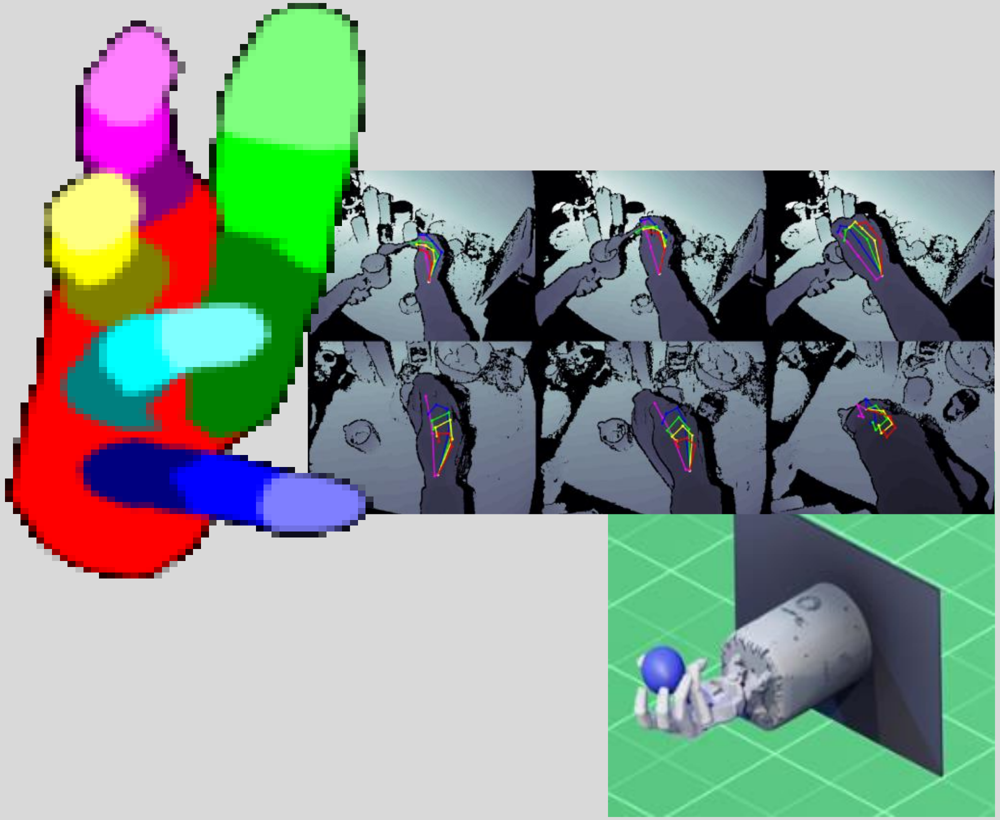
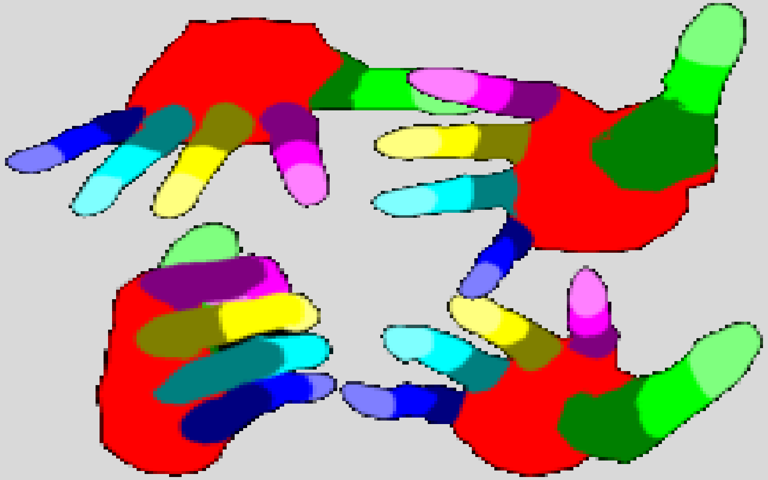
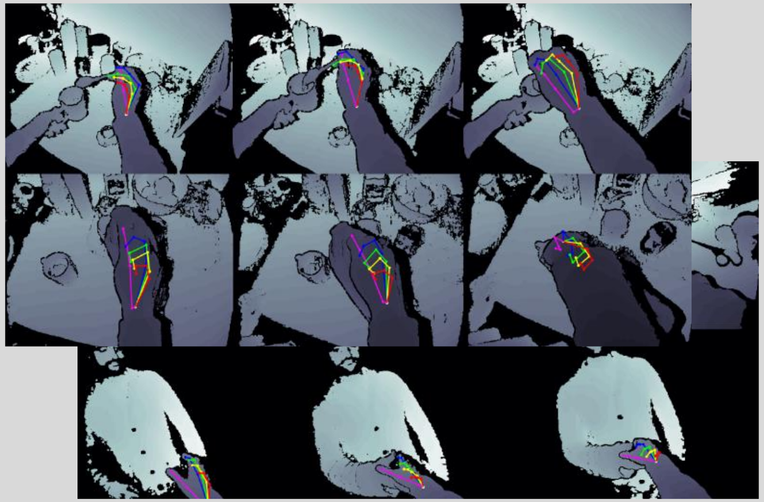
3D Hand Pose Estimation

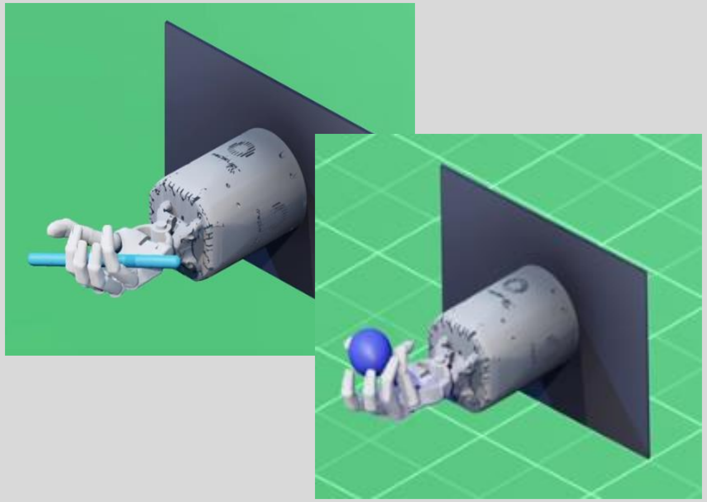


Tae-Kyun (T-K) Kim
Computer Vision and Learning Lab
School of Computing
KAIST

https://sites.google.com/view/tkkim/







aka Vision-based 3D Finger Tracking

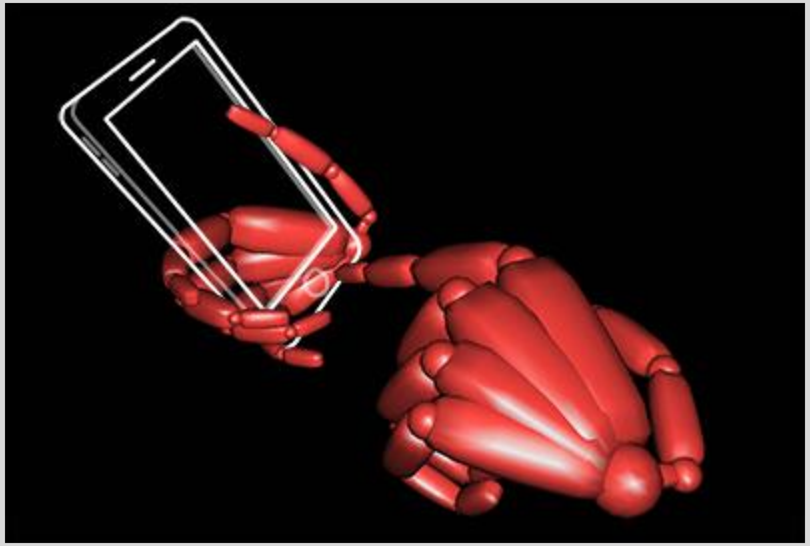
How to interact with AR/VR environment

[Oculus]

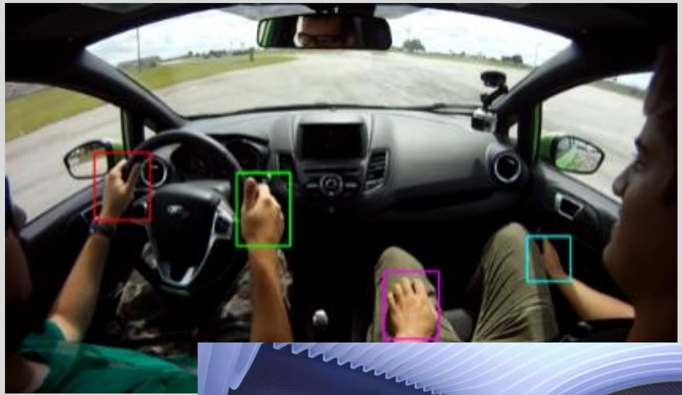
[Upload VR]

[Leap Motion]

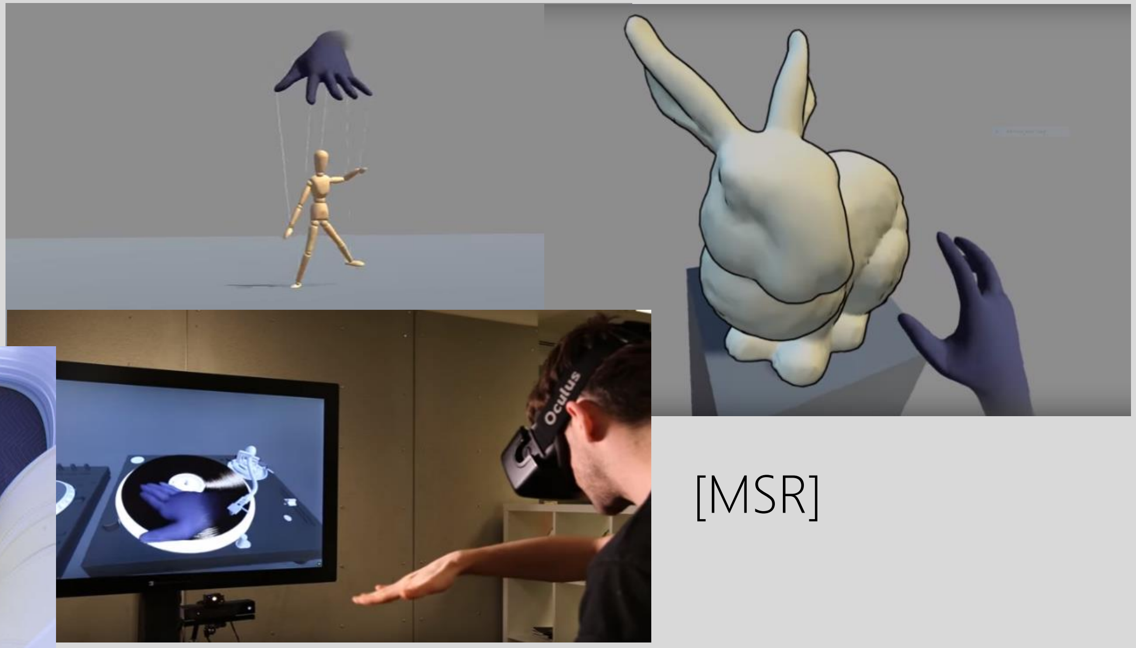
[NANSENSE]



More examples: AR/VR in autonomous cars



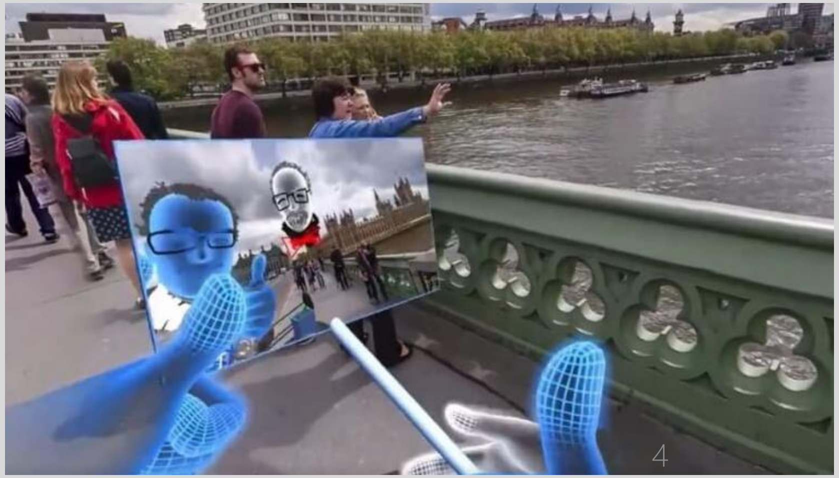
[UCSD]



[MSR]

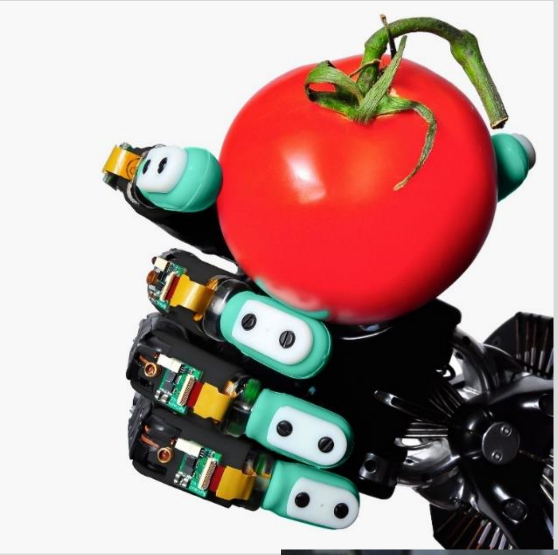
[Upload VR]

Driver-vehicle interaction



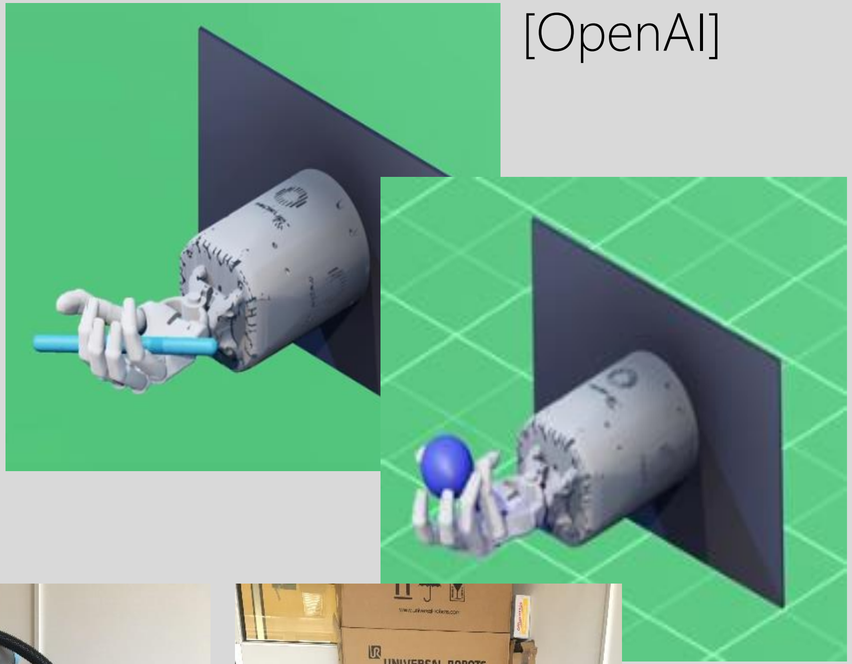
[Oculus]

Physical interactions and robotics



[SynTouch]

[Spread]



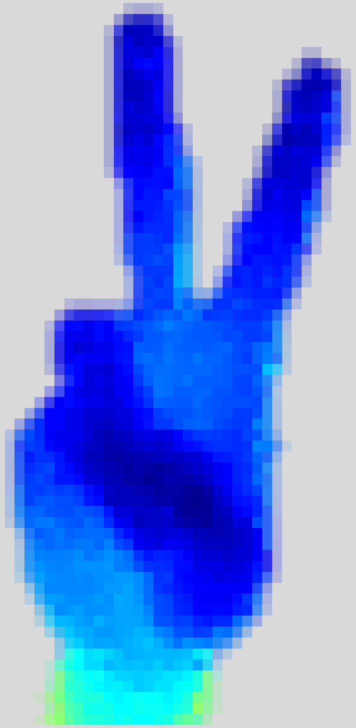
[OpenAI]

Robot-human interaction

[ICL & Samsung Research]

Problem statement

Input (RGB or Depth) Image



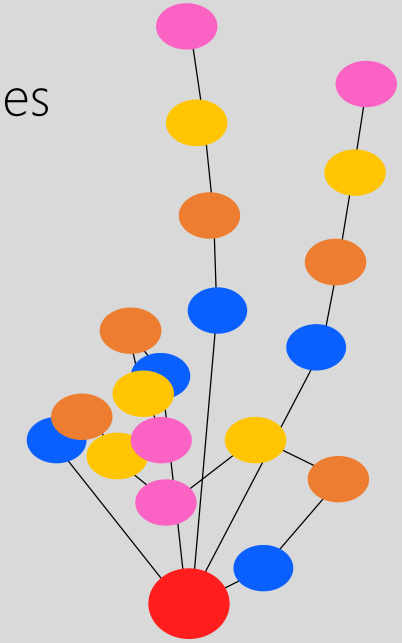
Z

Extract joint 3D locations/angles

$$\theta \in \mathbb{R}^d$$

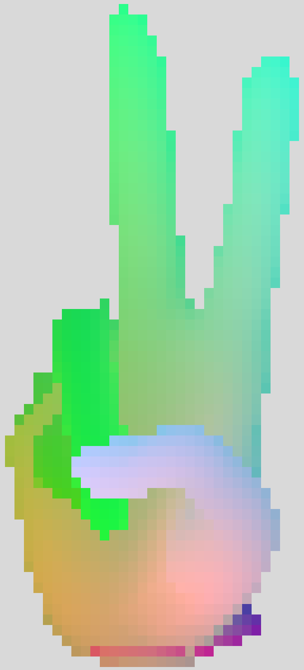
for current frame

Skeleton



θ

Rendered depth



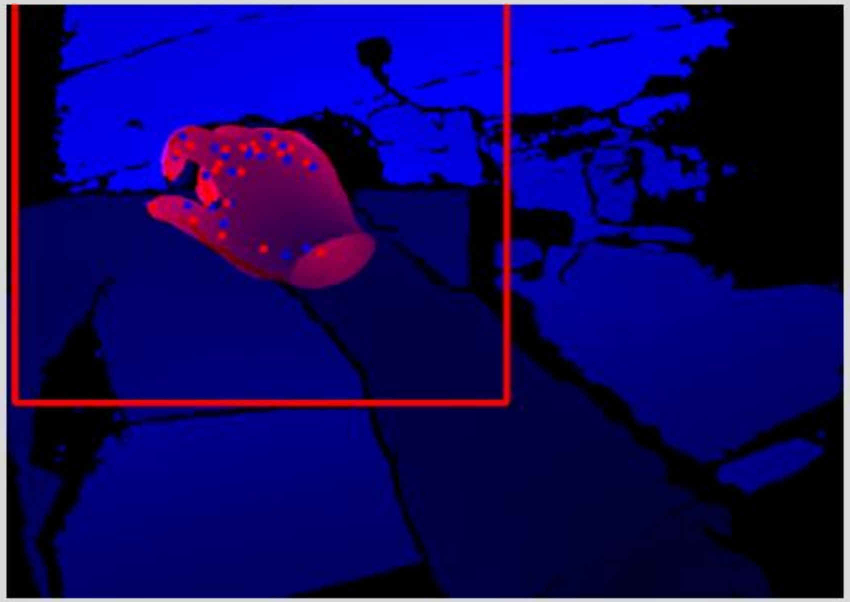
R_θ

Challenges:

- High degree of freedom (d=26)
- Viewpoint changes and self occlusions
- Fast movement
- Annotation difficulty
- Shape variation

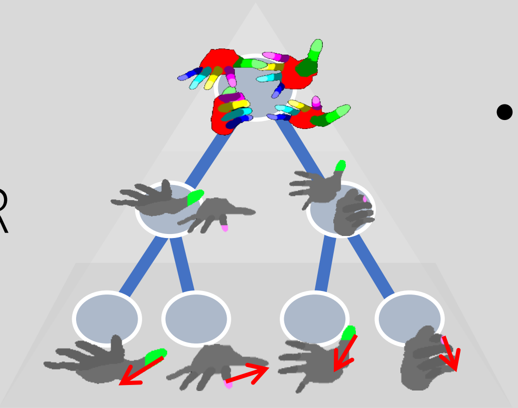
3D Dense pose prediction ECCV2020

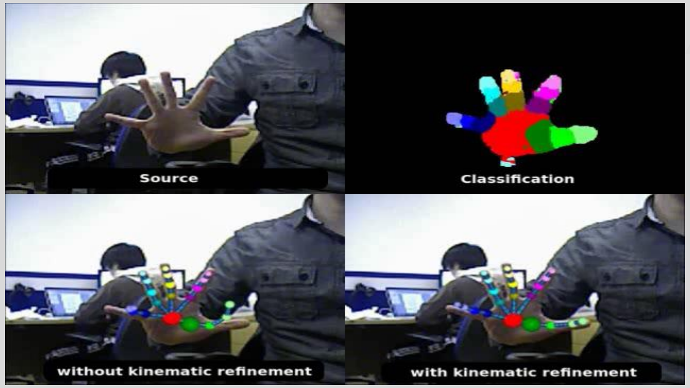
- HANDS19 Challenge @ ICCV includes: Hand-object interaction, depth and colour modalities, extrapolation capabilities, the use of synthetic data (MANO).
- Fitted mesh models to BigHand2.2M, F-PHAB, HO-3D datasets, are provided.



