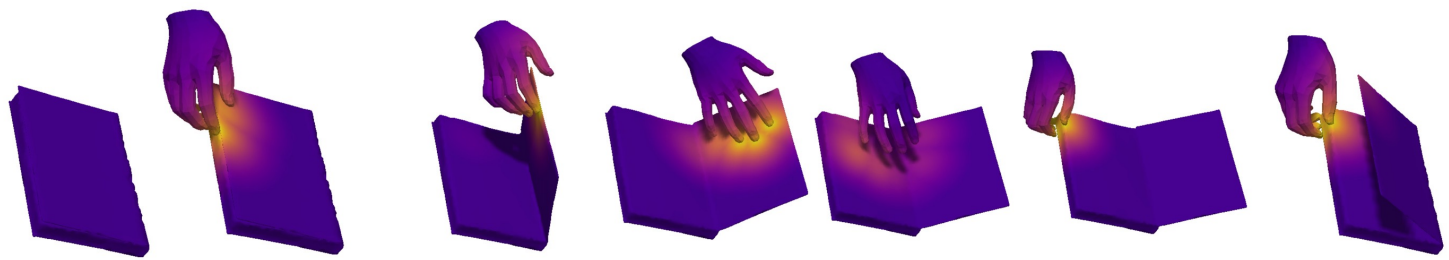
Monocular video

Articulated hand-object 3D motion (side view)

Interaction Field Estimation





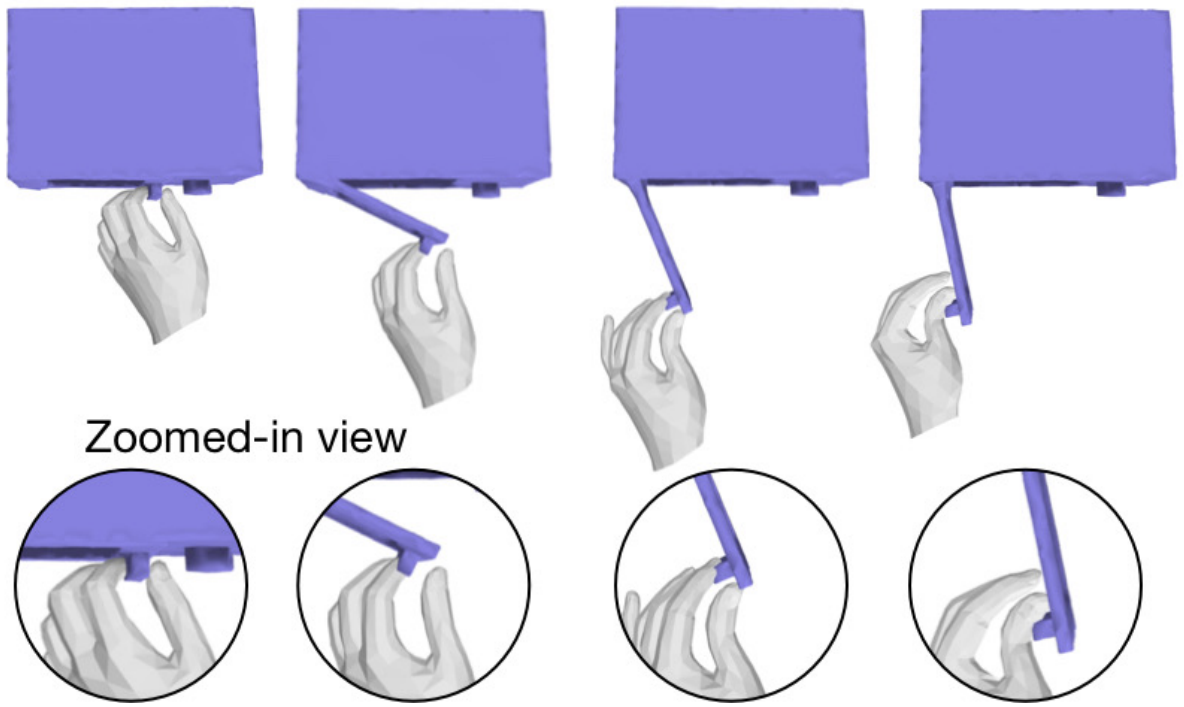


Consistent Motion Reconstruction

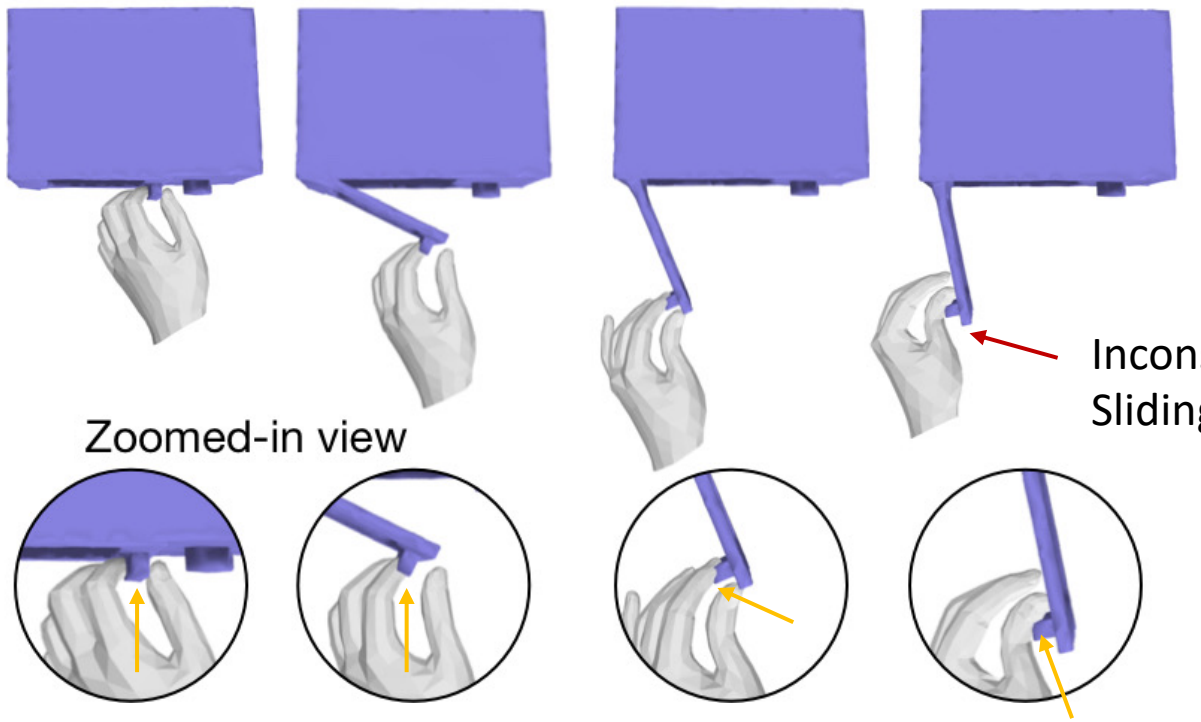
Monocular video

Articulated hand-object 3D motion (side view)

