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

Jihyun Lee, Minhyuk Sung, Honggyu Choi, Tae-Kyun Kim

KAIST · Imperial College London

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

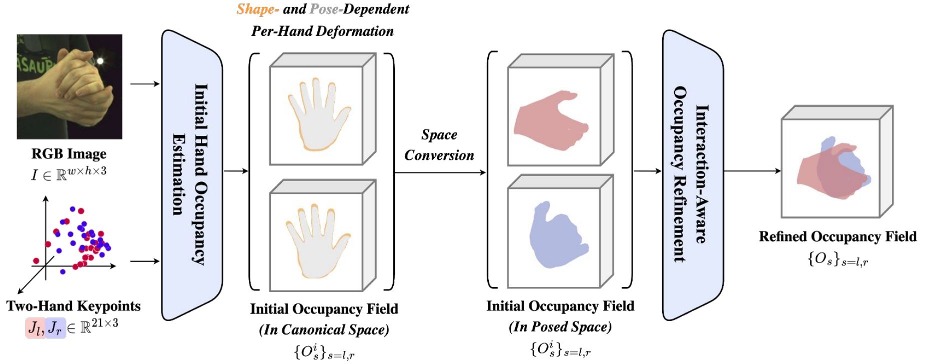
Motivation & Challenges

- **Existing two-hand reconstruction methods** model hands with low-resolution meshes with a fixed MANO[1] topology (|V| = 778).
- **Neural implicit representation** can model continuous shapes. It is also known to reconstruct shapes that are well-aligned to the input images.
  - ➔ However, implicitly modeling **complex articulations and interaction contexts between two hands** is highly challenging.

Method

- We propose two novel attention-based modules designed for:

1) **Initial per-hand occupancy estimation in the canonical space**, and
2) **Interaction-aware two-hand occupancy refinement in the original space**.



Initial Per-Hand Occupancy Estimation

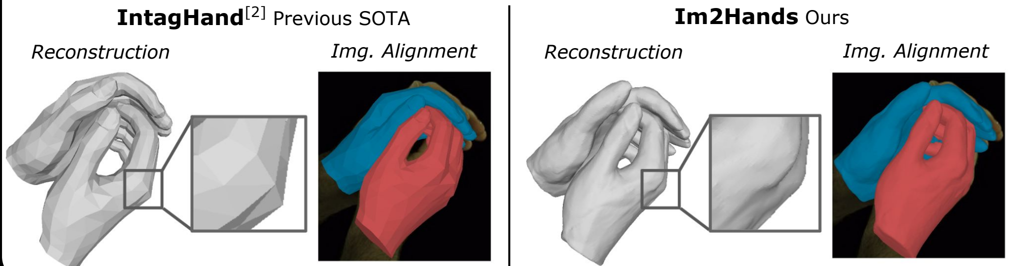
$$\mathcal{I}(x \,|\, I,\, J) = \max_{b=1,\dots,B} \{\bar{\mathcal{H}}_b(\mathbf{T}_b x,\ f_b^{\phi}, f_x^{\phi}, f_b^{\omega})\}$$

- $\bar{\mathcal{H}}_b$: Part occupancy network for bone b
- $\mathbf{T}_b x$: Canonicalized query point for bone b
- f_b^{ϕ}, f_b^{ω}: Per-bone shape and pose features[3]
- f_x^{ϕ}: Per-query shape (**query-image attention**) feature

Overview

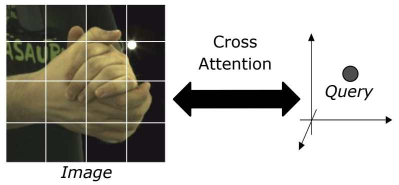
We propose **Im2Hands (Implicit Two Hands)**, the first neural implicit representation for two interacting hands.

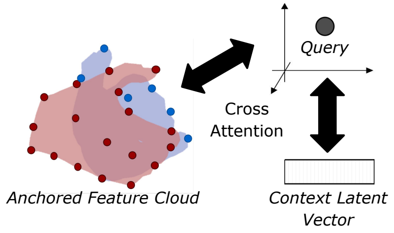
- ✅ Learns resolution-free two-hand geometries with high hand-to-hand and hand-to-image coherence
- ✅ Does not require vertex-wise shape correspondences or MANO[1] parameter annotations for training
- ✅ Achieves the state-of-the-art accuracy on two-hand reconstruction



Two-Hand Occupancy Refinement

- To encode the initial geometry of two hands, we represent them as anchored feature cloud (*i.e.* feature vectors of the points evaluated to be on surface by our initial occupancy network).
- We then apply **cross-attention between (1) a query, (2) anchored features, and (3) a context latent vector** to estimate the refined occupancy.