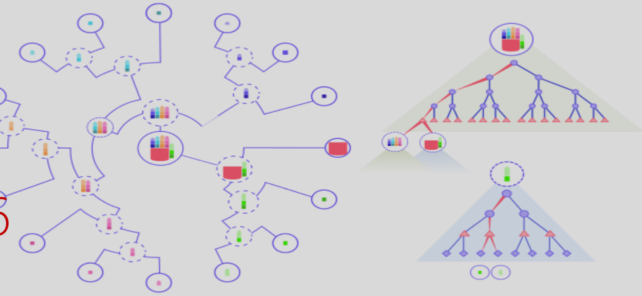
Advances so far

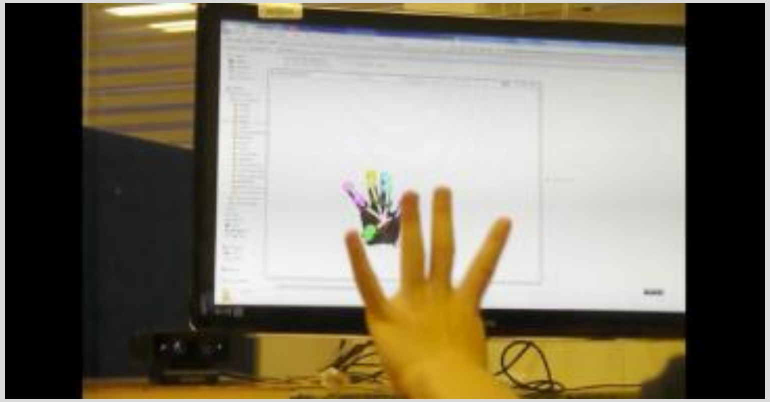
- Hierarchical, multi-task, transductive learning: STR forest ICCV13oral



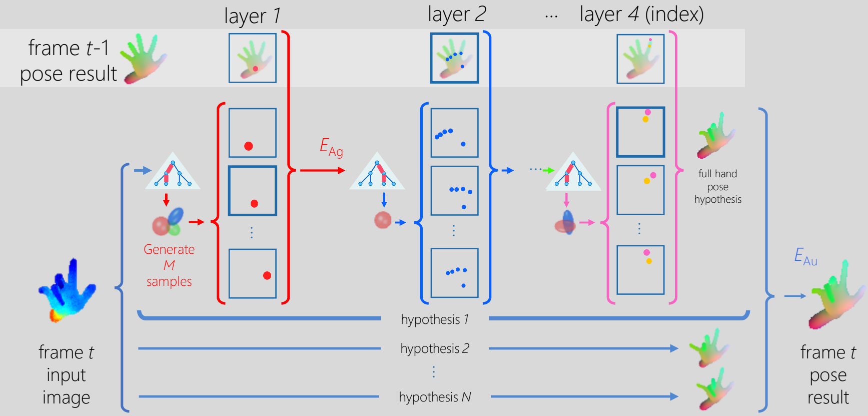


- Run-time speed 62.5 fps and ICVL dataset: LRF CVPR14oral/TPAMI16



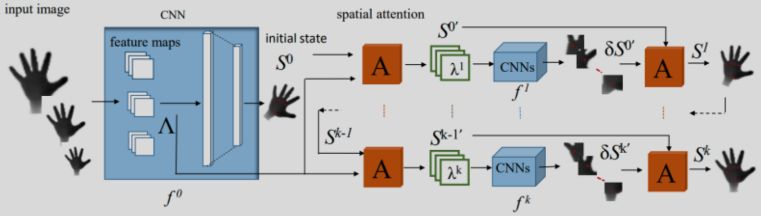


- Hierarchical sampling optimisation ICCV15oral/TPAMI18

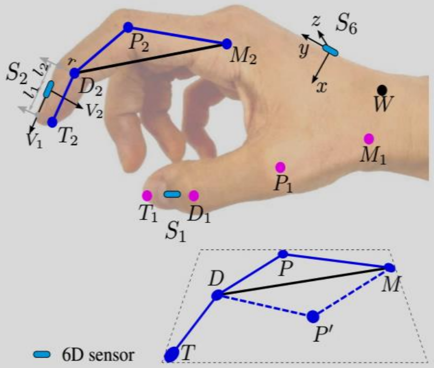


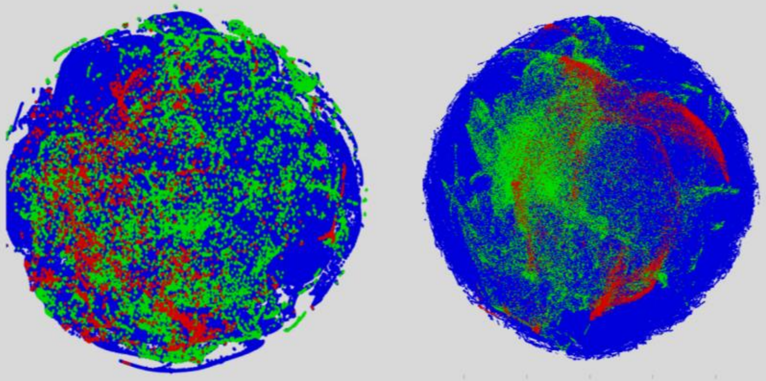
Advances so far

- Spatial attention deep net ECCV16



- BigHand2.2M benchmark CVPR17 [used by 116 unique institutions, 491 downloads]

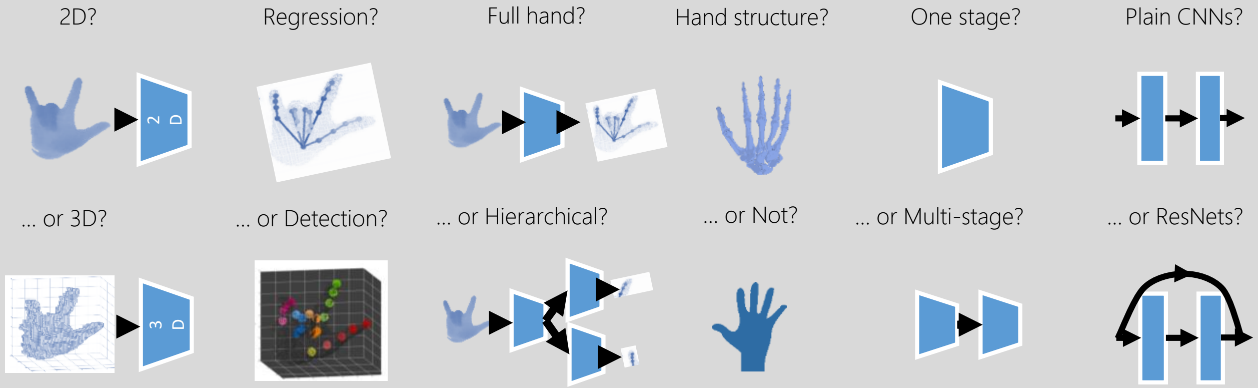


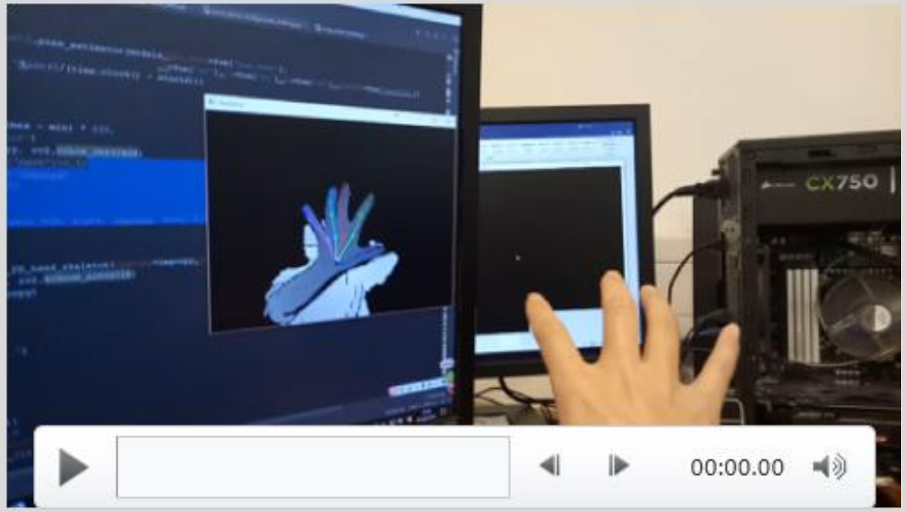


t-SNE embedding. *BigHand2.2M* (blue), ICVL (red), and NYU (green). global view point (left), articulation space in 25D (right)

- A comparative study CVPR18spotlight: of 17+ participating methods

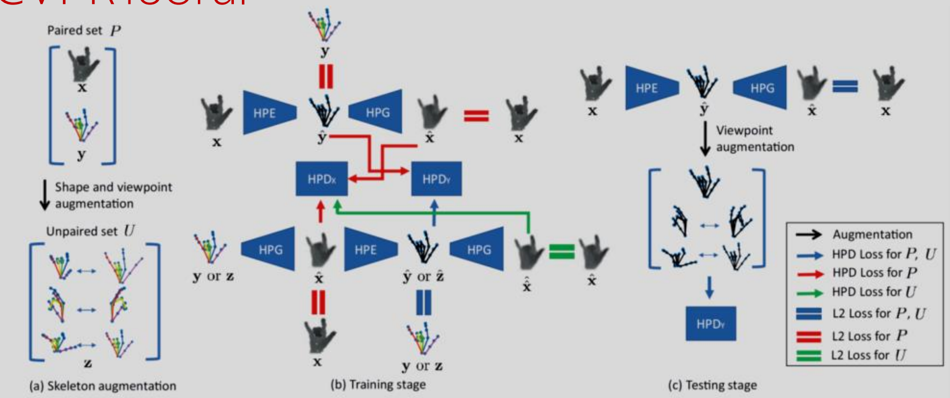
HANDS 2017
3rd International Workshop on Observing and Understanding Hands in Action



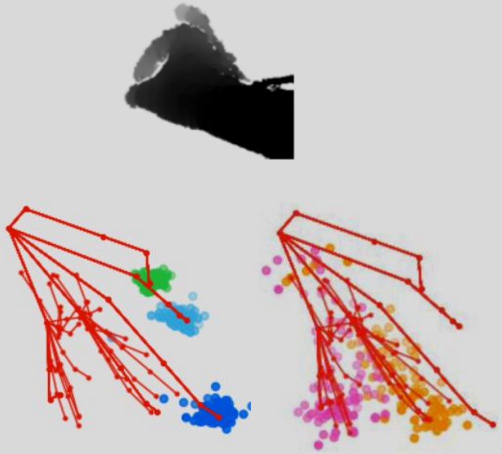


Advances so far

- Augmented Skeleton Space Transfer
CVPR18oral

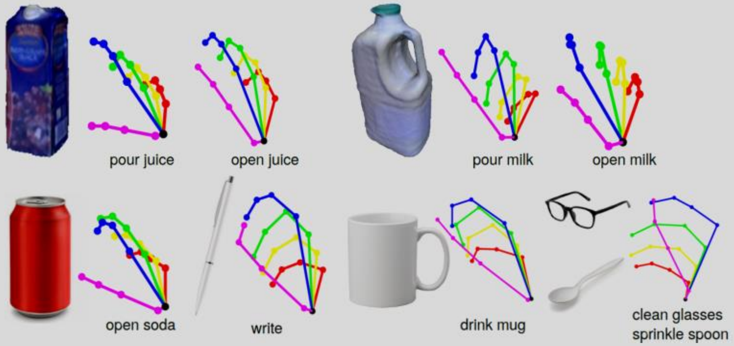


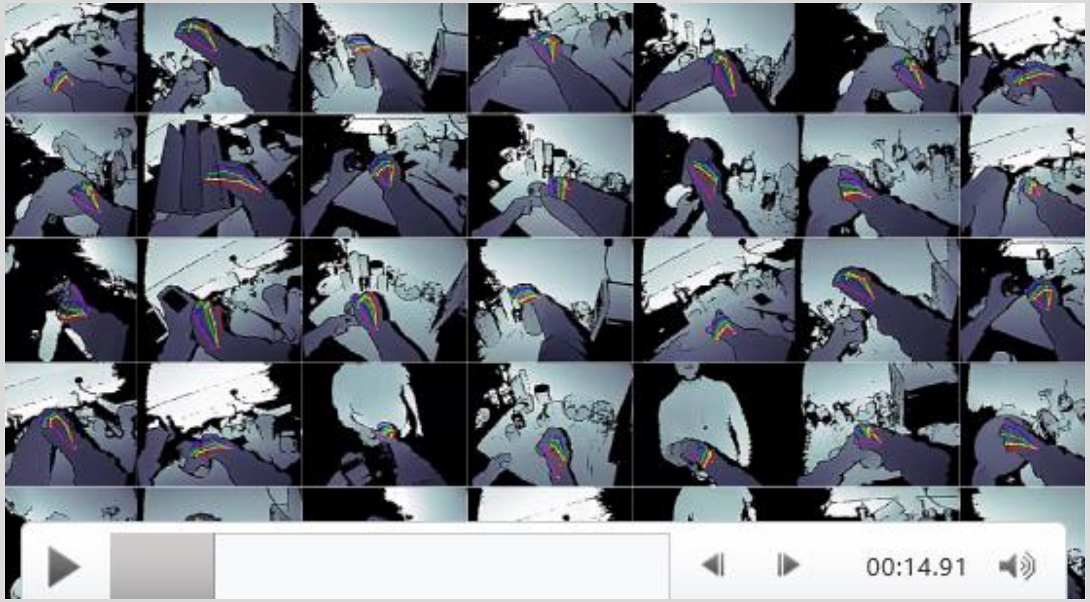
- Hierarchical Mixture Density Network ECCV18oral



- FPHA benchmark CVPR18:

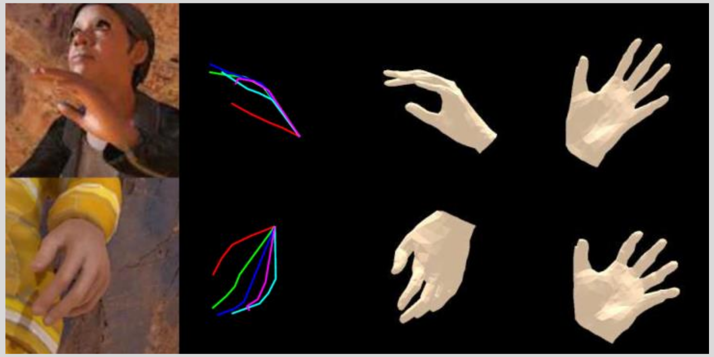
Egocentric views, Hand-object interaction, 3D hand/object pose, action labels, +1K sequences [used by 123 unique institutions: 236 licenses]



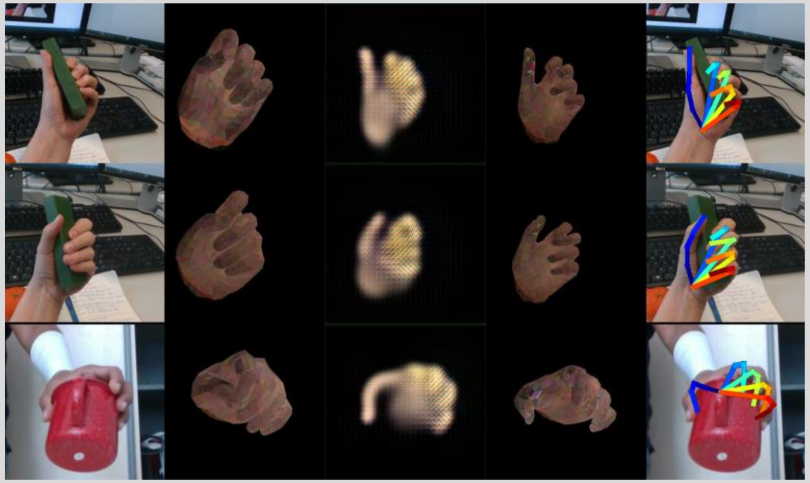


Advances so far

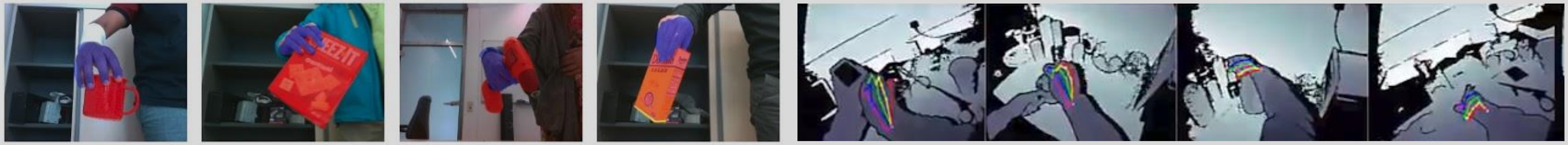
- RGB-based Dense 3D Hand Pose via Neural Rendering CVPR19