Example of inconsistent reconstruction



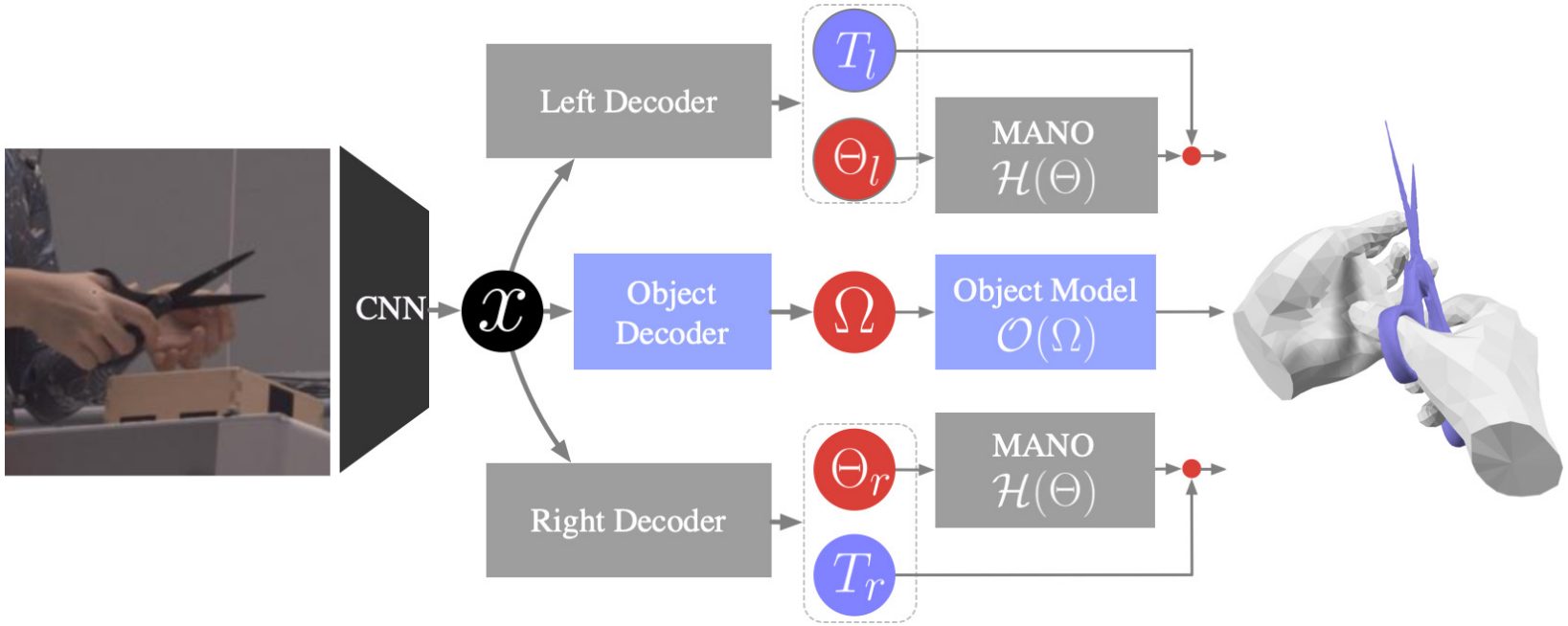
Example of inconsistent reconstruction



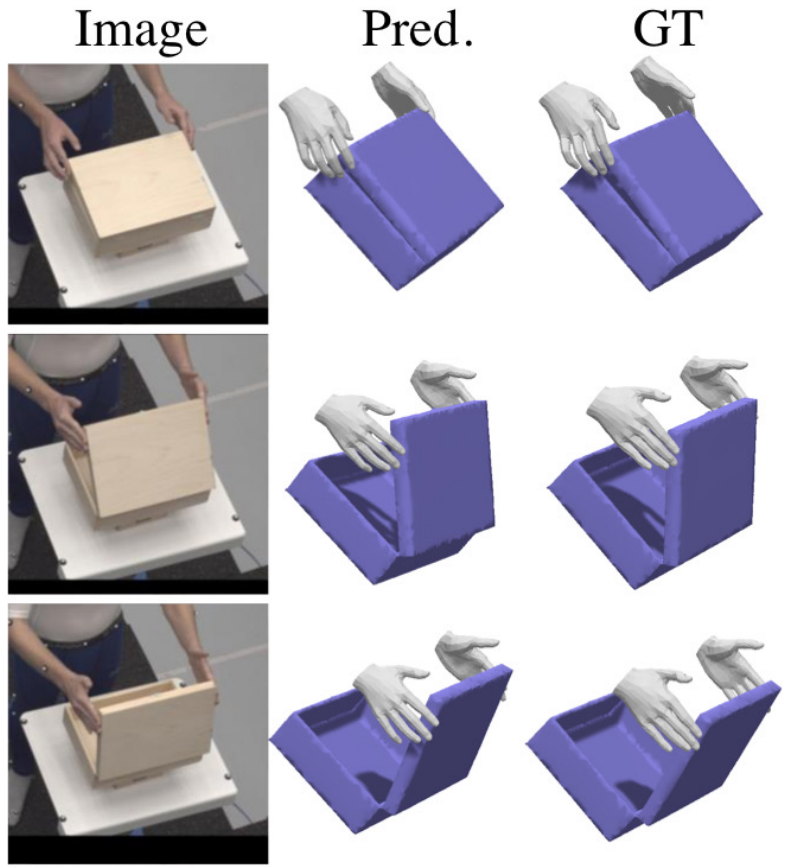
Inconsistent motion:
Sliding along the microwave door

Inconsistent contact:
Hand touches incorrect area on object.

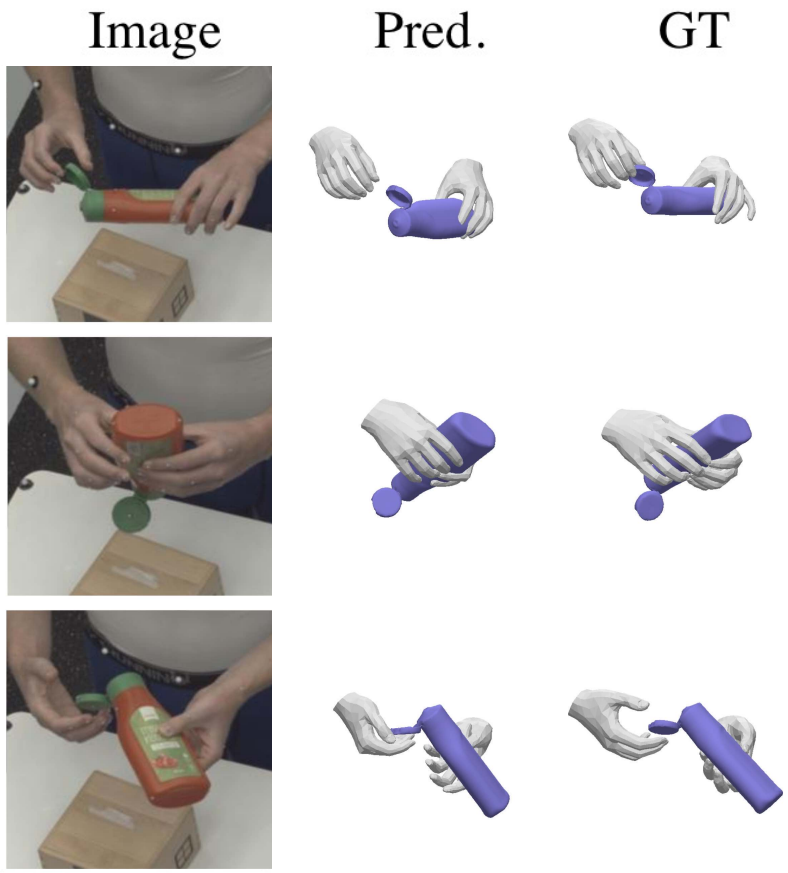
ArcticNet


CNN
x
Left Decoder
Object Decoder
Right Decoder
T_l
Θ_l
Ω
Θ_r
T_r
MANO $\mathcal{H}(\Theta)$
Object Model $\mathcal{O}(\Omega)$
MANO $\mathcal{H}(\Theta)$

ArcticNet LSTM – Qualitative Results



ArcticNet LSTM – Qualitative Results



Compare single-frame and temporal baselines

	Contact and Relative Position		Motion	
Method	CDev_{ho} $[mm]\downarrow$	$\text{MRRPE}_{rl/ro}$ $[mm]\downarrow$	MDev_{ho} $[mm]\downarrow$	$\text{ACC}_{h/o}$ $[m/s^2]\downarrow$
ArcticNet-SF	41.4	50.1/37.6	10.4	6.6/8.8
ArcticNet-LSTM	**38.8**	**47.1/36.8**	**8.9**	**5.6/6.9**

Compare single-frame and temporal baselines

	Contact and Relative Position		Motion	
Method	CDev_{ho} $[mm]\downarrow$	$\text{MRRPE}_{rl/ro}$ $[mm]\downarrow$	MDev_{ho} $[mm]\downarrow$	$\text{ACC}_{h/o}$ $[m/s^2]\downarrow$
ArcticNet-SF	41.4	50.1/37.6	10.4	6.6/8.8
ArcticNet-LSTM	**38.8**	**47.1/36.8**	**8.9**	**5.6/6.9**