📌 We also proposed an **optional keypoint refinement module** for an image-based reconstruction scenario.

Please check the paper for more details.

Experiments

- Im2Hands achieves SOTA reconstruction results on InterHand2.6M[4].

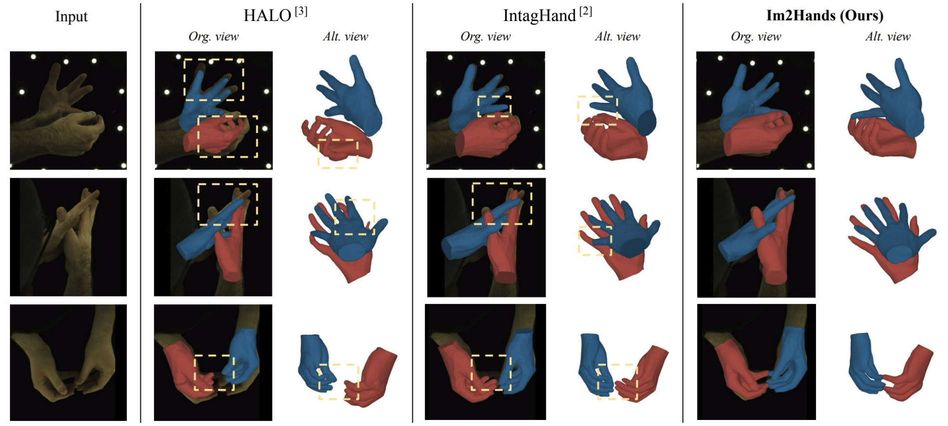
Using Image and Keypoint Inputs

Method	Inputs	IoU (%) ↑	CD (mm) ↓
Two-Hand-Shape-Pose[5]	$\mathcal{I}, \mathcal{L}$	54.8	5.51
IntagHand[2]	$\mathcal{I}, \mathcal{L}$	67.0	3.88
HALO[3]	$\mathcal{J}$	74.7	2.62
HALO*[3]	$\mathcal{I}, \mathcal{J}$	75.8	2.51
Im2Hands (Ours)	$\mathcal{I}, \mathcal{J}$	**77.8**	**2.30**

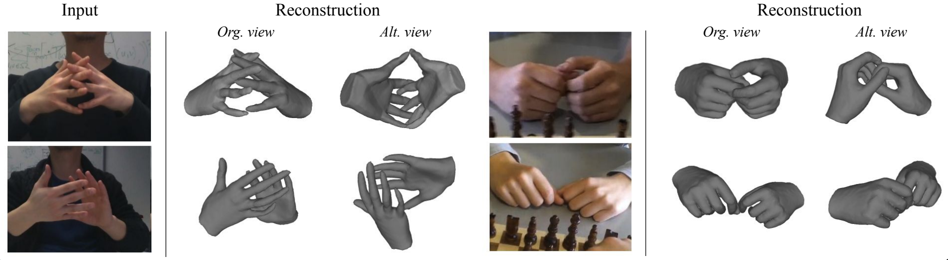
Using Image Inputs Only (+ Predicted Keypoints)

Method	IoU (%) ↑	CD (mm) ↓
Two-Hand-Shape-Pose[5]	48.4	6.09
IntagHand[2]	59.0	4.69
DIGIT[6]+ HALO[3]	45.1	7.64
IntagHand[2]+ HALO[3]	53.8	5.38
DIGIT[6]+ Im2Hands (Ours)	**59.4**	**4.75**
IntagHand[2]+ Im2Hands (Ours)	**62.1**	**4.35**

Qualitative Results on Image-Based Two-Hand Reconstruction



- We also show generalization test results on RGB2Hands[7] and EgoHands[8] datasets.



References

[1] J. Romero *et al*. Embodied hands: Modeling and capturing hands and bodies together. TOG, 2017.
[2] M. Li *et al.* Interacting attention graph for single image two-hand reconstruction. In CVPR, 2022.
[3] K. Karunratanakul *et al*. A skeleton-driven neural occupancy representation for articulated hands. In 3DV, 2021.
[4] G. Moon *et al*. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In ECCV, 2020.
[5] B. Zhang *et al*. Interacting two-hand 3d pose and shape reconstruction from single color image. In ICCV, 2021.
[6] Z. Fan *et al.* Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In 3DV, 2021.
[7] J. Wang *et al*. Rgb2hands: Real-time tracking of 3d hand interactions from monocular rgb video. TOG, 2020.
[8] S. Bambach *et al*. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In ICCV, 2015.