- Domain Adaptation via GAN and 3D Mesh Model CVPR2020 bestpaperfinalist



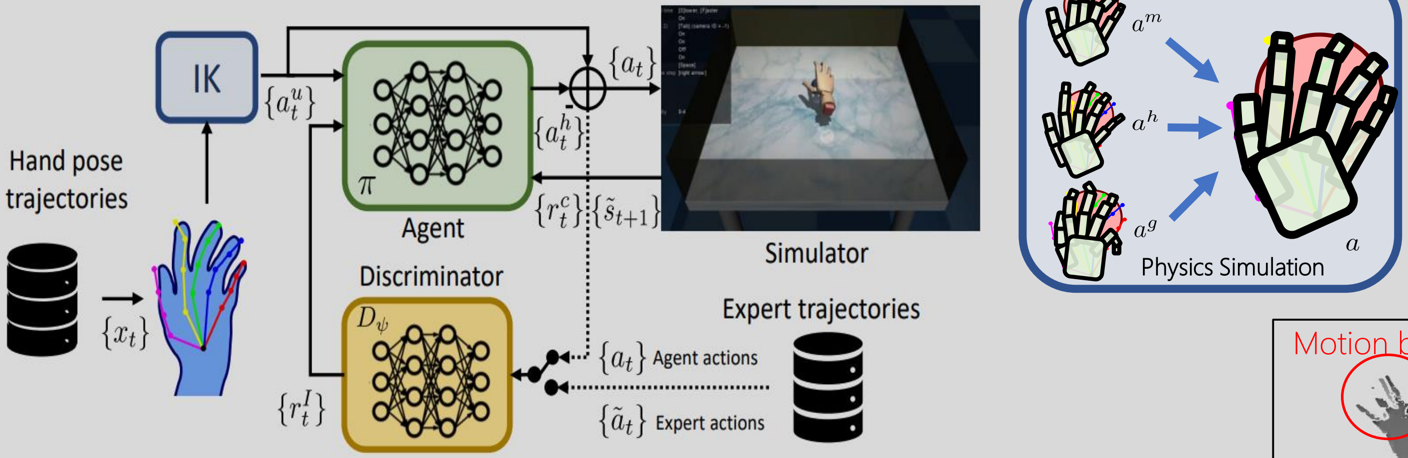
- HANDS19 Challenge ECCV2020: Hand-Object Interaction, Dense Pose, Fitted mesh models



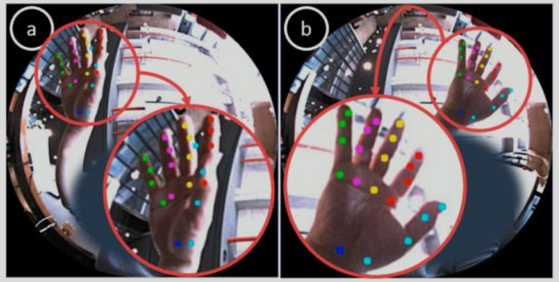
- We have co-organised 6 CVPR/ICCV/ECCV workshops (2015-2022) and 2 challenge on 3D hand pose.

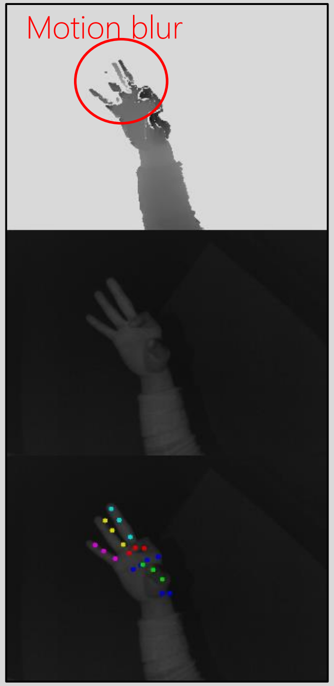
Applications

- Physics-Based Dexterous Manipulations
  IROS2020

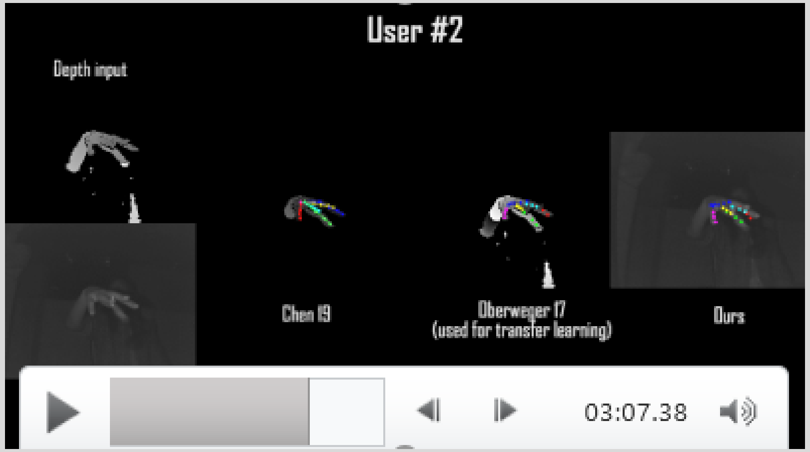


- DeepFisheye
  UIST2020





- With a Single Infrared Camera via Domain Transfer Learning
  ISMAR2020



Latest Innovations

- Im2Hands: Learning Attentive Implicit Representation of Interacting Two-Hand Shapes, CVPR2023

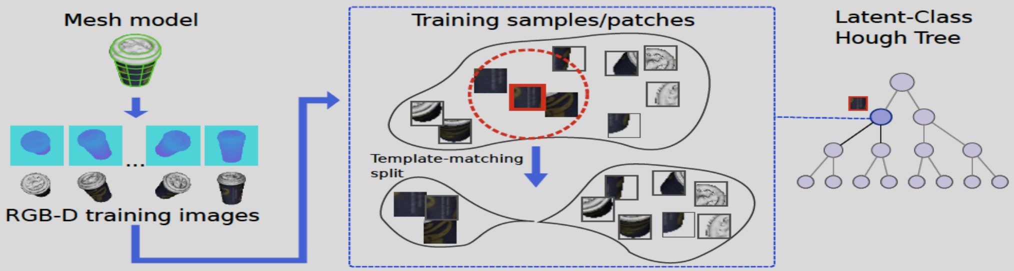
Input

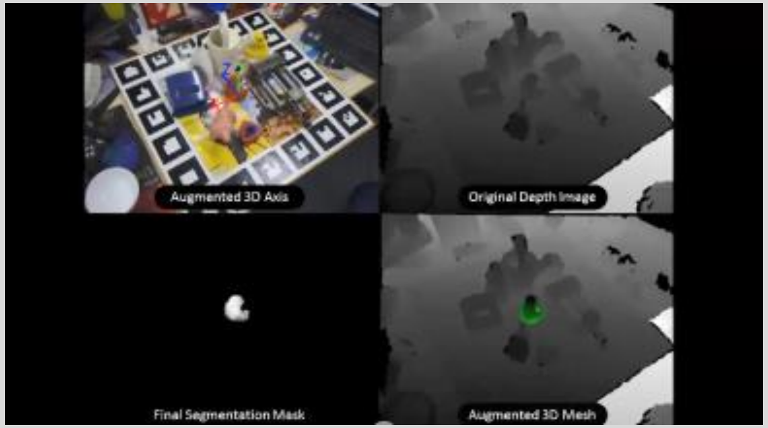
Input Alignment *Original Viewpoint* *Another Viewpoint*

Reconstruction

6D object pose

- Latent Hough Forest ECCV14/TPAMI16 : novel template-matching based splitting, one-class learning



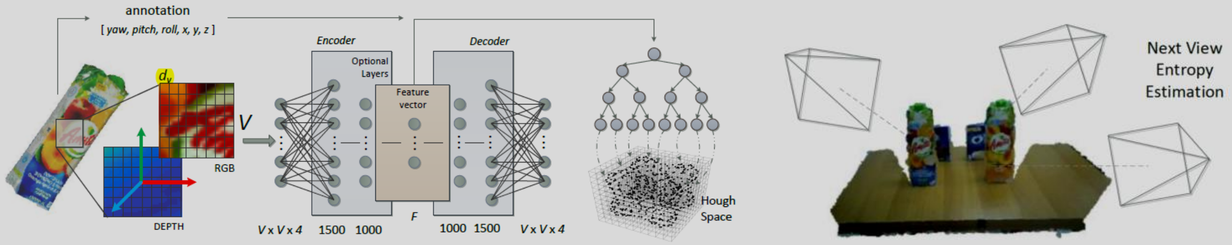


- Autonomous unfolding clothes ICRA14 bestpaperaward: regression forests, probabilistic active planning

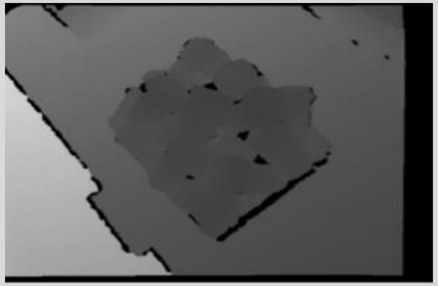
- Active Forest ECCV14 : multi-task learning, next-best view learning in RF

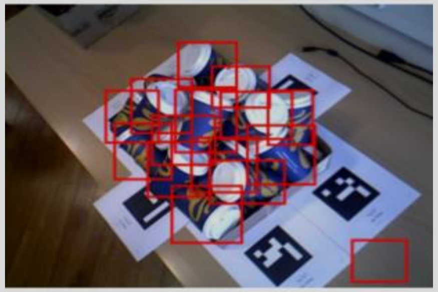
6D object pose

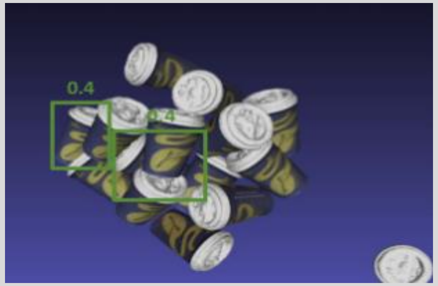
- 6D Object Detection and Next-Best-View Prediction in the Crowd CVPR16



- Multi-task Deep Network and Joint Registration BMVC18

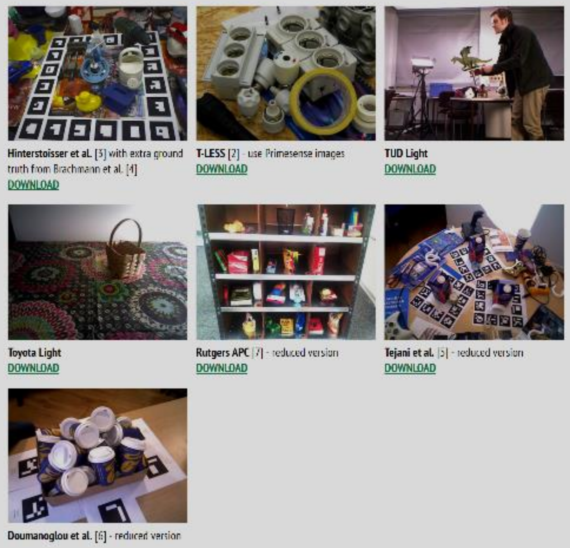




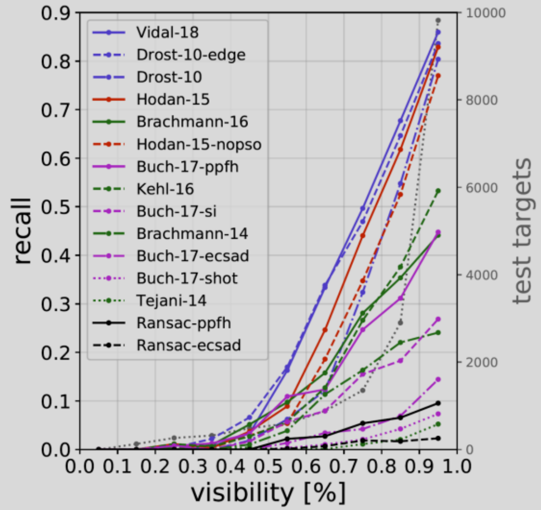


6D object pose

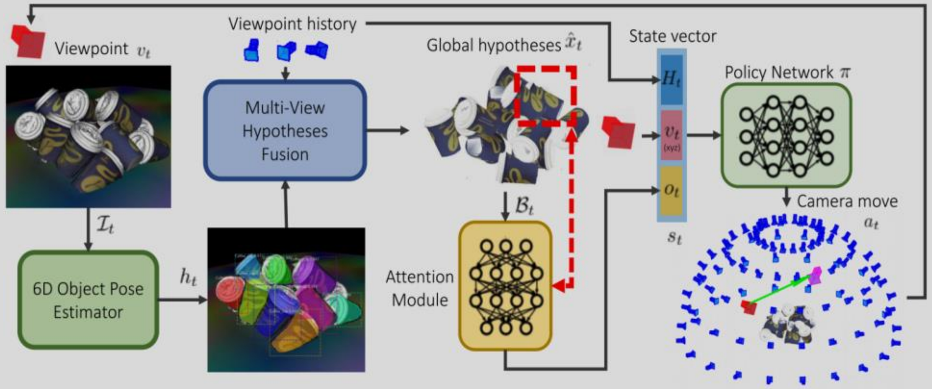
- BOP: Benchmark for 6D Object Pose Estimation ECCV18 89 object models, 62K test images, 110K test objects, 15+ participating methods



- We have co-organised 5 ICCV/ECCV workshops (2015-2019) and SIXD challenge at ICCV17 on 6D object pose.

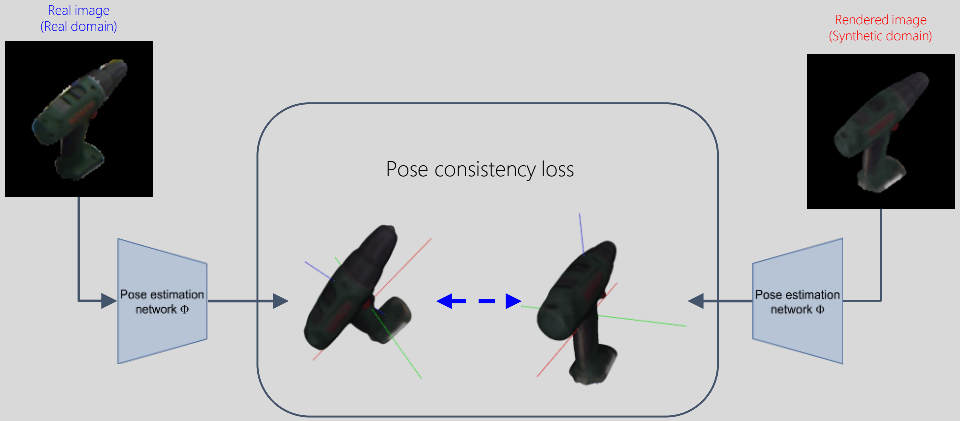


- Active 6D Pose Estimation by Deep Reinforcement Learning IROS2020

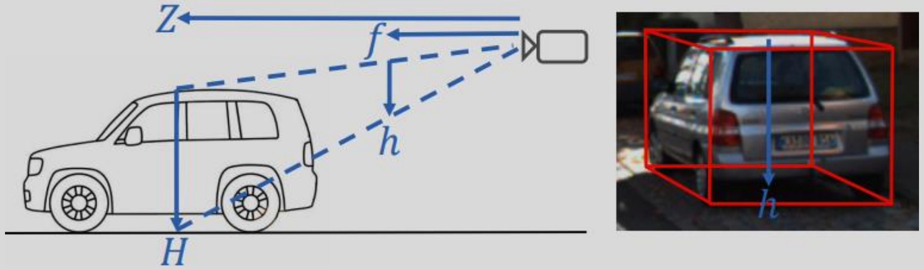


6D object pose

- Self-Supervised 6D Object Pose Estimation
  3DV2020: Pose Consistency, Warp-Alignment



- Geometry-based Distance Decomposition for Monocular 3D Object Detection, ICCV2021



- Distance-Normalized Unified Representation for Monocular 3D Object Detection, ECCV2020

Im2Hands: Learning Attentive Implicit Representation of Interacting Two-Hand Shapes

KAIST
Imperial College London

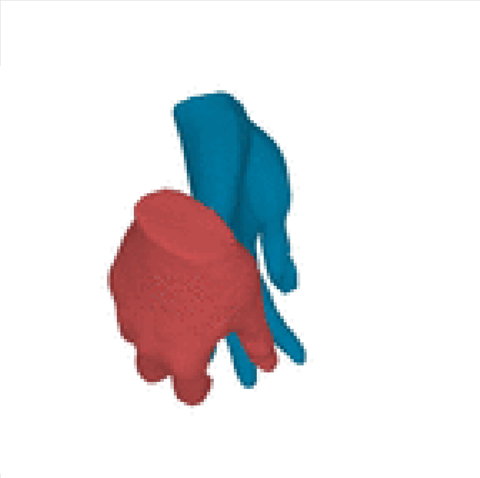
Jihyun Lee, Minhyuk Sung, Honggyu Choi, Tae-Kyun (T-K) Kim

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Input

Input Alignment

Original Viewpoint

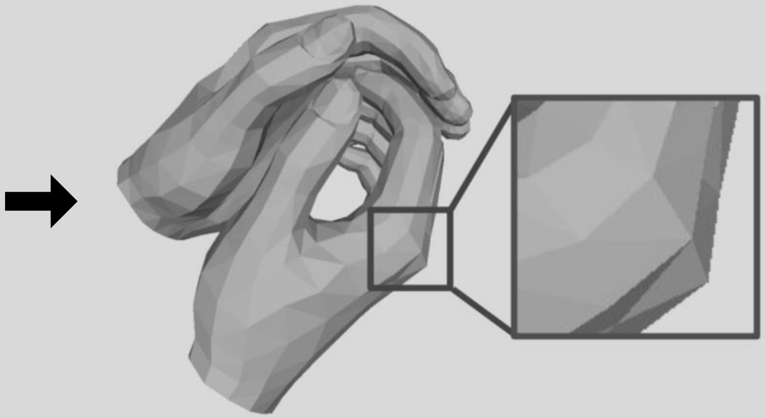


Another Viewpoint

Reconstruction

Motivation: Existing Hand Reconstruction Methods

Mesh-Based
Two-Hand Representations
[1, 2]



- Existing methods directly regress MANO [3] parameters *or* vertex positions of MANO meshes.
- However, they model hands with low-resolution meshes with a fixed MANO topology (|V| = 778).