Hand-object contact

Compare single-frame and temporal baselines

	Contact and Relative Position		Motion	
Method	CDev_{ho} $[mm]\downarrow$	$\text{MRRPE}_{rl/ro}$ $[mm]\downarrow$	MDev_{ho} $[mm]\downarrow$	$\text{ACC}_{h/o}$ $[m/s^2]\downarrow$
ArcticNet-SF	41.4	50.1/37.6	10.4	6.6/8.8
ArcticNet-LSTM	**38.8**	**47.1/36.8**	**8.9**	**5.6/6.9**

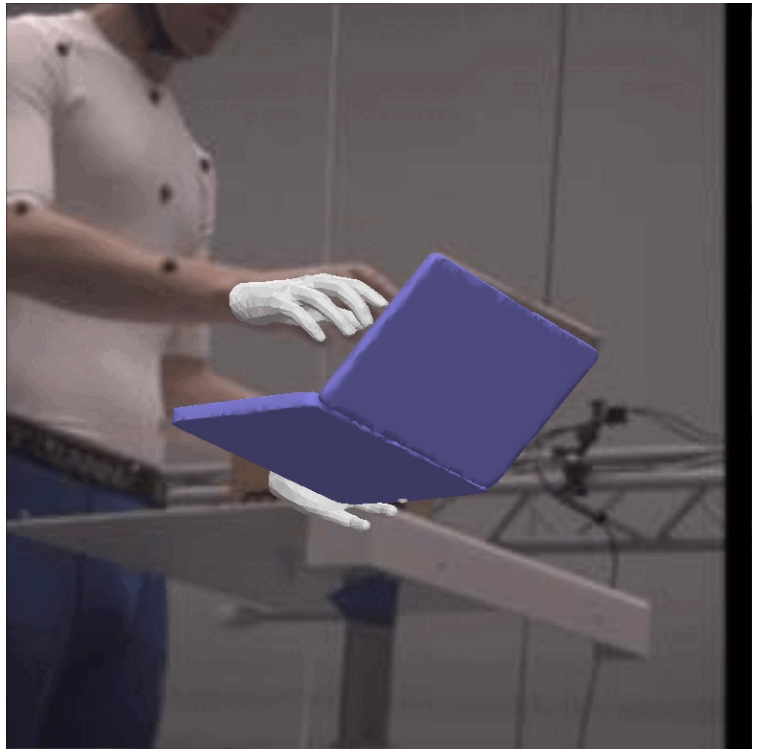
Consistency in hand-object motion

Compare single-frame and temporal baselines

	Contact and Relative Position		Motion	
Method	CDev_{ho} $[mm]\downarrow$	$\text{MRRPE}_{rl/ro}$ $[mm]\downarrow$	MDev_{ho} $[mm]\downarrow$	$\text{ACC}_{h/o}$ $[m/s^2]\downarrow$
ArcticNet-SF	41.4	50.1/37.6	10.4	6.6/8.8
ArcticNet-LSTM	**38.8**	**47.1/36.8**	**8.9**	**5.6/6.9**

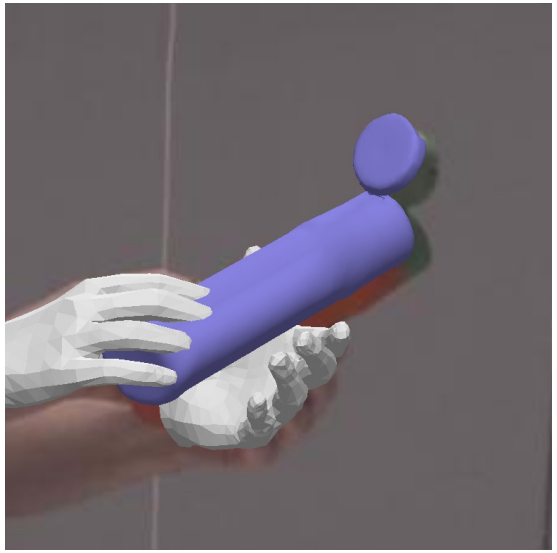
Temporal smoothness in prediction

ArcticNet LSTM – Failure Cases



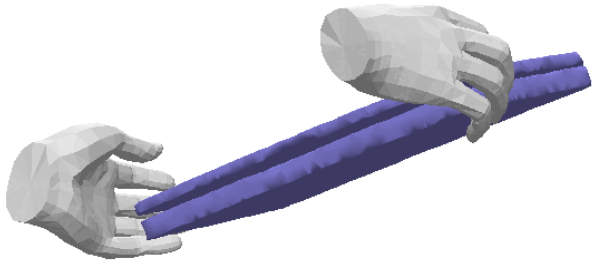
Jitter in laptop prediction (see video)

ArcticNet LSTM – Failure Cases



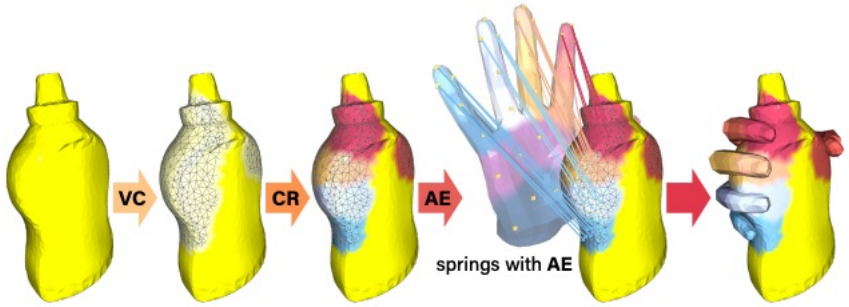
Imperfect 2D alignment

ArcticNet LSTM – Failure Cases

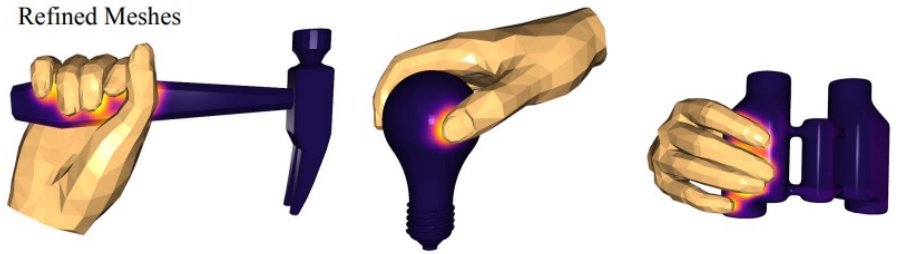


Inconsistency of hand-object contact (left hand)

Contact is important.