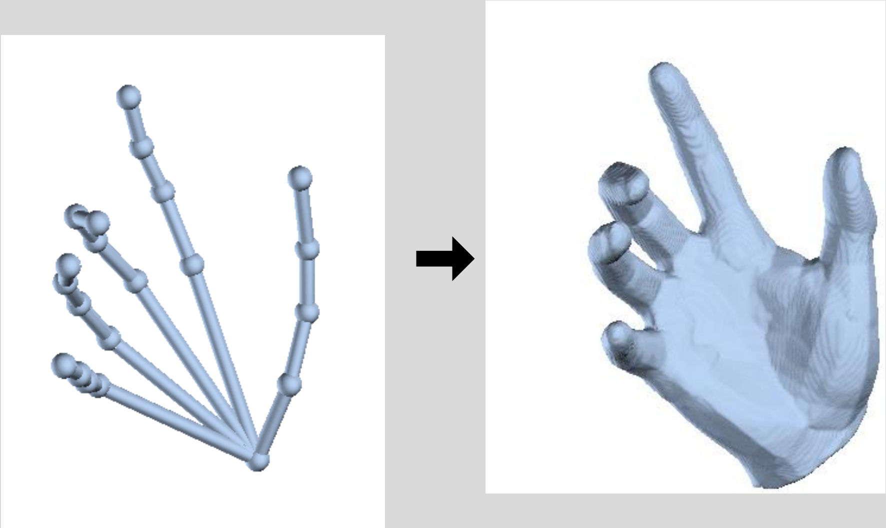
[1] Li *et al*. Interacting attention graph for single image two-hand reconstruction. In CVPR, 2022.
[2] Zhang *et al*. Interacting two-hand 3d pose and shape reconstruction from single color image. In ICCV, 2021.
[3] Romeo *et al*. Embodied Hands: Modeling and Capturing Hands and Bodies Together. In SIGGRAPH Asia, 2017.

Motivation: Existing Hand Reconstruction Methods

Implicit

Single-Hand Representation

[1]



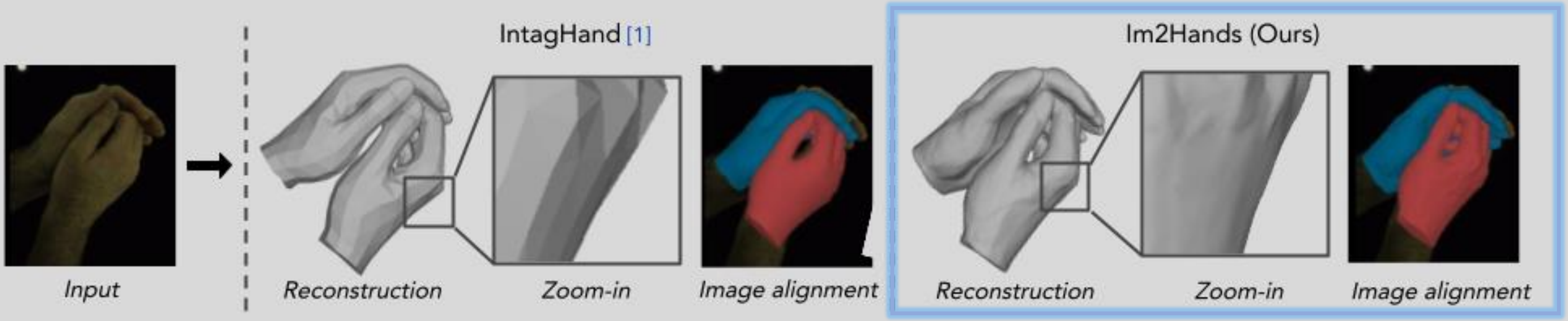
- Existing method predicts hand occupancy field conditioned on pose inputs.
- However, they cannot incorporate two-hand interaction contexts or perform image-based reconstruction.

[1] Karunratanakul *et al*. A skeleton-driven neural occupancy representation for articulated hands. In 3DV, 2021.

Proposed Method

Im2Hands: The first neural implicit representation of two interacting hands

- It learns resolution-free geometry of two-hands with high hand-to-hand and hand-to-image coherency.
- It can produce two-hand meshes with an arbitrary resolution.
- It does not require dense vertex correspondences or MANO parameter annotations for training.



[1] Li *et al.* Interacting attention graph for single image two-hand reconstruction. In CVPR, 2022.

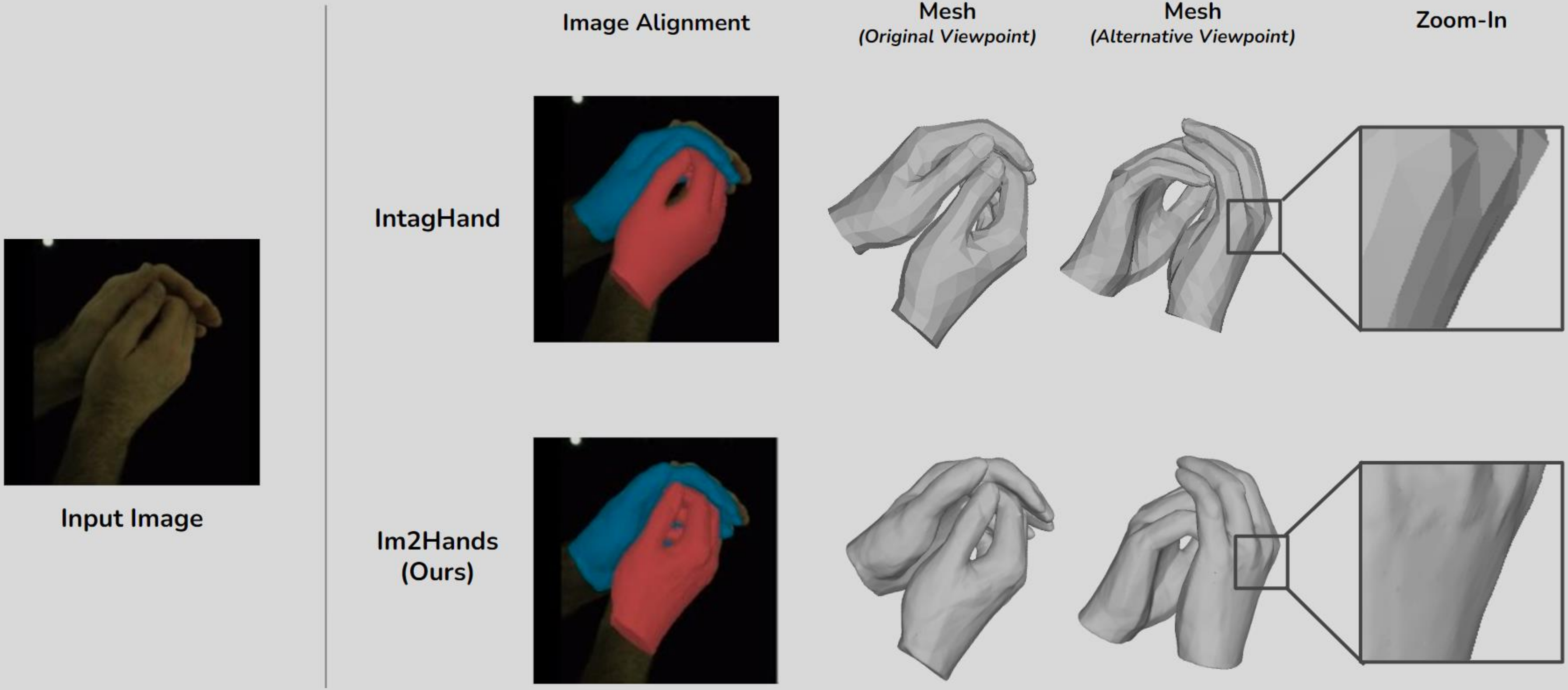


Image Alignment
Mesh
(Original Viewpoint)
Mesh
(Alternative Viewpoint)
Zoom-In
IntagHand
Input Image
Im2Hands
(Ours)

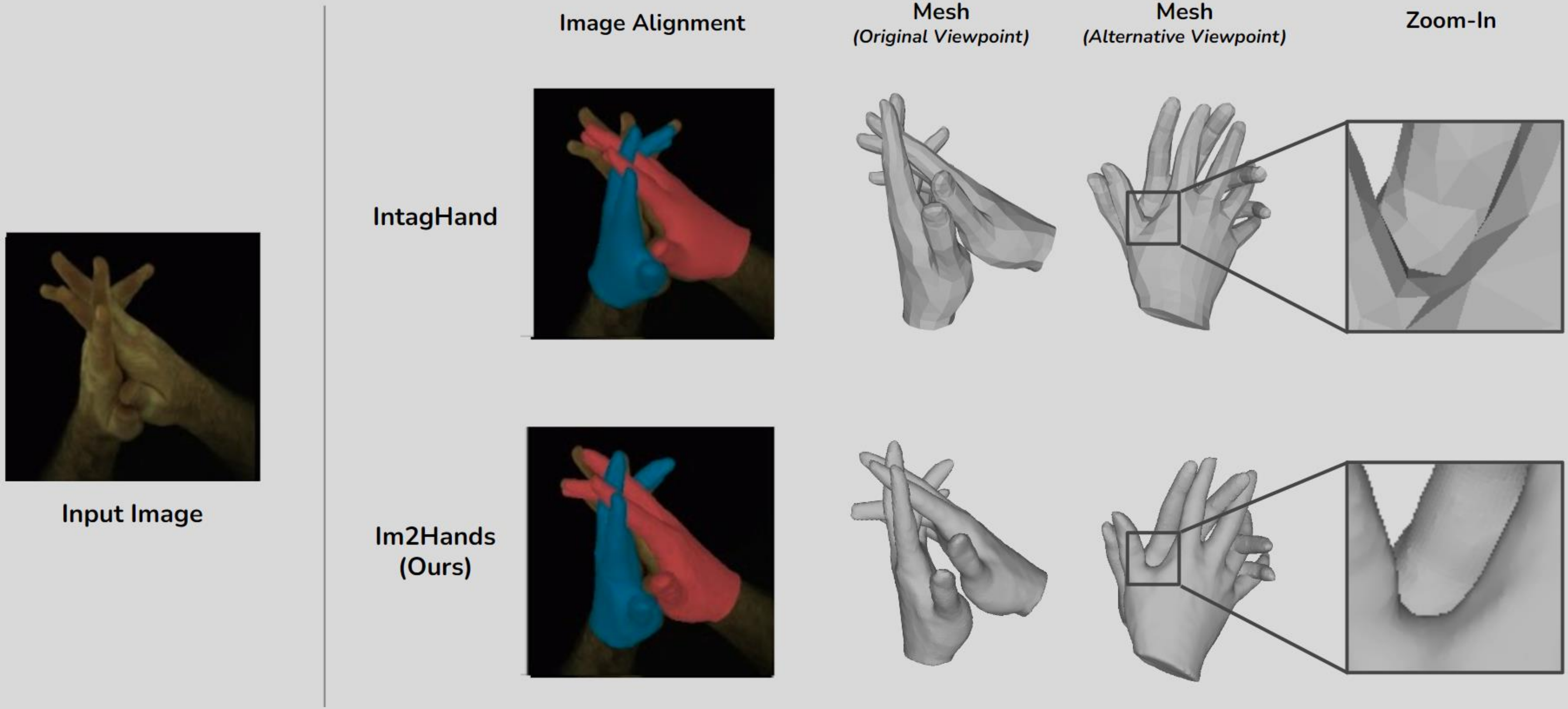


Image Alignment
Mesh
(Original Viewpoint)
Mesh
(Alternative Viewpoint)
Zoom-In
IntagHand
Input Image
Im2Hands
(Ours)

Method Overview (1/2)

- Our goal is to learn continuous 3D occupancy field of interacting two-hand geometry:

$$\mathcal{O}(x \,|\, \alpha,\, \beta) \rightarrow [o_l,\, o_r],$$

where $\mathcal{O}$ is a neural network that maps a query point $x \in \mathbb{R}^3$ to occupancy probabilities for each hand $o_l,\, o_r \in [0, 1]$

- Note that our two-hand occupancy is learned conditioned on a shape observation α and a pose β observation n ,

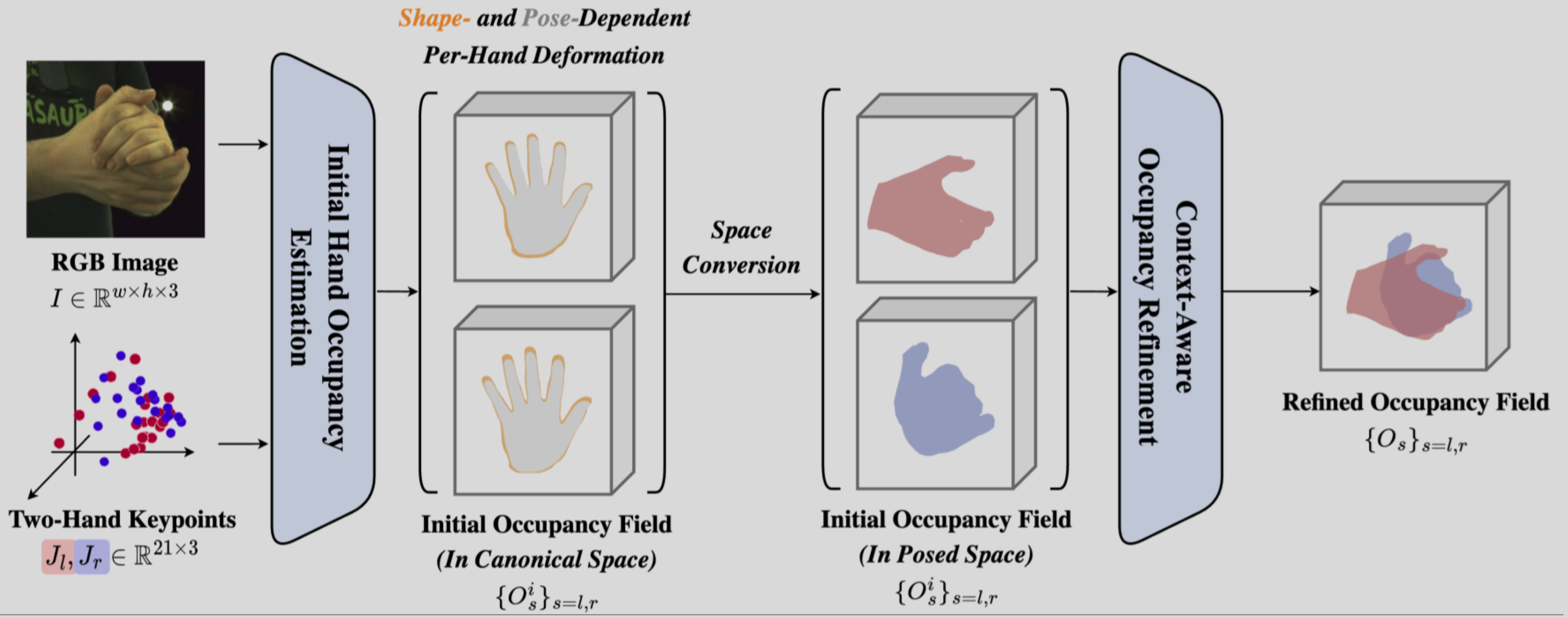
  which are represented as an RGB image and sparse two-hand 3D keypoints, respectively.

Method Overview (2/2)

- To effectively handle the shape complexity and interaction context between two hands, we propose two novel attention-based modules that performs:

  1) Initial occupancy estimation in the hand canonical space, and

  2) Interaction context-aware occupancy refinement in the original posed space.



Initial Hand Occupancy Estimation (1/2)

- We first estimate initial occupancy probabilities for each hand in the **canonical space**.

- Motivated by existing articulated implicit functions[1, 2], we train part occupancy networks to predict occupancy values for each bone transformed to canonical pose:

$$\mathcal{I}(x \,|\, I,\, J) = \max_{b=1,\,\ldots,\,B} \{\bar{\mathcal{H}}_b(\mathbf{T}_b x,\; f_b^{\phi}, f_x^{\phi}, f_b^{\omega})\}$$

where $\bar{\mathcal{H}}_b$ is an MLP-based part occupancy network for bone *b*, and $\mathbf{T}_b$ is the canonicalization matrix for bone *b* computed using the input 3D hand keypoints.

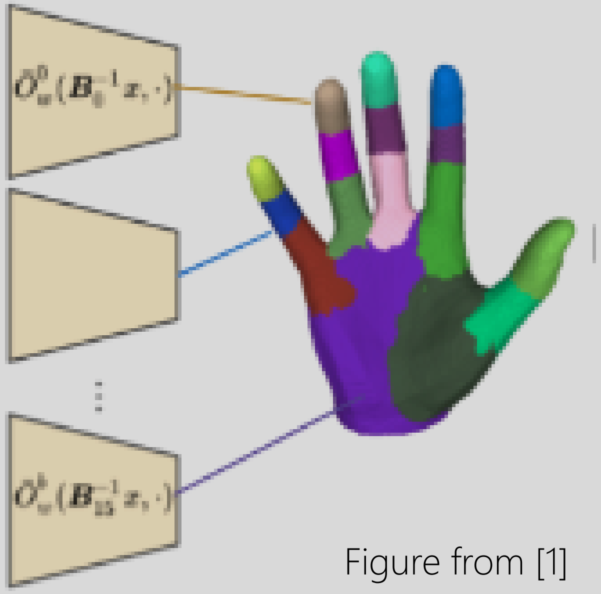


Figure from [1]

[1] Karunratanakul *et al.* A skeleton-driven neural occupancy representation for articulated hands. In 3DV, 2021.
[2] Deng, *et al.* Neural articulated shape approximation. In ECCV, 2020.

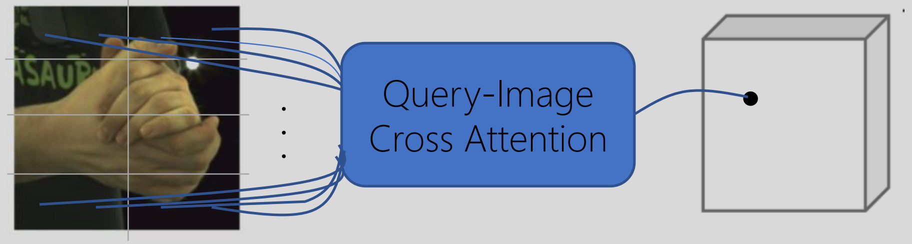
Initial Hand Occupancy Estimation (2/2)

- When estimating part occupancies, we use hand features for modeling shape- and pose- dependent deformations.

$$\mathcal{I}(x \mid I,\, J) = \max_{b=1,\,\dots,\,B} \{\bar{\mathcal{H}}_b(\underbrace{\mathbf{T}_b x}_{\text{Canonicalized query}},\, \underbrace{f_b^{\phi},\, f_x^{\phi},\, f_b^{\omega}}_{\text{Shape and pose features}})\}$$

- Unlike existing articulated implicit functions[1, 2] that use **bone-wise** global features (*i.e.*, bone length feature f_b^{ϕ} and canonicalization matrix feature f_b^{ω}), we propose to use additional **query-wise** shape feature f_x^{ϕ} to recover fine-grained shape details observed from an image.

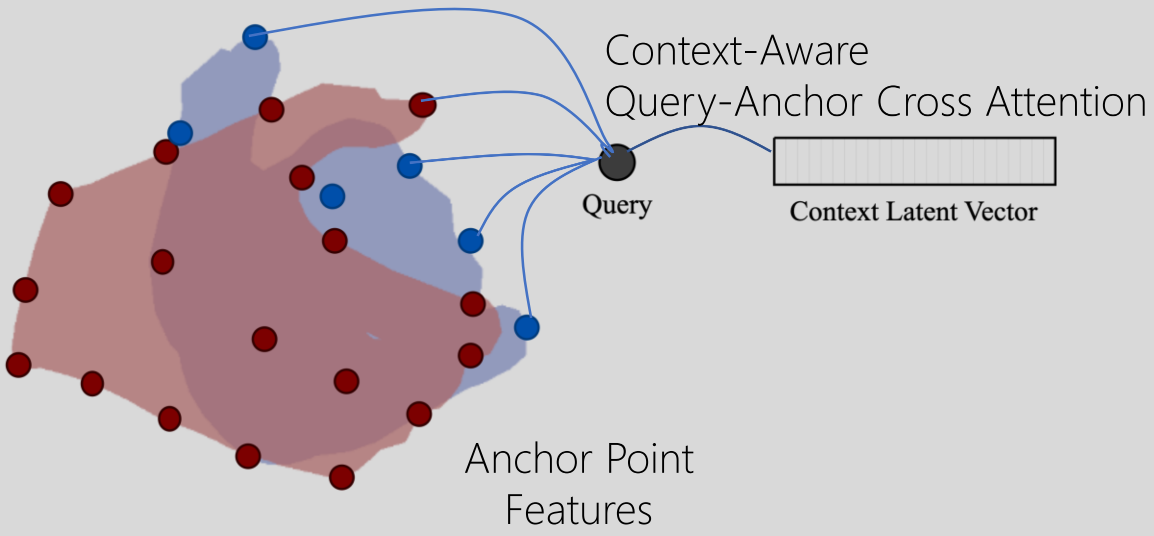
- To this end, we introduce **query-image cross attention** to extract a per-query feature while attending to image regions informative for estimating occupancy at the query *x*.



[1] Karunratanakul *et al.* A skeleton-driven neural occupancy representation for articulated hands. In 3DV, 2021.
[2] Deng, *et al.* Neural articulated shape approximation. In ECCV, 2020.

Context-Aware Occupancy Refinement

- We additionally propose to perform two-hand occupancy refinement in the **original posed space**.
- To encode the initial geometry of two hands, we represent them as anchored feature cloud (*i.e.*, feature vectors of queries evaluated to be on surface by our initial occupancy network).
- We then apply cross-attention between (1) a query, (2) anchor features, and (3) a context latent vector (extracted from global features of initial two-hand shape and the input image) to estimate a refined occupancy value.



Input Keypoint Refinement (Optional)

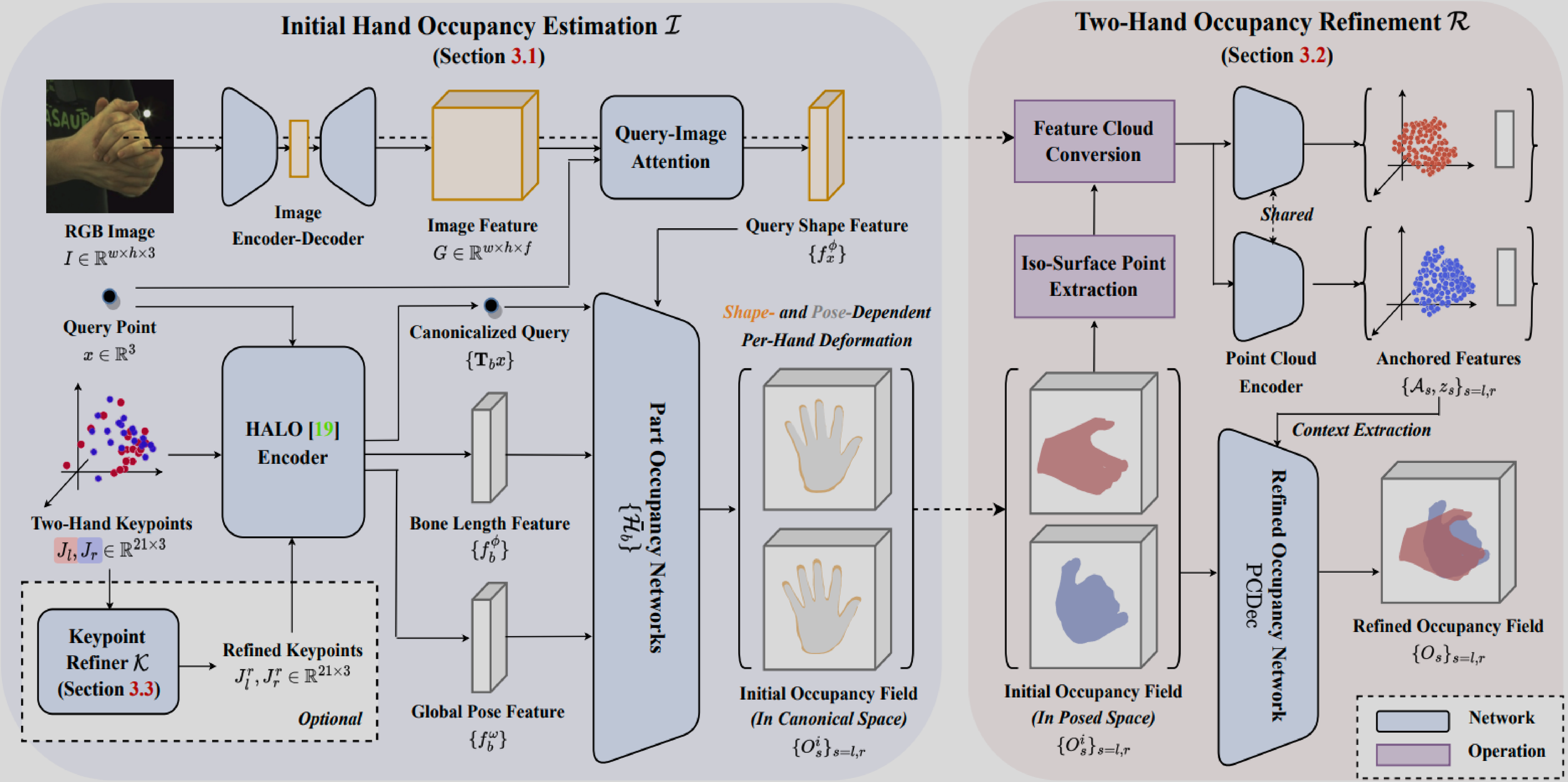
- We further consider image-based two-hand reconstruction using Im2Hands, where no ground truth hand keypoints are available as inputs.
- To enable robust shape reconstruction from keypoints predicted from an off-the-shelf image-based two-hand keypoint estimator, we introduce an optional keypoint refinement module that can alleviate input keypoint noise.

$$\mathcal{K}(J, I) = \mathrm{MSA}([\mathrm{GCN}(\mathrm{KptEnc}(J)),\ \mathrm{ImgEnc}(I)]),$$

where the refined keypoints are estimated via multi-headed attention (MSA) between keypoint features encoded using a graph convolutional network $\mathrm{GCN}(\mathrm{KptEnc}(J))$ and input image features $\mathrm{ImgEnc}(I)$

.

Architecture details



Loss Functions

- For our initial occupancy network and context-aware occupancy refinement network, we use MSE loss that measures deviation between the ground truth and the predicted occupancy values.

- For context-aware occupancy refinement network, we additionally use **penetration loss** to penalize the refined two-hand occupancy values that are estimated to be occupied in both hands at the same query position.

$$\mathcal{L}_{pen} = \frac{1}{|\mathcal{X}|} \sum_{(I,J)\in\mathcal{X}} \sum_{x\in\mathcal{P}} \mathcal{R}_l(x|I,\, J) \cdot \mathcal{R}_r(x|I,\, J),$$

$$\textbf{where } \mathcal{R}_l(\cdot) > 0.5 \textbf{ and } \mathcal{R}_r(\cdot) > 0.5.$$

In the above equation, $\mathcal{X}$ is a set of training samples, $\mathcal{P}$ is a set of training query points, and $\mathcal{R}_l$ and $\mathcal{R}_r$ are functions that return the refined occupancy probabilities for left and right hand, respectively.

Two-Hand Reconstruction Results (1/2)

- Results on InterHand2.6M dataset [1] using image and ground truth keypoint inputs

Method	Inputs	IoU (%) ↑	CD (mm) ↓
Two-Hand-Shape-Pose [42]	**I, L**	54.8	5.51
IntagHand [23]	**I, L**	67.0	3.88
HALO [18]	**J**	74.7	2.62
HALO (modified) [18]	**I, J**	75.8	2.51
Im2Hands (Ours)	**I, J**	**77.8**	**2.30**

Two-Hand Reconstruction Results (2/2)

- Results on InterHand2.6M dataset [1] using image inputs only

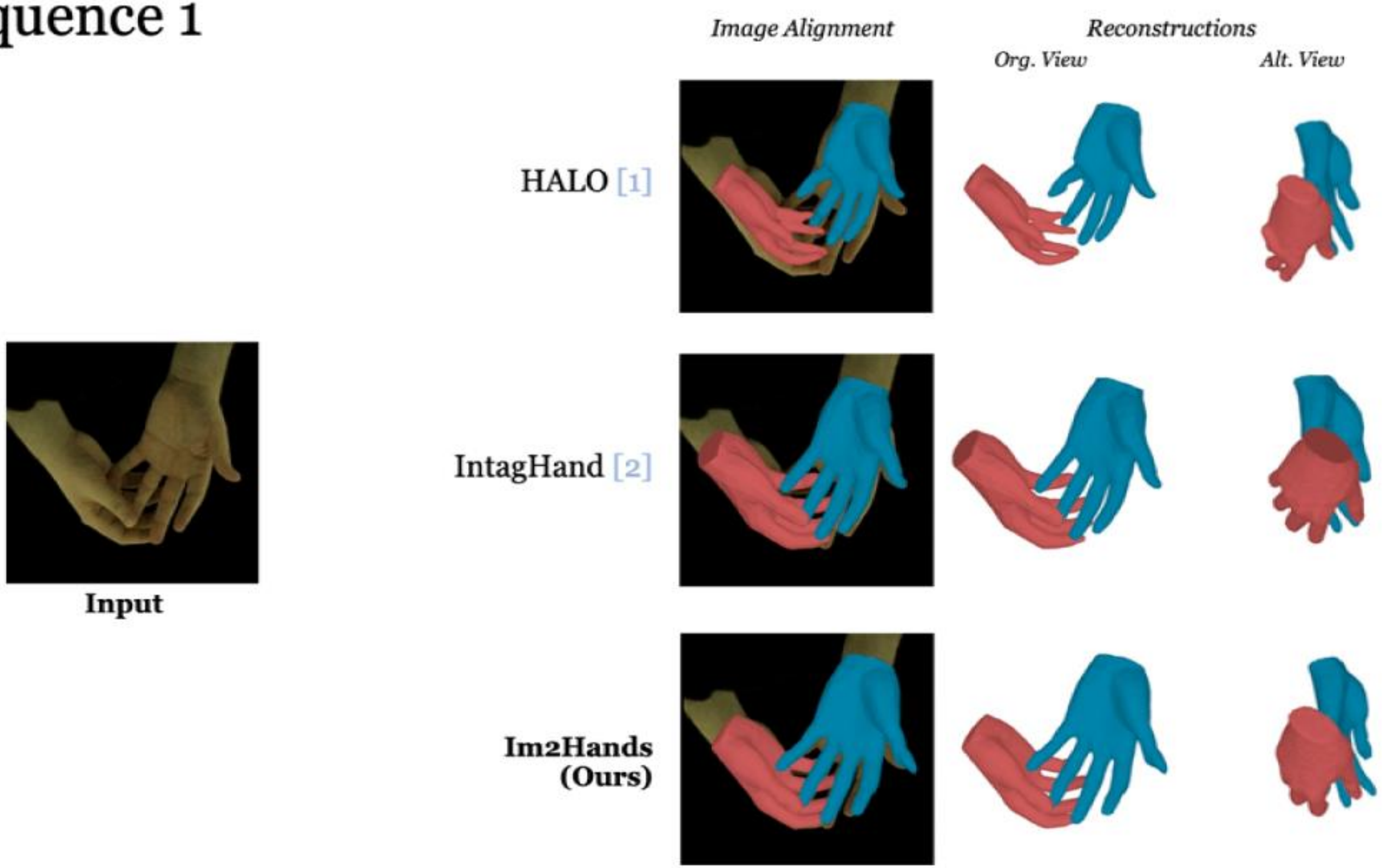
Shape Reconstruction Results

Method	IoU (%) ↑	CD (mm)↓
Two-Hand-Shape-Pose [42]	48.4	6.09
IntagHand [23]	59.0	4.69
DIGIT [11] + HALO [18]	45.1	7.64
IntagHand [23] + HALO [18]	53.8	5.38
DIGIT [11] + Im2Hands (Ours)	**59.4**	**4.75**
IntagHand [23]+ Im2Hands (Ours)	**62.1**	**4.35**

Keypoint Refinement Results

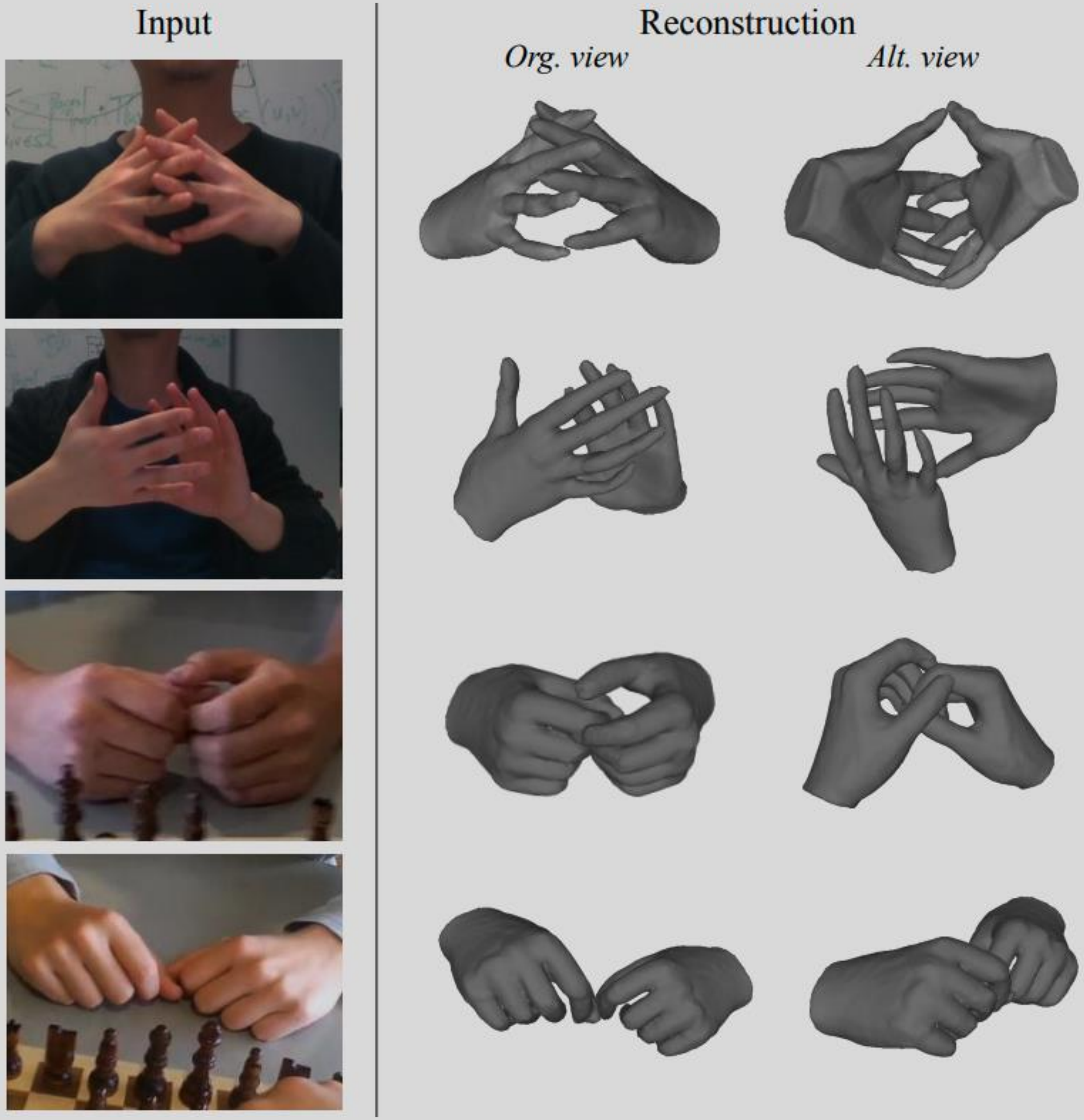
Method	MPJPE (mm) ↓
DIGIT [11]	16.75
DIGIT [11] + $\mathcal{K}$ (Ours)	**10.70**
IntagHand [23]	10.13
IntagHand [23] + $\mathcal{K}$ (Ours)	**9.68**

Sequence 1



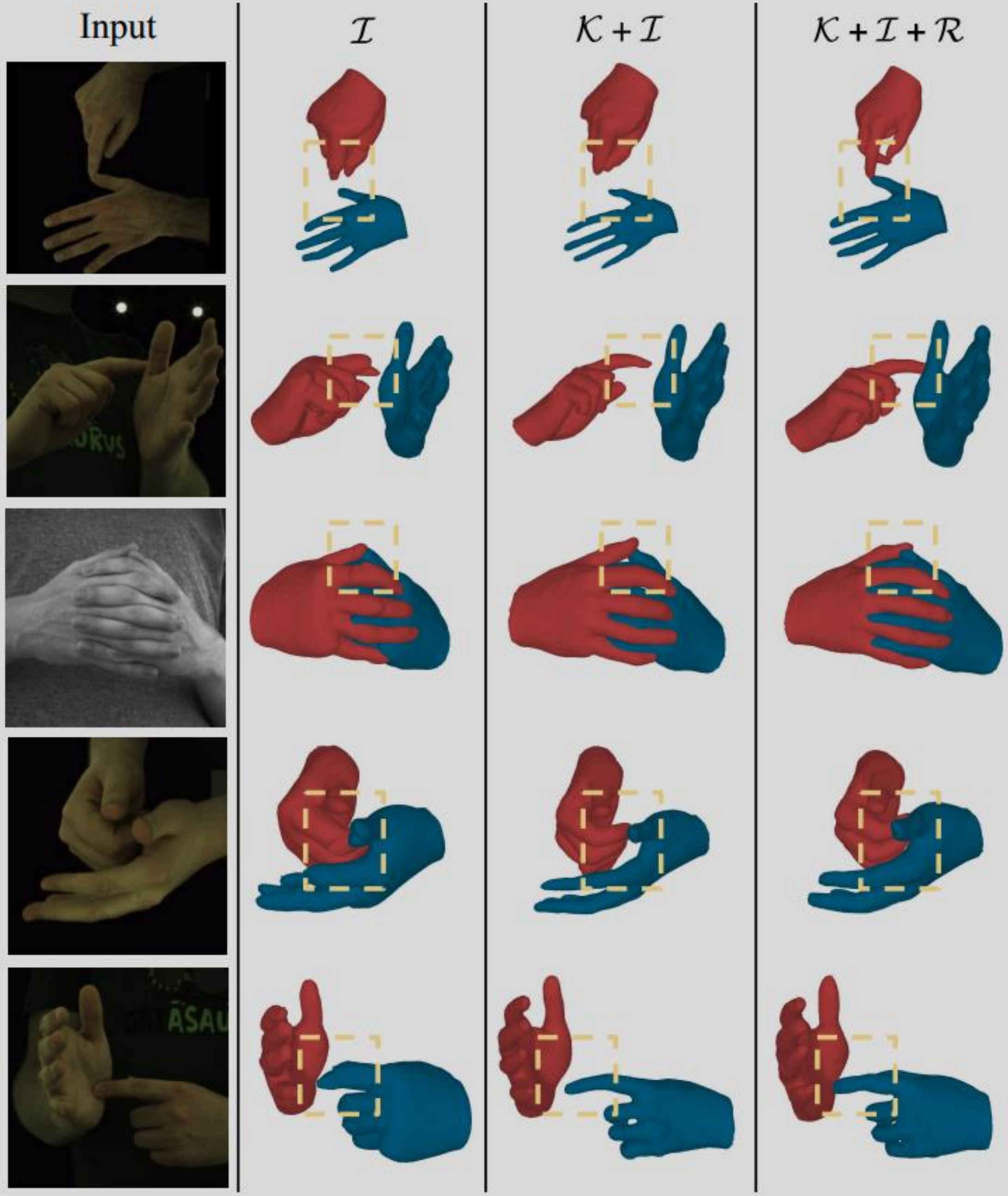
[1] Karunratanakul *et al.* A skeleton-driven neural occupancy representation for articulated hands. In 3DV, 2021.