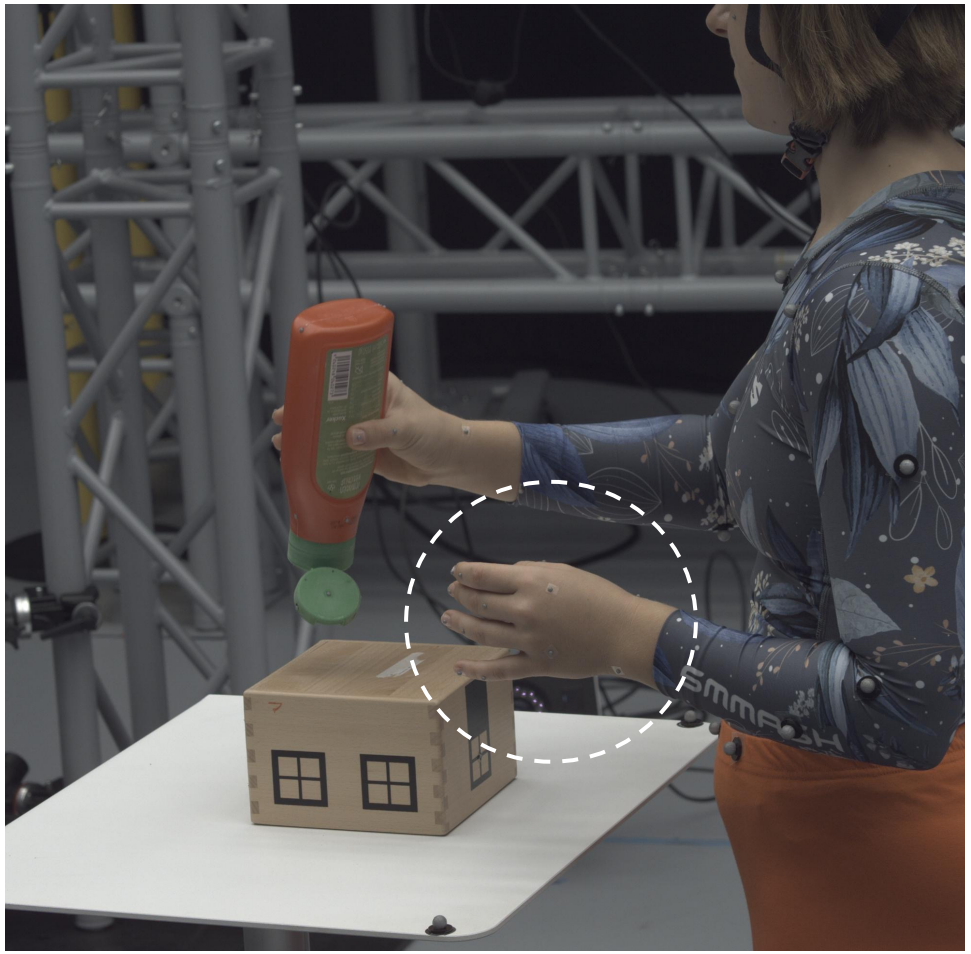


Yang et al., 2021

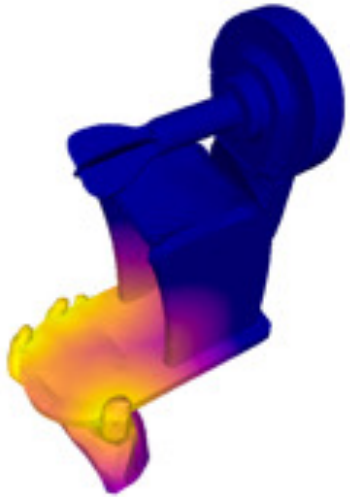


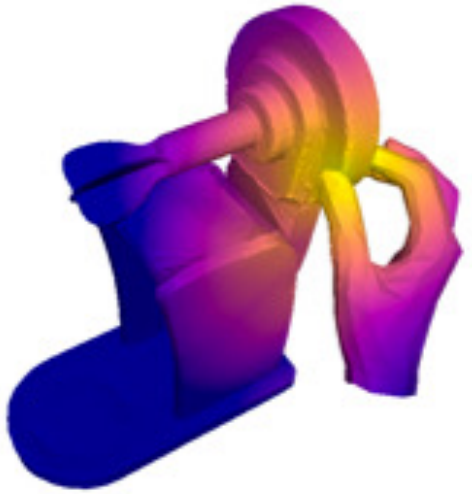
Grady et al. 2021

Binary Contact Estimation



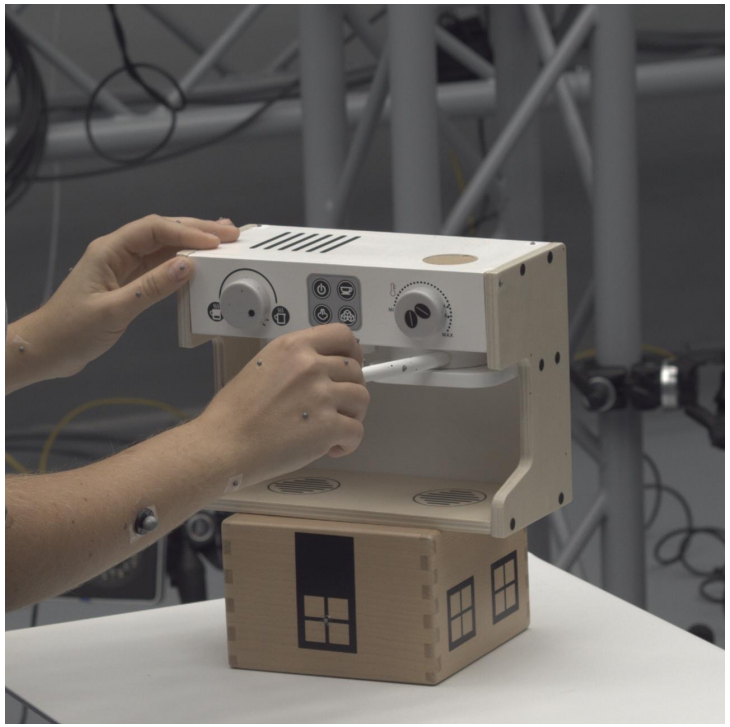
Interaction Field Estimation



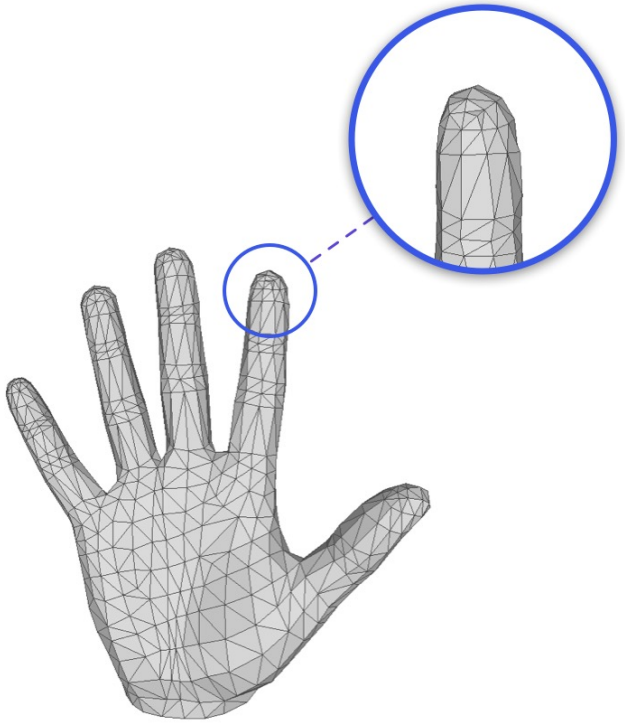


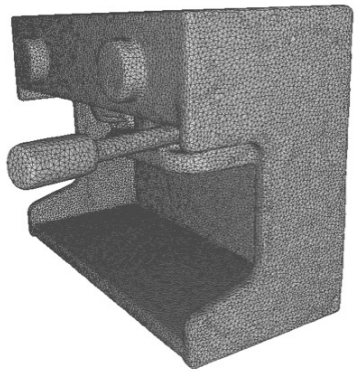
Interaction Field Estimation

Input:



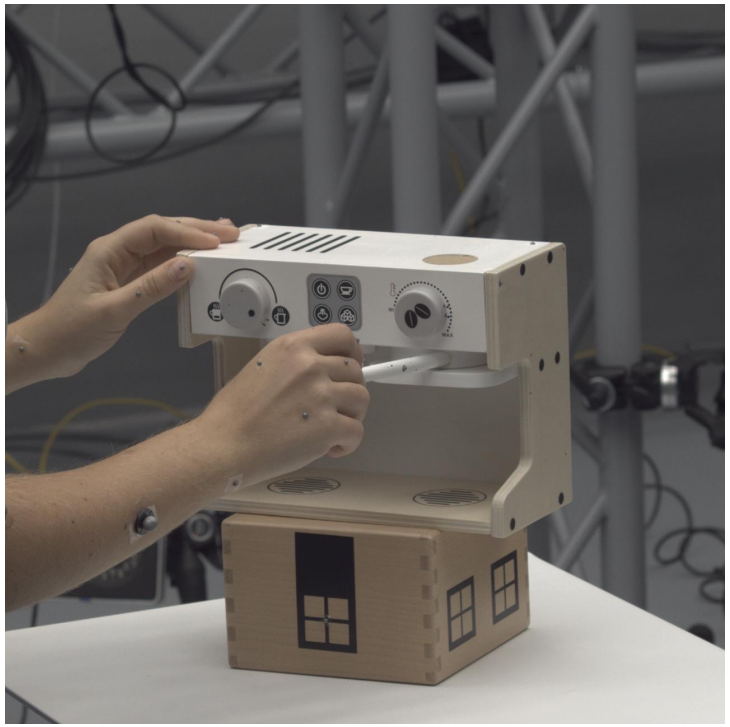
Output:



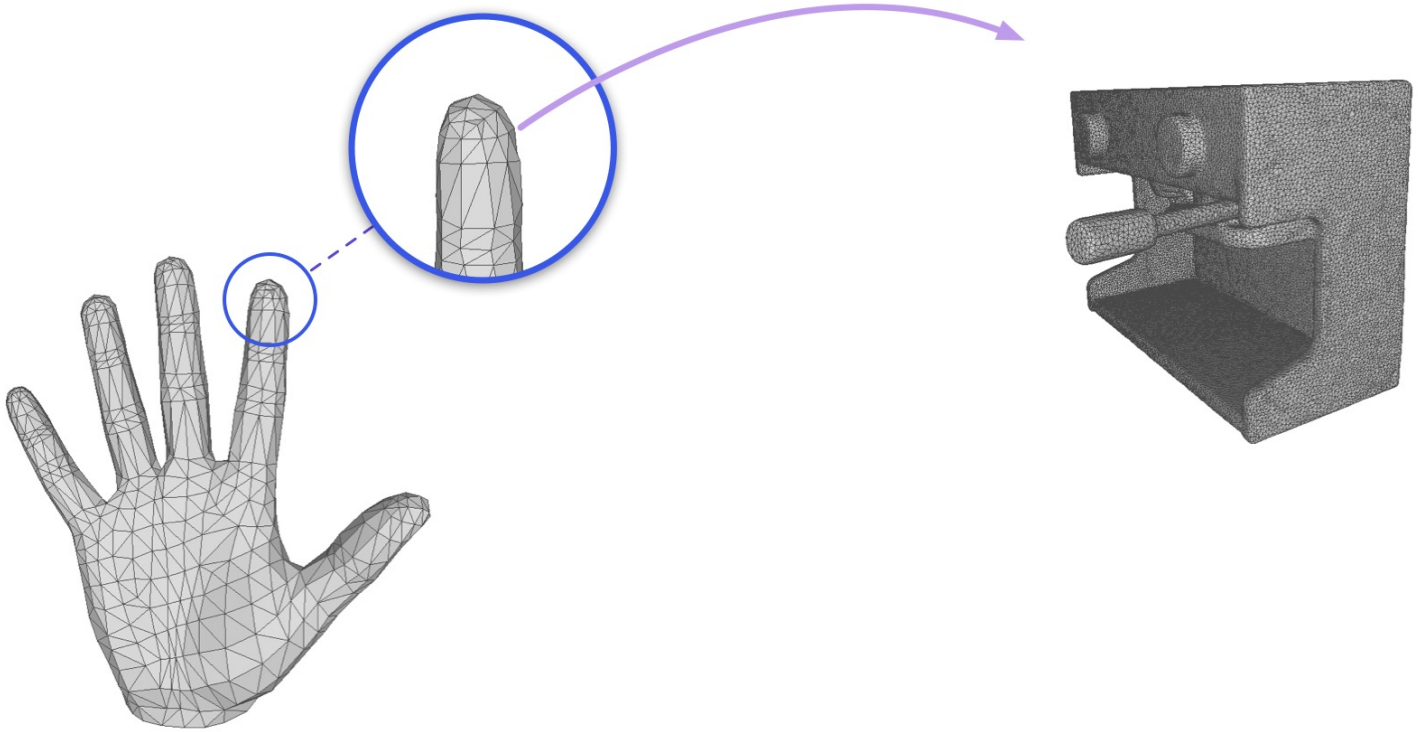


Interaction Field Estimation

Input:

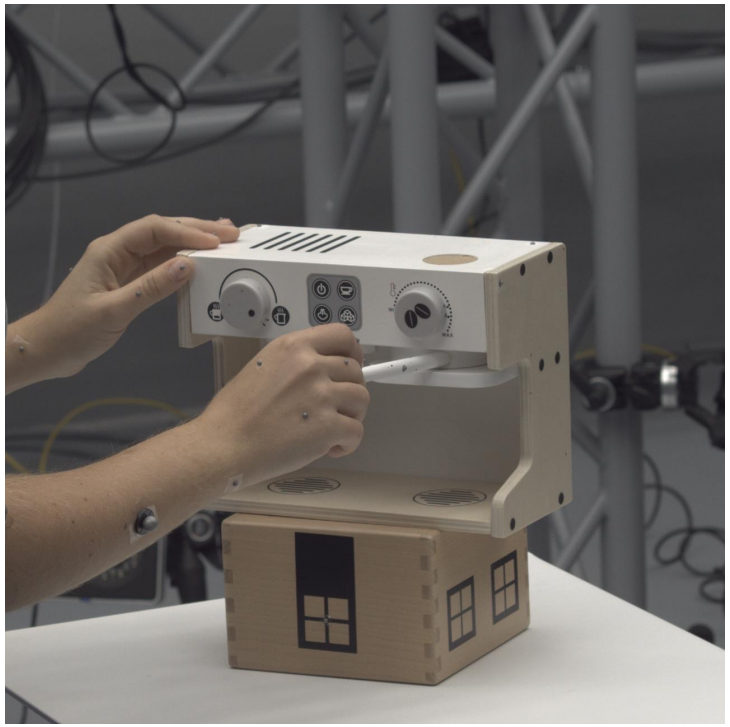


Output:

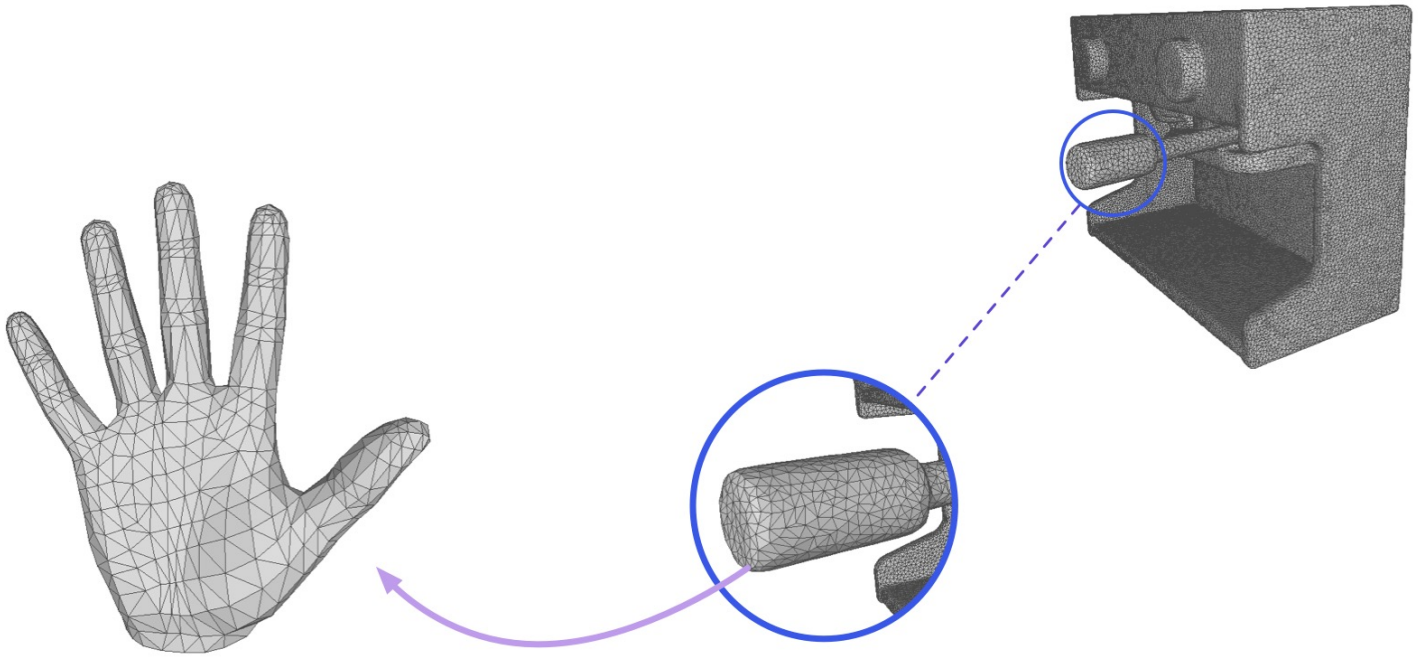


Interaction Field Estimation

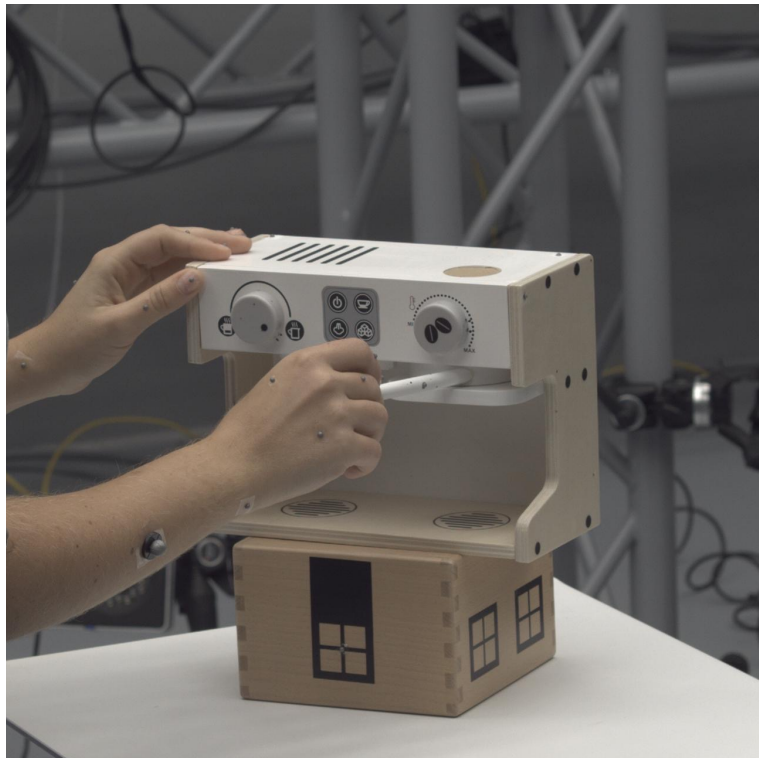
Input:

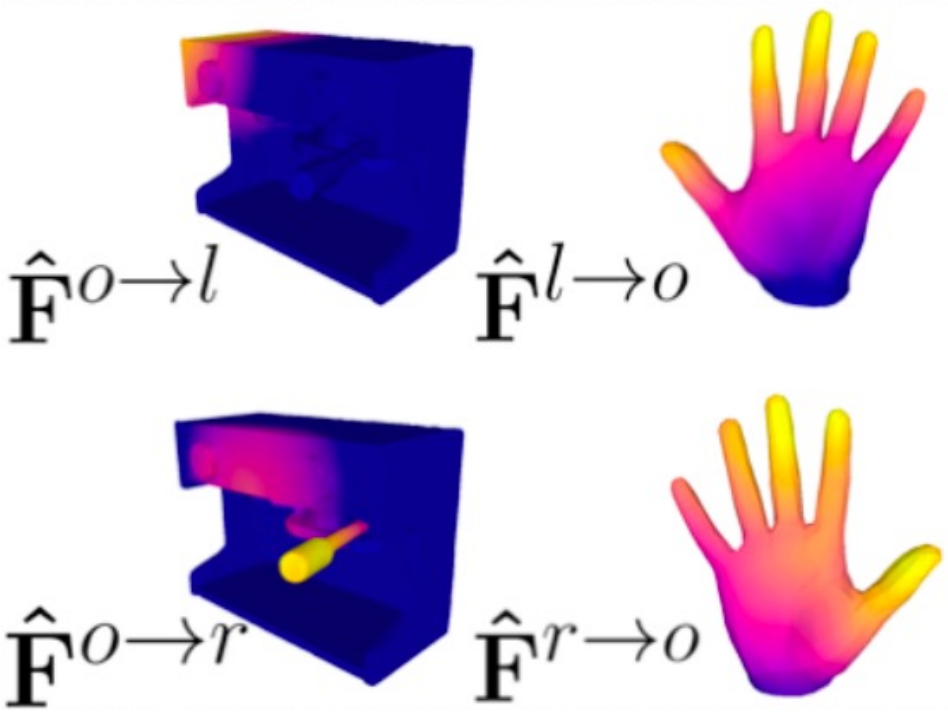


Output:

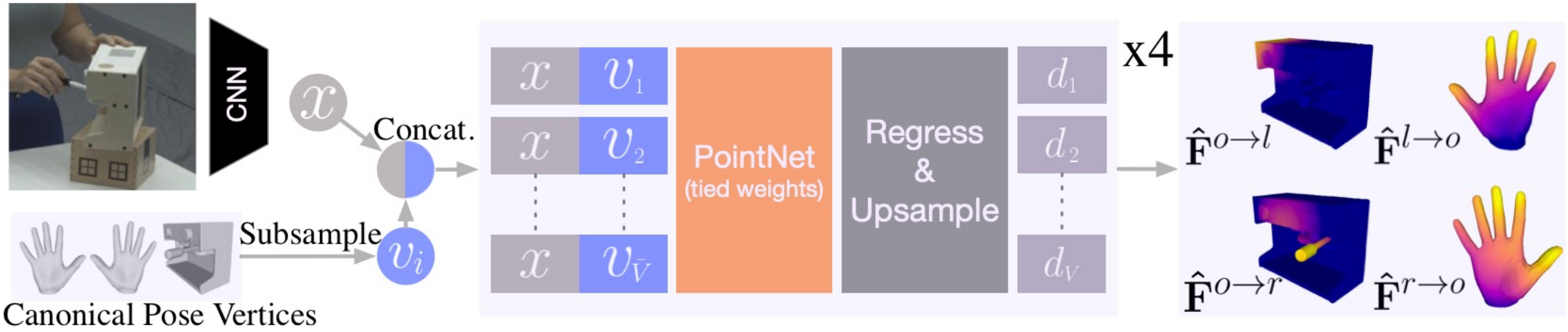


Interaction Field Estimation

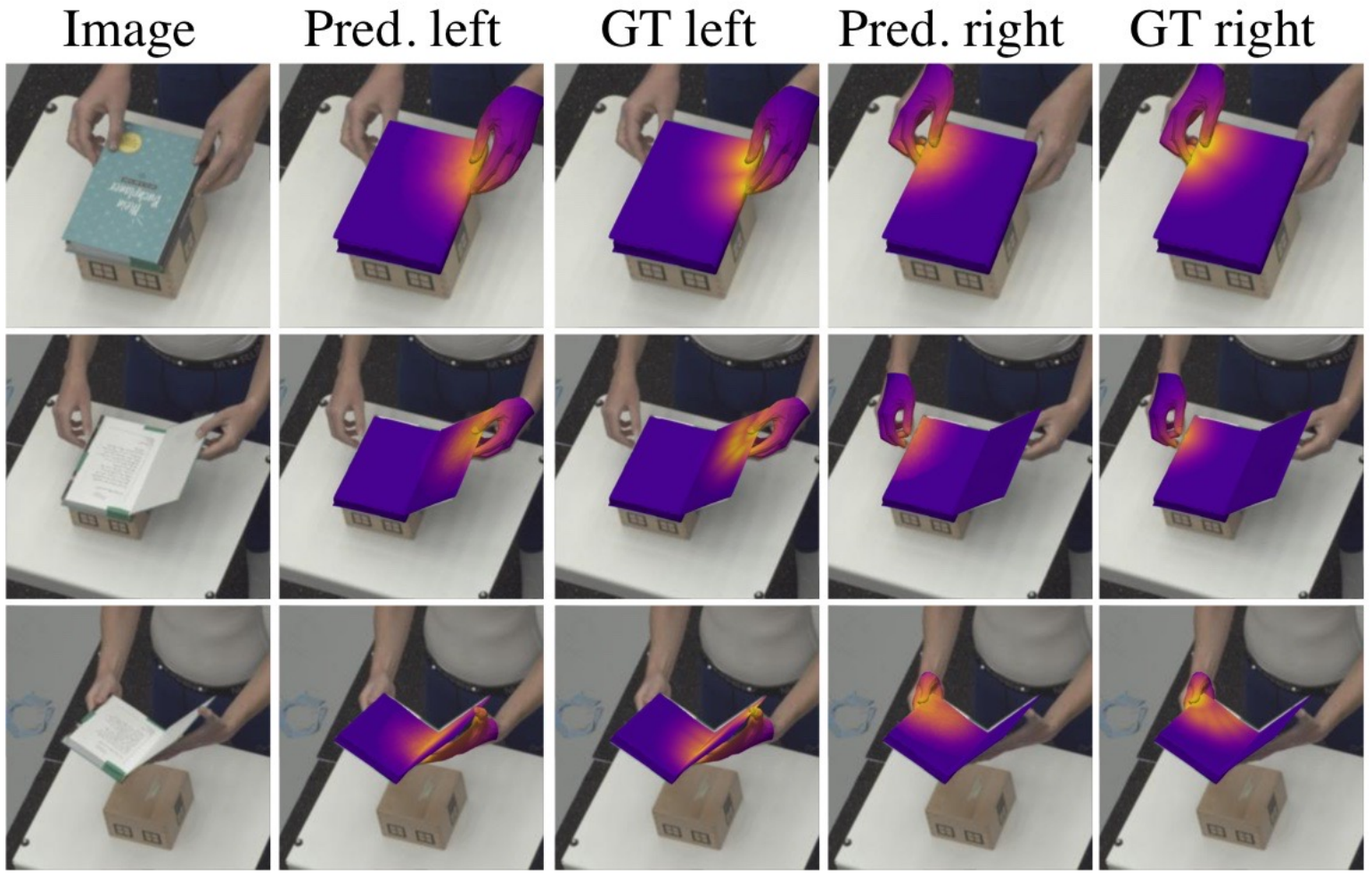




InterField

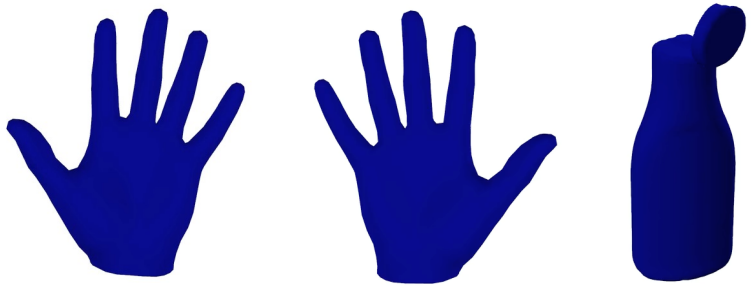

CNN
x
Concat.
Subsample
v_i
Canonical Pose Vertices
x v_1
x v_2
x $v_{\bar{V}}$
PointNet
(tied weights)
Regress
&
Upsample
d_1
d_2
d_V
x4
$\hat{\mathbf{F}}^{o\to l}$
$\hat{\mathbf{F}}^{l\to o}$
$\hat{\mathbf{F}}^{o\to r}$
$\hat{\mathbf{F}}^{r\to o}$

InterField - Qualitative Results



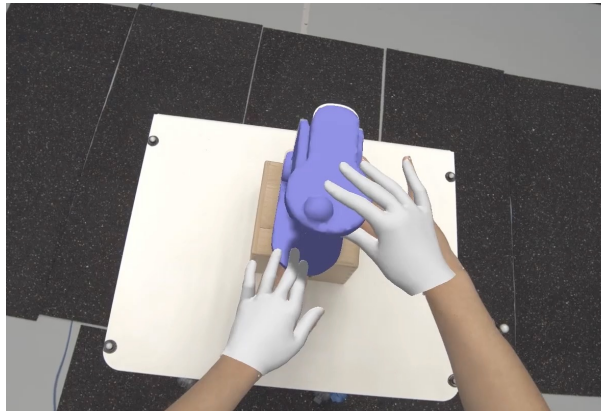
Conclusion

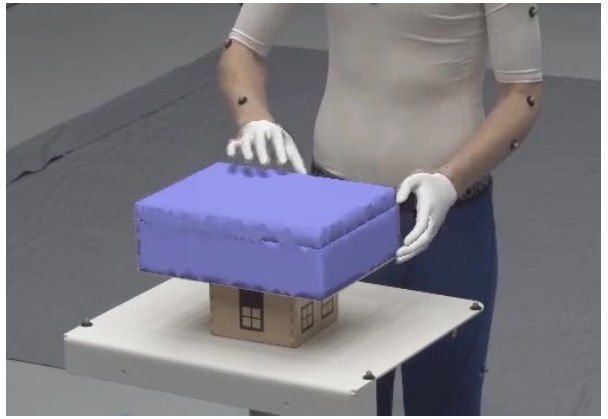
ARCTIC

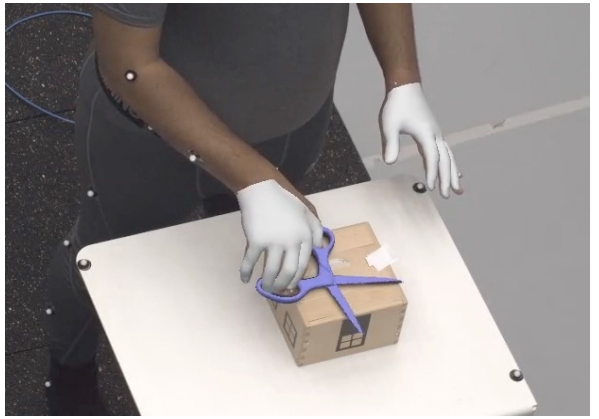


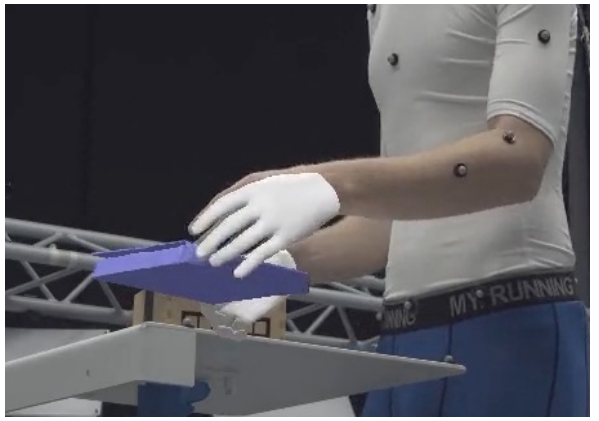
Left Hand | Right Hand | Object

Hand-Object Distance (Brighter: closer)



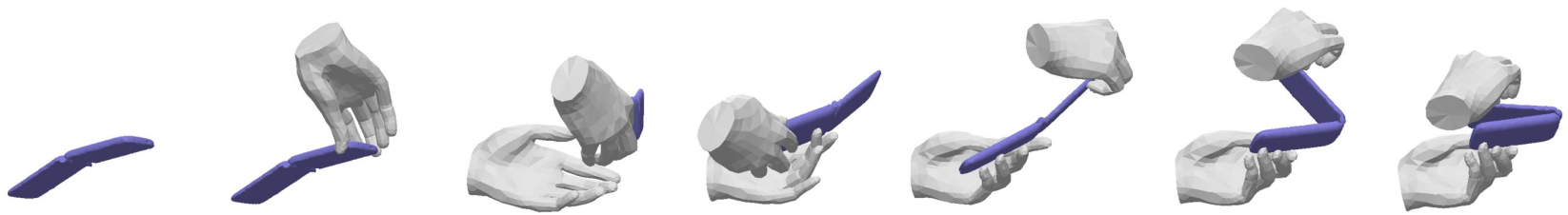






Monocular video

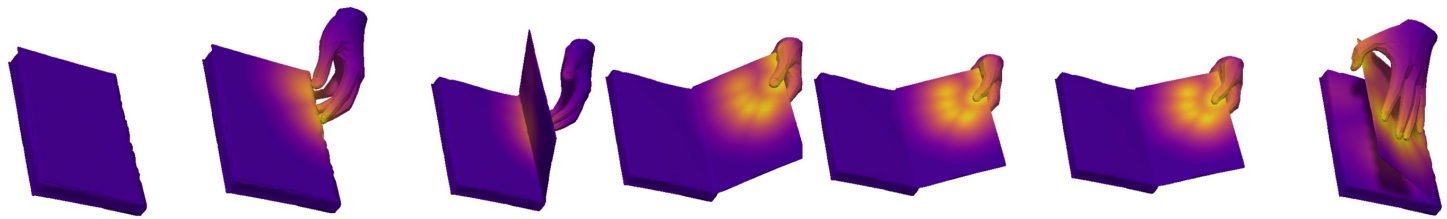
Articulated hand-object 3D motion (side view)