Generalizability test on RGB2Hands and EgoHands



- Top two rows show examples from RGB2Hands and bottom two rows show examples from EgoHands

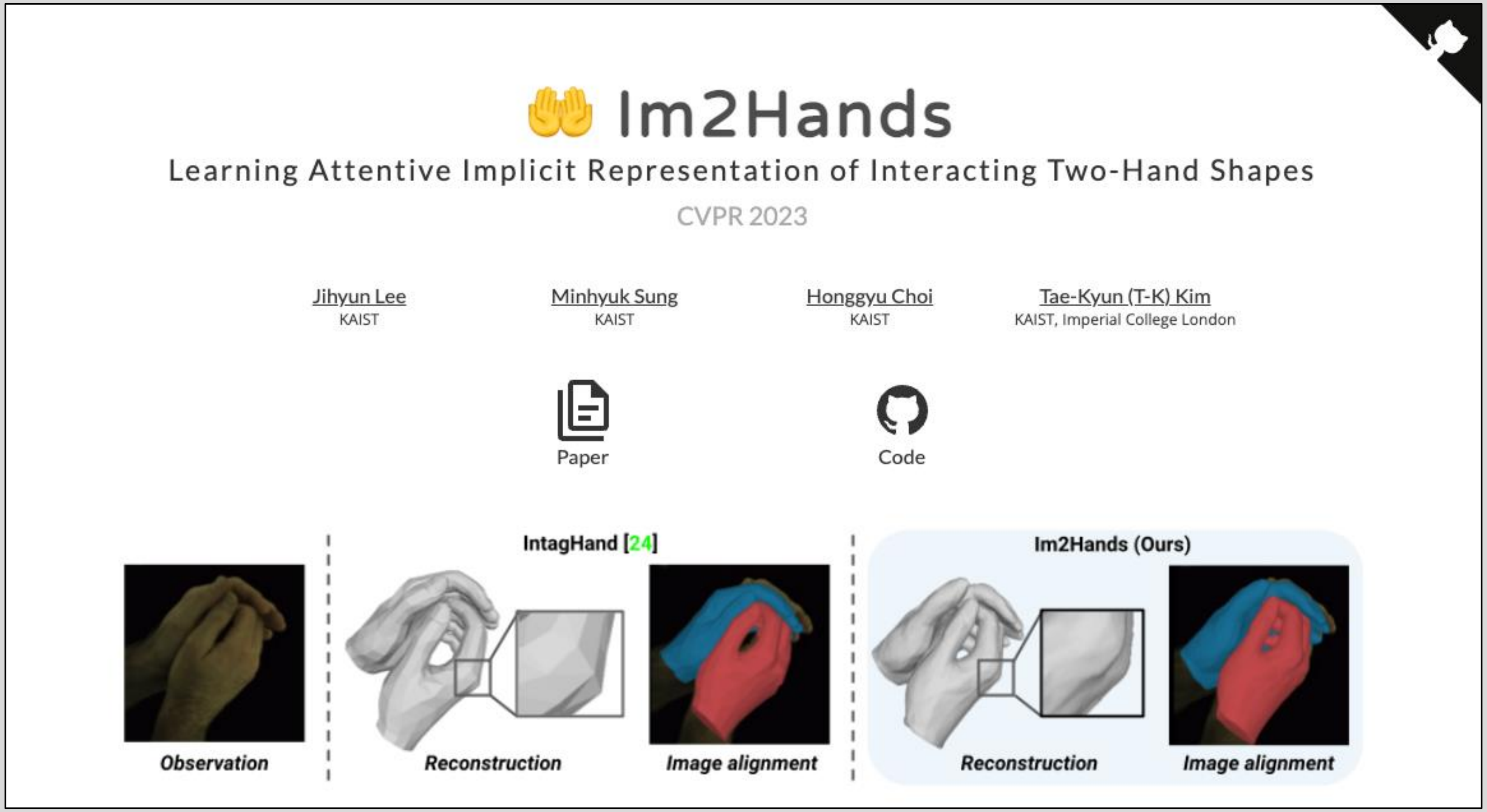
Qualitative ablation study on InterHand2.6M



- I, R and K denotes Initial Hand Occupancy Network, Two-Hand Occupancy Refinement Network, and Input Keypoint Refinement Network, respectively.

Project Website & Code

https://jyunlee.github.io/projects/implicit-two-hands

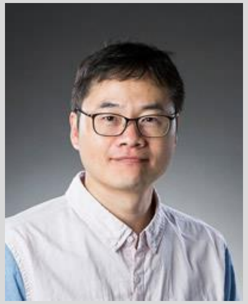


Weakly-supervised Domain Adaptation via GAN and Mesh Model for Estimating 3D Hand Poses Interacting Objects

(CVPR20 oral, best paper finalist)

Seungryul Baek

Imperial College London



Kwang In Kim

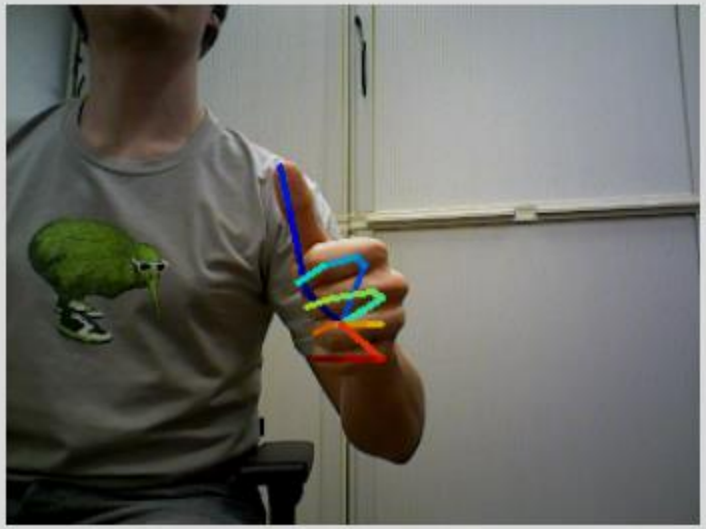
UNIST
ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

Tae-Kyun Kim

KAIST

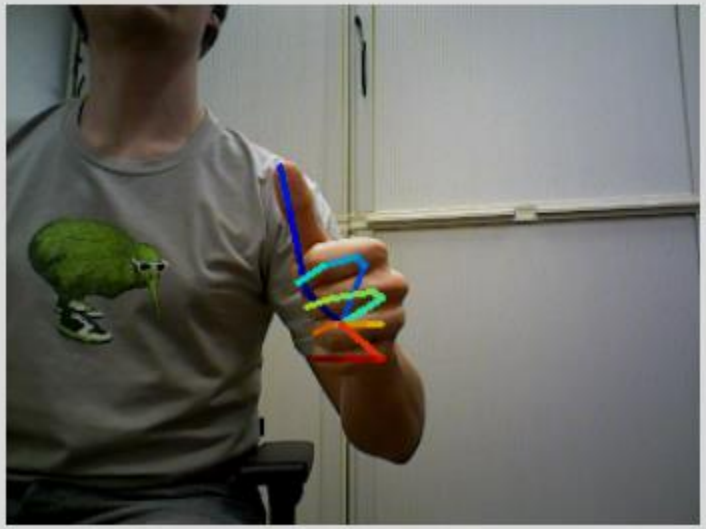
Imperial College London

Objective

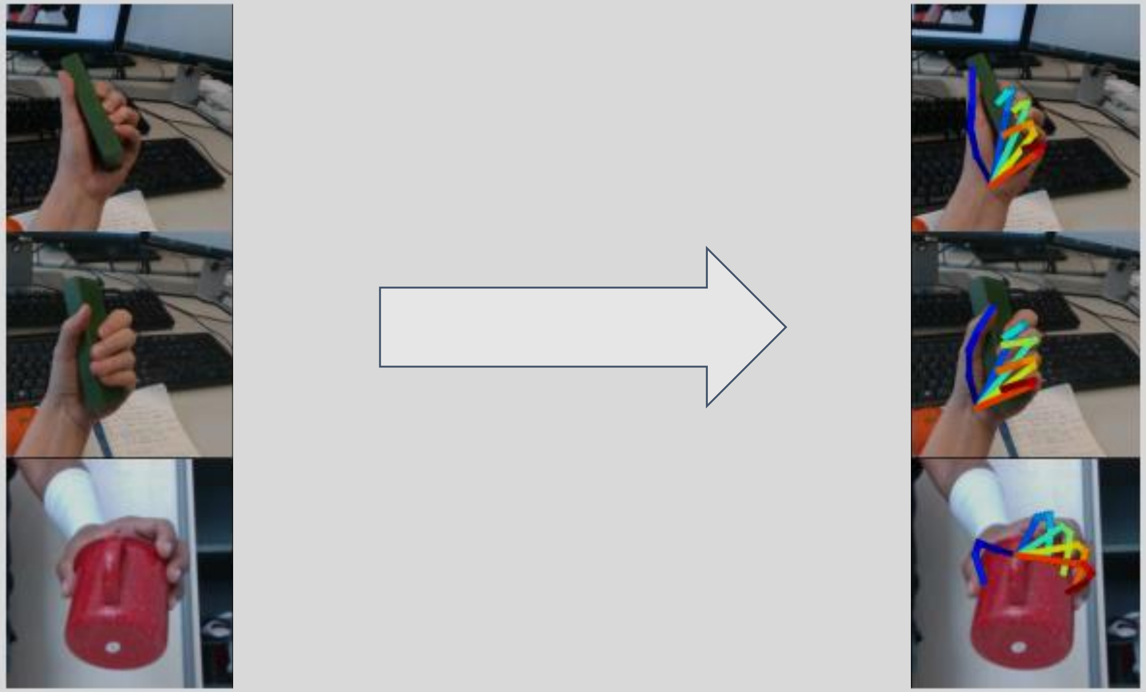


Hand pose estimation for Hand-only scenario.

Objective



Hand pose estimation for Hand-only scenario.



Hand pose estimation from single RGB images under hand object interaction (HOI) scenario.

Related works

Most previous works tackle the HOI problem by collecting a **new dataset.**

Dexter+Object (ECCV’16)

EgoDexter (ICCV’17)

[**Real** dataset – Few in quantity, inaccurate/insufficient 3D annotation]

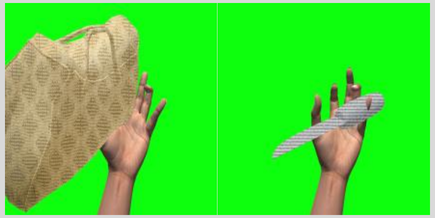
Related works

Most previous works tackle the HOI problem by collecting a **new dataset.**

Dexter+Object (ECCV'16)

Obman (CVPR'19)

EgoDexter (ICCV'17)



SynthHands (ICCV'17)

[**Real** dataset]

[**Synthetic** dataset – gap to real dataset]

Related works

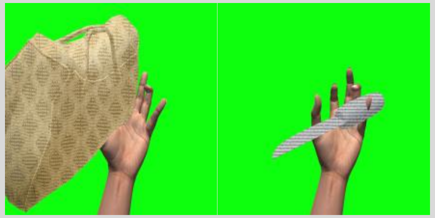
Most previous works tackle the HOI problem by collecting a **new dataset.**

Dexter+Object (ECCV'16)

Obman (CVPR'19)

FPHA (CVPR'18)

EgoDexter (ICCV'17)



SynthHands (ICCV'17)

GANerated (CVPR'18)

[**Real** dataset]

[**Synthetic** dataset]

[Using **GAN/Sensors** – Still limited]

Related works

Most previous works tackle the HOI problem by collecting a **new dataset.**

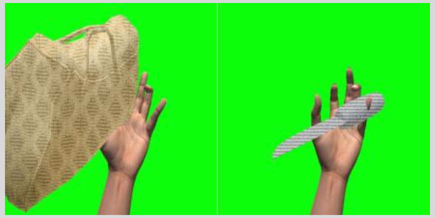
Dexter+Object (ECCV'16)

Obman (CVPR'19)

FPHA (CVPR'18)

HO3D (ArXiv'19)

EgoDexter (ICCV'17)



SynthHands (ICCV'17)

GANerated (CVPR'18)

FreiHand (ICCV'19)

[**Real** dataset] [**Synthetic** dataset] [Using **GAN/Sensors**] [Iterative **3D model** fitting – #sample]

Challenges

STB (ICIP'17)

RHD (ICCV'17)

[Real and synthetic Hand-only data]

[Diverse objects]

Challenges

STB (ICIP'17)

RHD (ICCV'17)

[Real and synthetic Hand-only data]

[Diverse objects]

HO3D (ArXiv'19)
Real, <15000 frame, 6 objects.

FreiHand (ICCV'19)
Real, <3000 frame, <30 objects.

Key Idea

Image-level supervision with **HOI images**:

Real HOI

Synthetic HOI

We exploit only easily available **Real and synthetic hand-only data**, **Real HOI images with segmentation masks**, **Synthetic hand-only and HOI image pairs.**

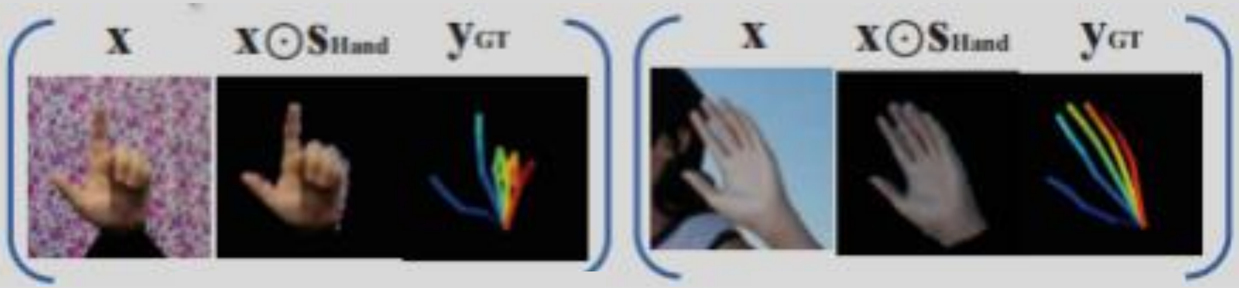
Key Idea

Image-level supervision with **HOI images**:

Real HOI

Synthetic HOI

3D supervision with **Hand-only data**:



Real Hand-only

Synthetic Hand-only

We exploit only easily available **Real and synthetic hand-only data**, **Real HOI images with segmentation masks**, **Synthetic hand-only and HOI image pairs.**

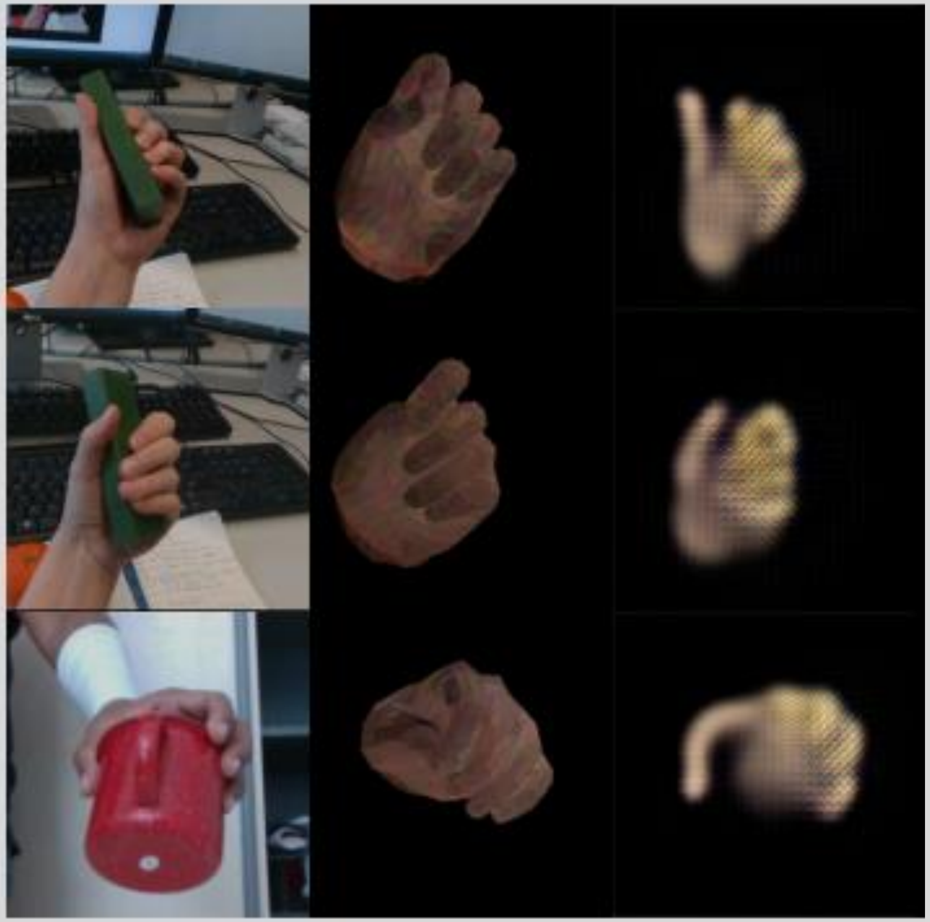
Key Idea

We gradually synthesize hand-only images using Mesh model and GAN with a weak image-level supervision.

Key Idea

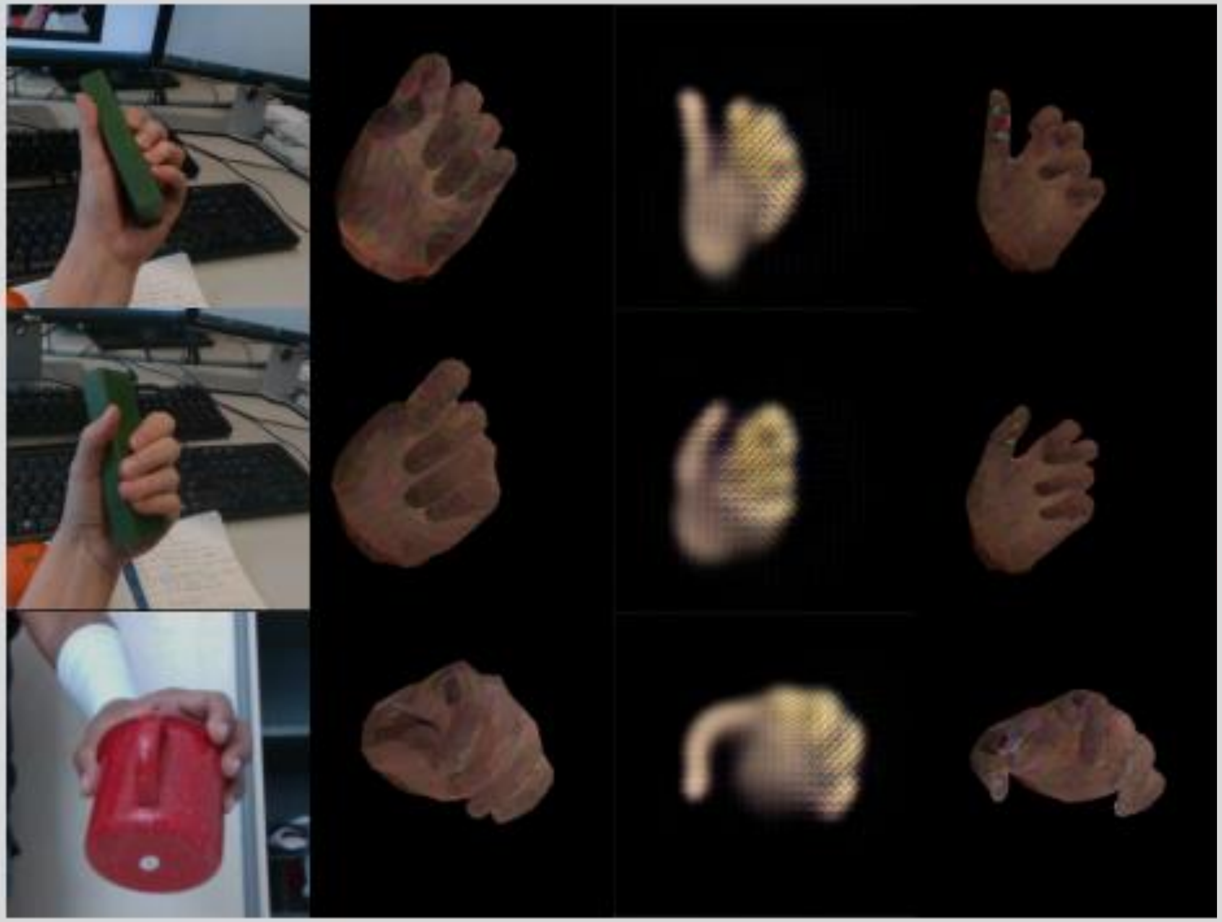
We gradually synthesize hand-only images using Mesh model and GAN with a weak image-level supervision.

Key Idea



We gradually synthesize hand-only images using Mesh model and GAN with a weak image-level supervision.

Key Idea

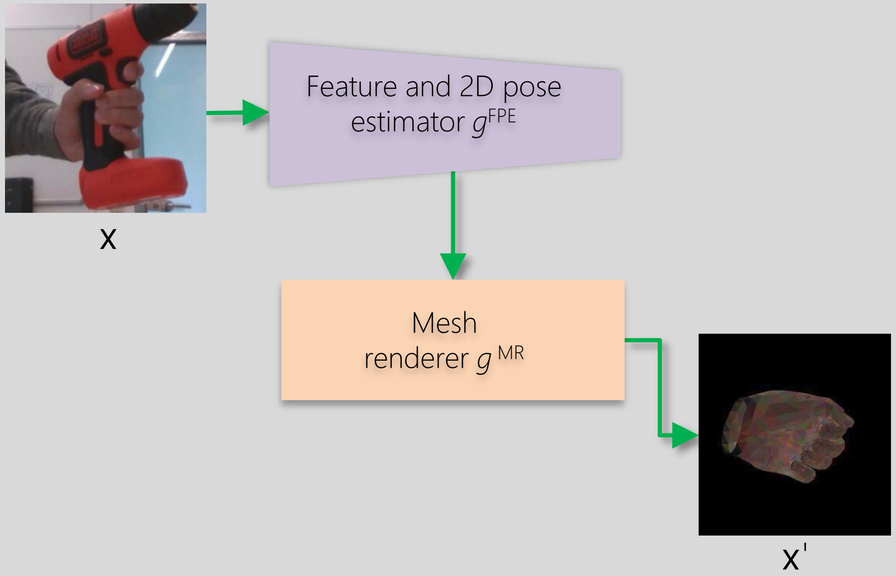


Then, we learn the hand mesh estimation using the translated images.

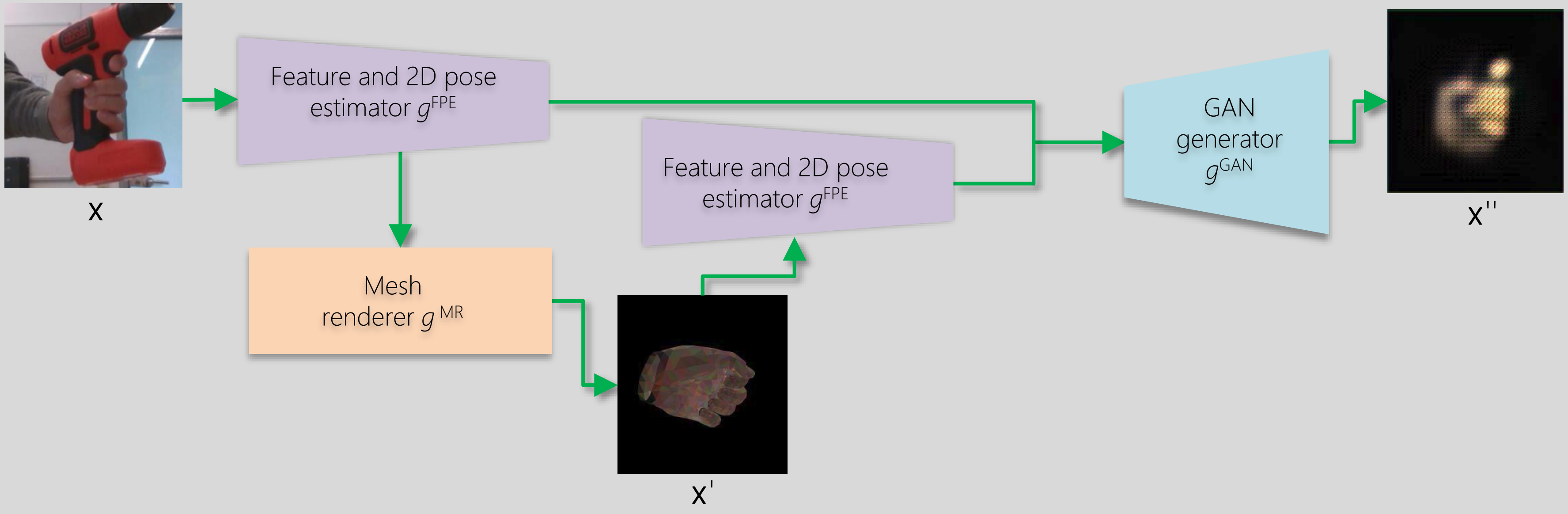
Key Idea

Finally, we obtain the skeletons from the mesh.

Pipeline

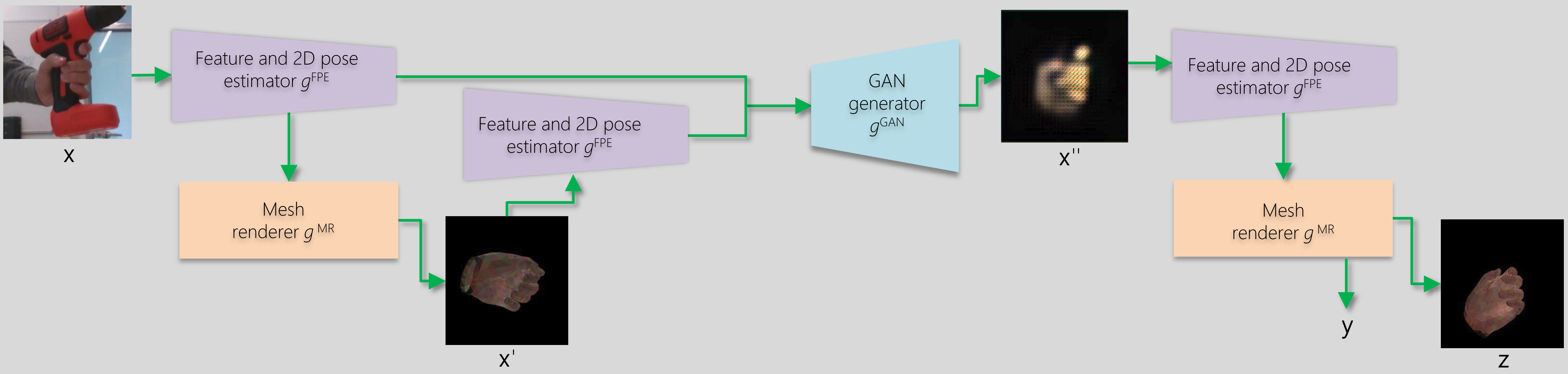


Pipeline



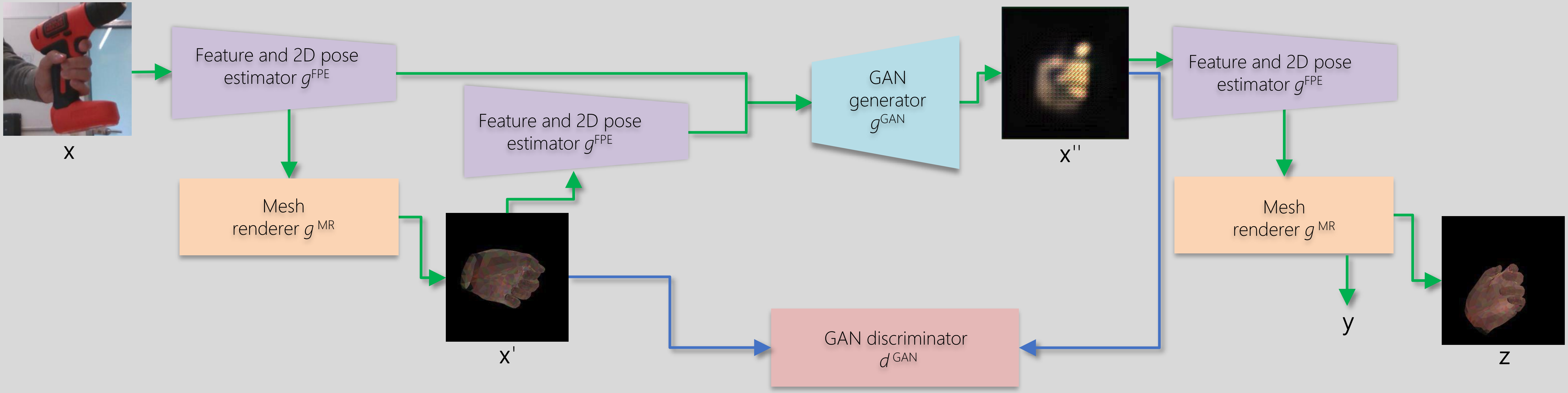
x
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
x'
Feature and 2D pose estimator g^{FPE}
GAN generator g^{GAN}
x''

Pipeline



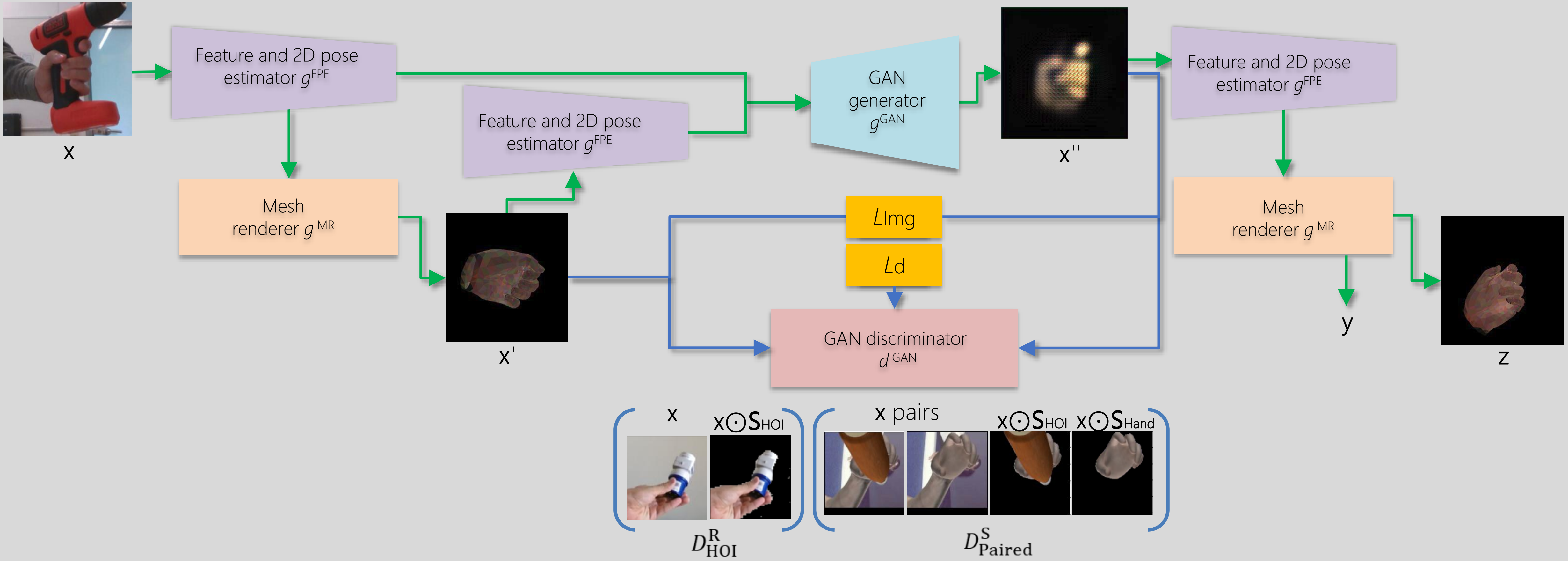
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
y
z

Pipeline



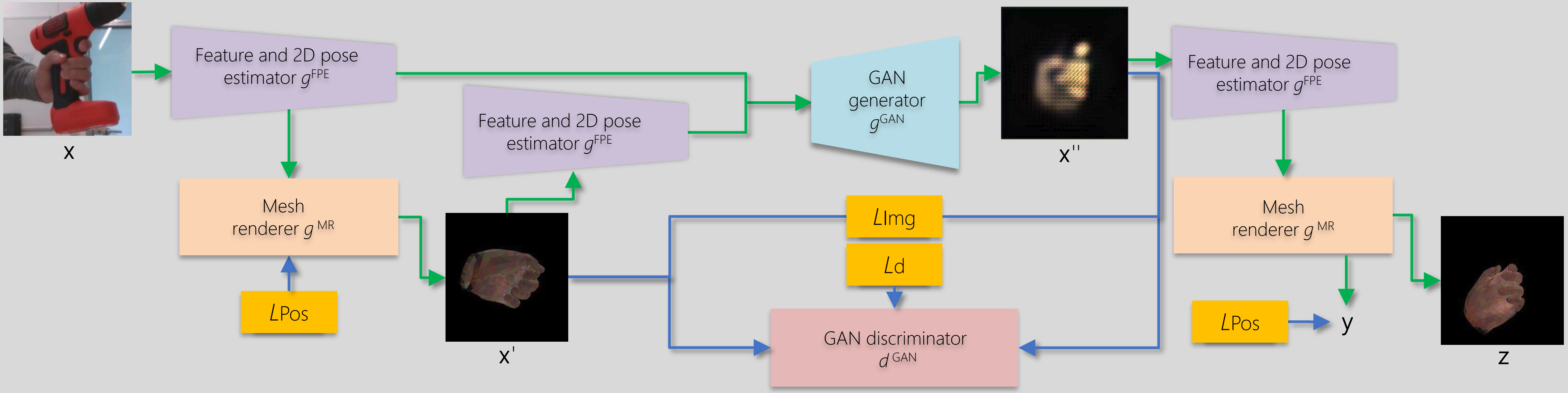
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
GAN discriminator d^{GAN}
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
y
z

Pipeline



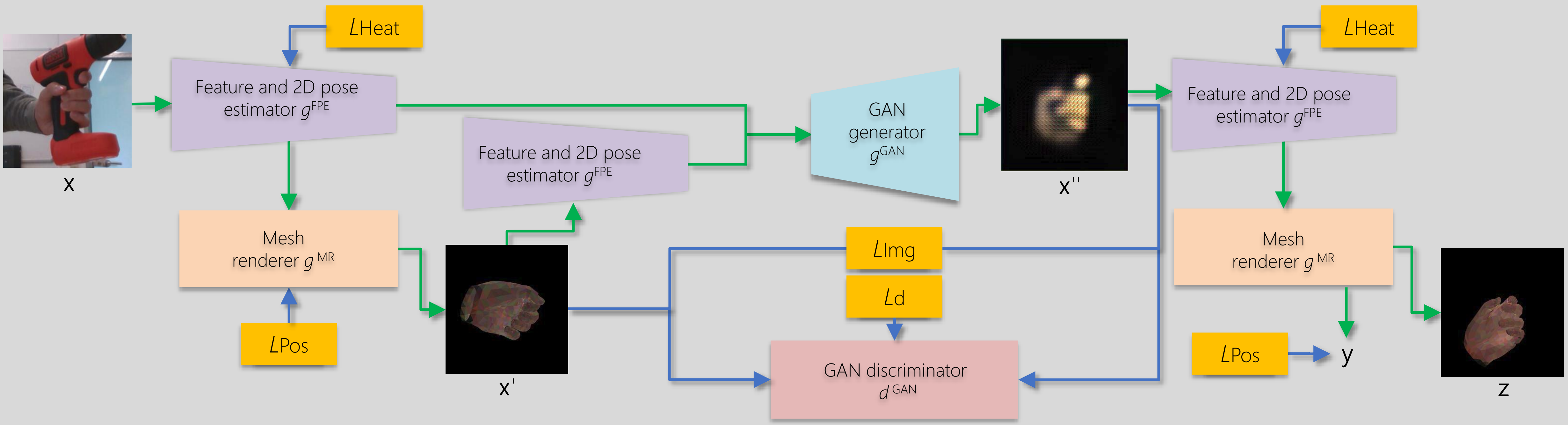
x
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
x'
Feature and 2D pose estimator g^{FPE}
GAN generator g^{GAN}
x''
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
y
z
LImg
Ld
GAN discriminator d^{GAN}
x
x⊙S HOI
D^{R}_{HOI}
x pairs
x⊙S HOI
x⊙S Hand
D^{S}_{Paired}