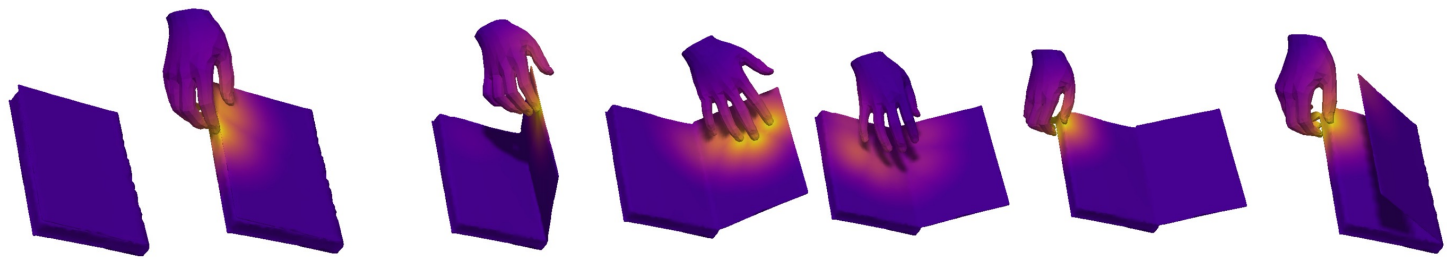
Consistent motion reconstruction

Monocular video

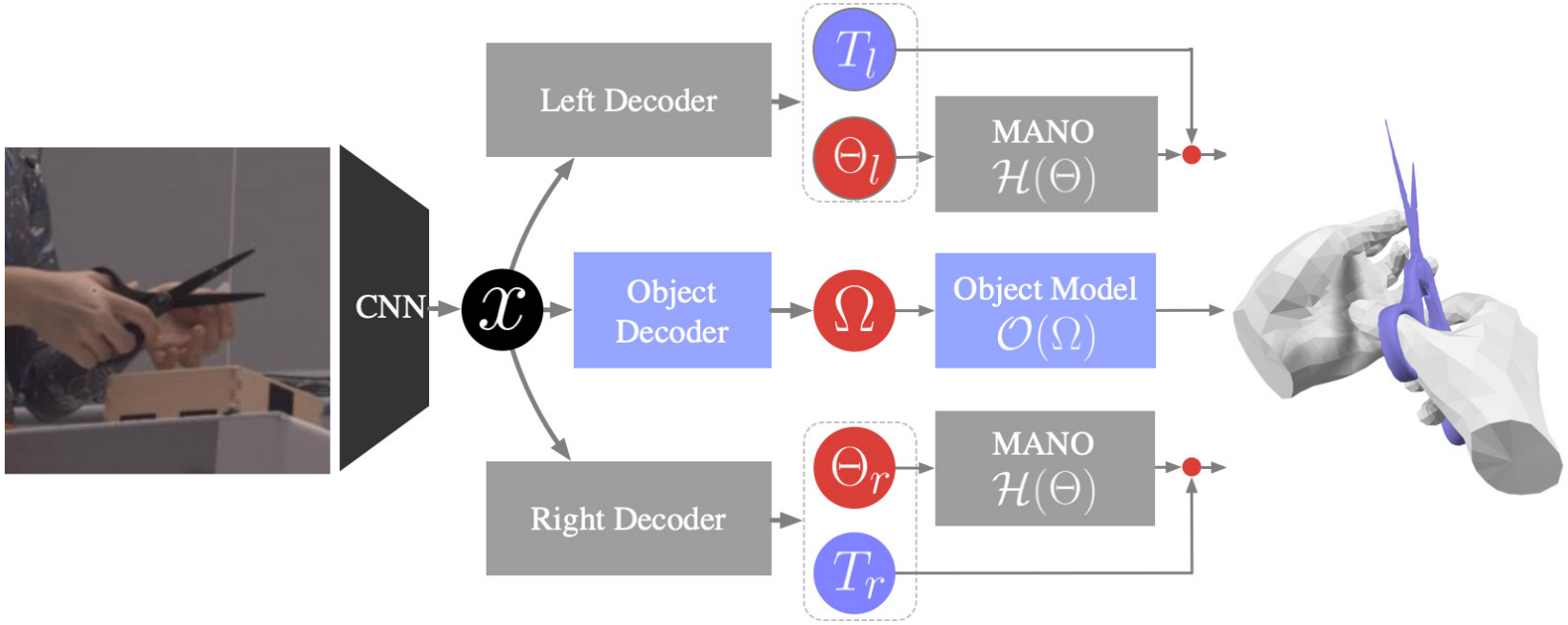
Interaction Field (Left)



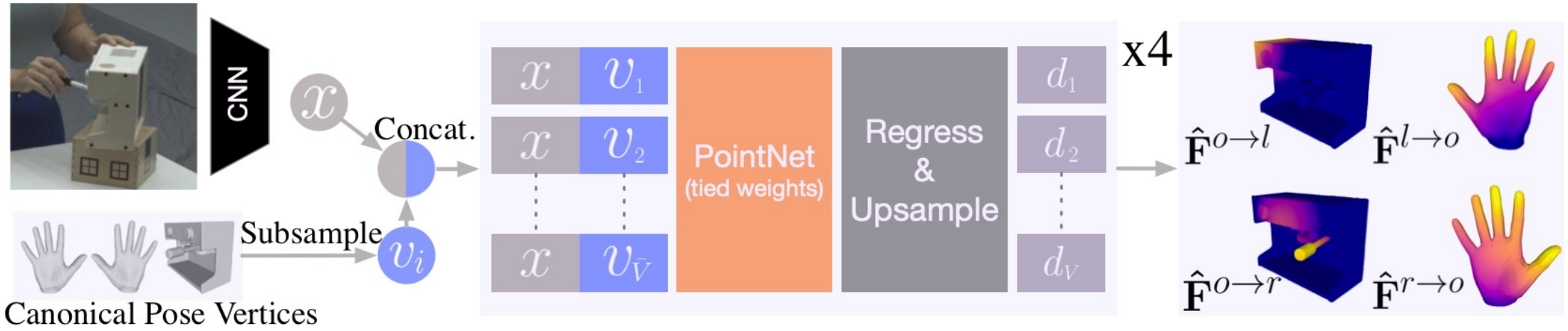
Interaction Field (Right)



Interaction field estimation


CNN
$\mathcal{X}$
Left Decoder
Object Decoder
Right Decoder
T_l
Θ_l
Ω
Θ_r
T_r
MANO $\mathcal{H}(\Theta)$
Object Model $\mathcal{O}(\Omega)$
MANO $\mathcal{H}(\Theta)$

ArcticNet


CNN
x
Concat.
Subsample
v_i
Canonical Pose Vertices
x v_1
x v_2
x $v_{\bar{V}}$
PointNet
(tied weights)
Regress
&
Upsample
d_1
d_2
d_V
x4
$\hat{\mathbf{F}}^{o\to l}$
$\hat{\mathbf{F}}^{l\to o}$
$\hat{\mathbf{F}}^{o\to r}$
$\hat{\mathbf{F}}^{r\to o}$

InterField

A Dataset for Dexterous Bimanual Hand-Object Manipulation

Zicong Fan[1,2], Omid Taheri[2], Dimitrios Tzionas[2], Muhammed Kocabas[1,2], Manuel Kaufmann[1], Michael J. Black[2], Otmar Hilliges[1]

[1]ETH Zürich, Switzerland
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

arctic.is.tue.mpg.de