Pipeline



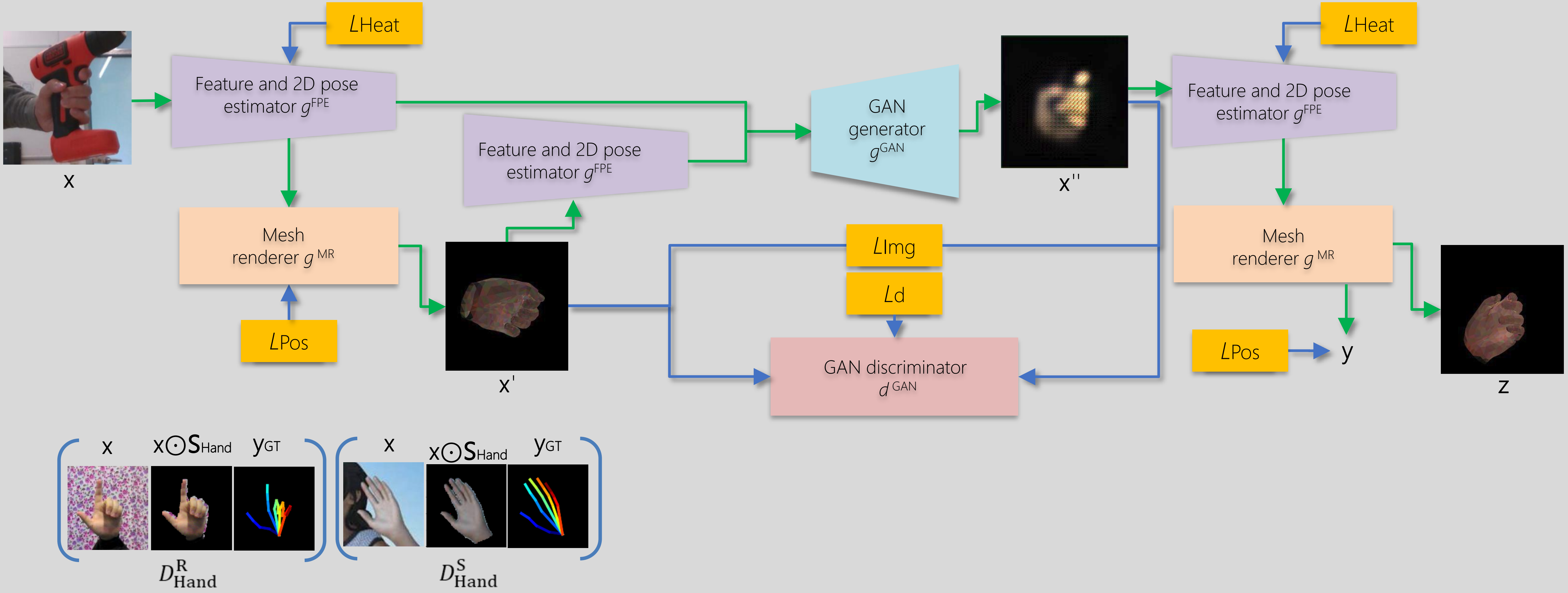
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
LPos
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
LImg
Ld
GAN discriminator d^{GAN}
x''
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z

Pipeline



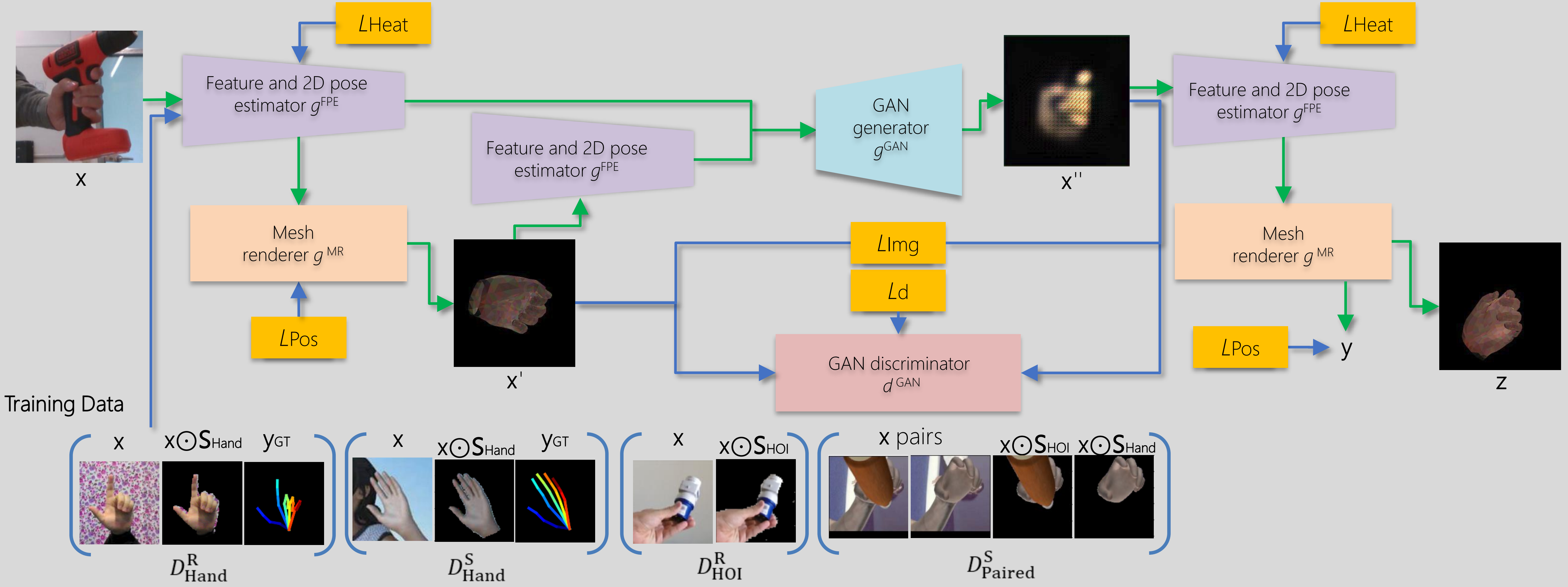
x
Feature and 2D pose estimator g^{FPE}
*L*Heat
Mesh renderer g^{MR}
*L*Pos
x'
Feature and 2D pose estimator g^{FPE}
GAN generator g^{GAN}
x''
*L*Img
*L*d
GAN discriminator d^{GAN}
*L*Heat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
*L*Pos
y
z

Pipeline



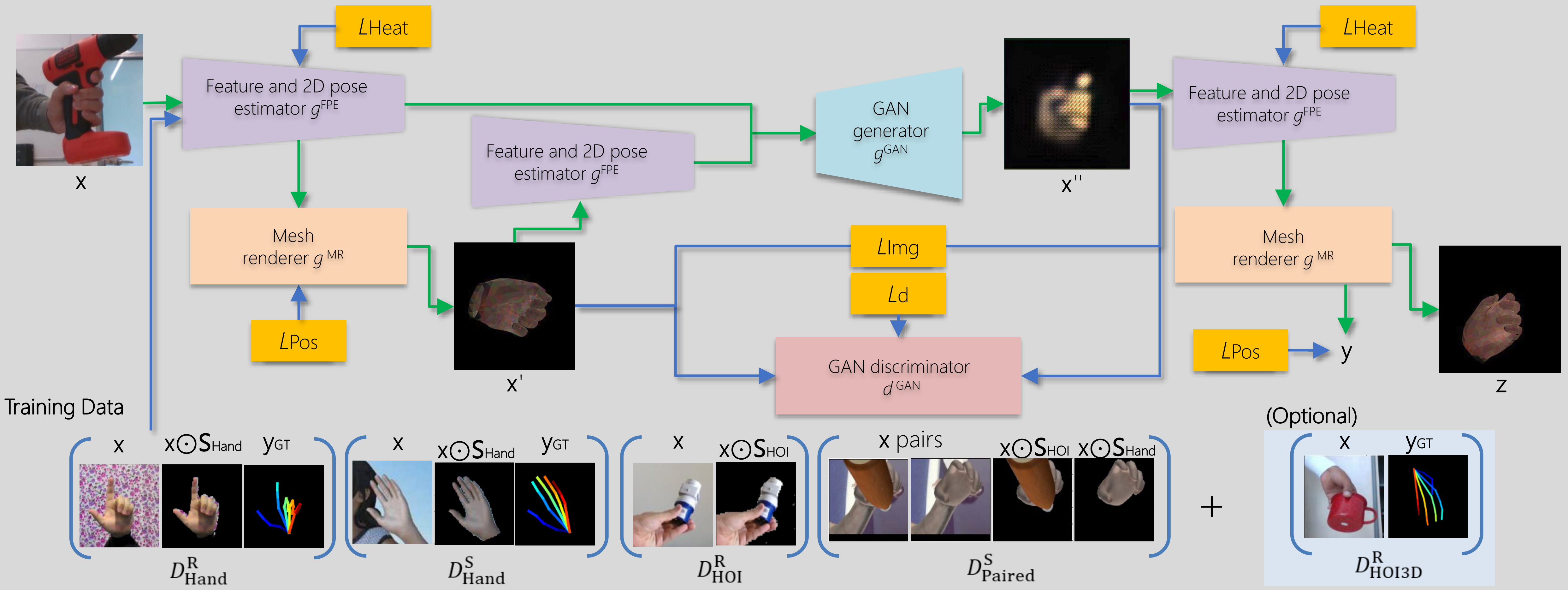
LHeat
Feature and 2D pose estimator g^{FPE}
x
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
x'
GAN generator g^{GAN}
x''
LImg
Ld
GAN discriminator d^{GAN}
LHeat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z
x
x⊙SHand
yGT
D^{R}_{Hand}
x
x⊙SHand
yGT
D^{S}_{Hand}

Pipeline



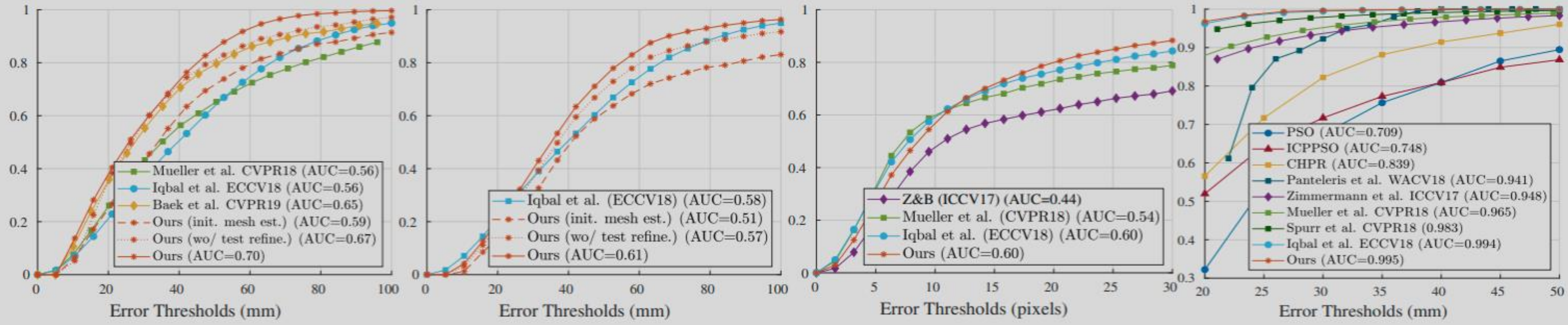
LHeat
Feature and 2D pose estimator gFPE
x
Feature and 2D pose estimator gFPE
Mesh renderer g MR
LPos
x'
GAN generator gGAN
x''
LImg
Ld
GAN discriminator d GAN
LHeat
Feature and 2D pose estimator gFPE
Mesh renderer g MR
LPos
y
z
Training Data
x
x⊙SHand
yGT
D^{R}_{Hand}
x
x⊙SHand
yGT
D^{S}_{Hand}
x
x⊙SHOI
D^{R}_{HOI}
x pairs
x⊙SHOI
x⊙SHand
D^{S}_{Paired}

Pipeline



LHeat
Feature and 2D pose estimator g^{FPE}
x
Mesh renderer g^{MR}
LPos
Feature and 2D pose estimator g^{FPE}
x'
GAN generator g^{GAN}
x''
LImg
Ld
GAN discriminator d^{GAN}
LHeat
Feature and 2D pose estimator g^{FPE}
Mesh renderer g^{MR}
LPos
y
z
Training Data
x
x⊙SHand
yGT
D^R_{Hand}
x
x⊙SHand
yGT
D^S_{Hand}
x
x⊙SHOI
D^R_{HOI}
x pairs
x⊙SHOI
x⊙SHand
D^S_{Paired}
+
(Optional)
x
yGT
D^R_{HOI3D}

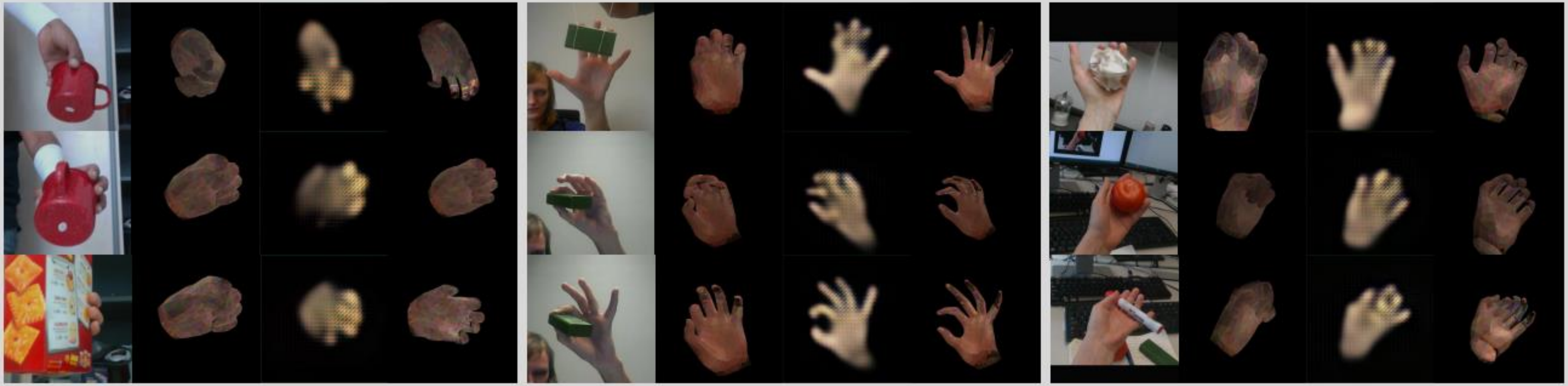
Results



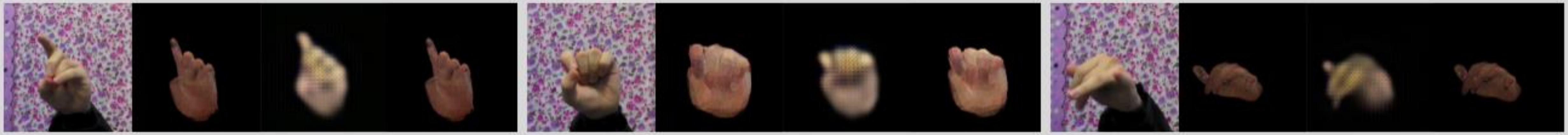
(a) 3D PCK on *DO* (b) 3D PCK on *ED* (c) 2D PCK on *ED* (d) 3D PCK on *STB*

- We obtained state-of-the-art performance, with our weakly supervised approach in challenging HOI datasets (DO, ED).
- We maintained the state-of-the-art performance in hand-only dataset (STB).

Results



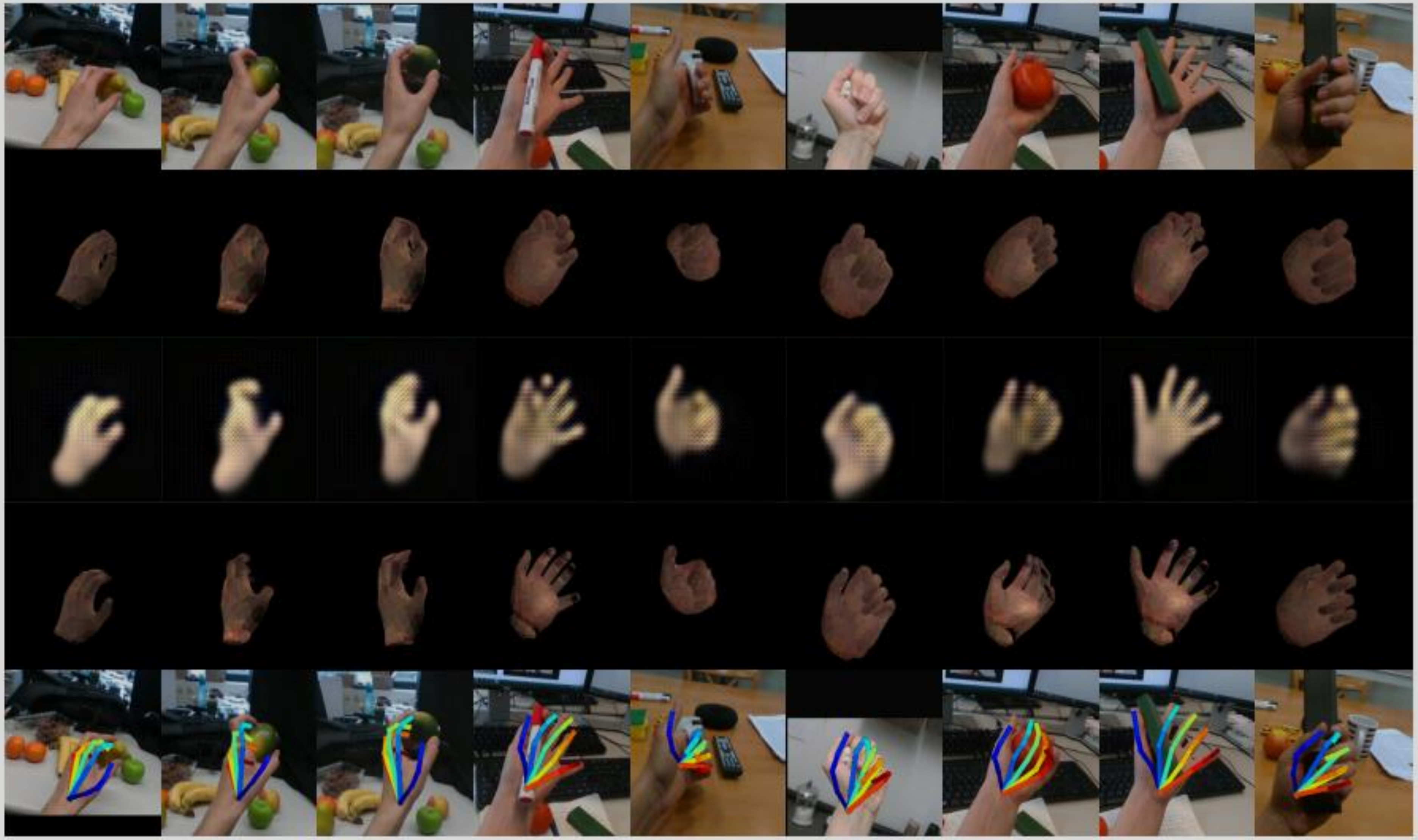
Hand-Object Interaction examples (Input/Init. Mesh/GAN output/2nd Mesh).



Hand-only examples (Input/Init. Mesh/GAN output/2nd Mesh).

Results

Results



Results