THU-AM-052

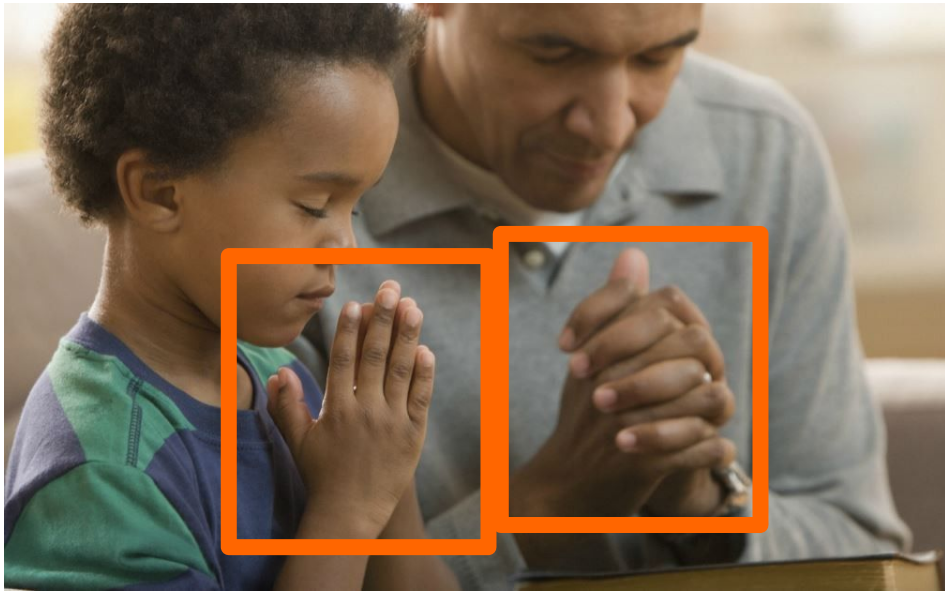
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Bringing Inputs to Shared Domains for 3D Interacting Hands Recovery *in the Wild*

Gyeongsik Moon

Meta

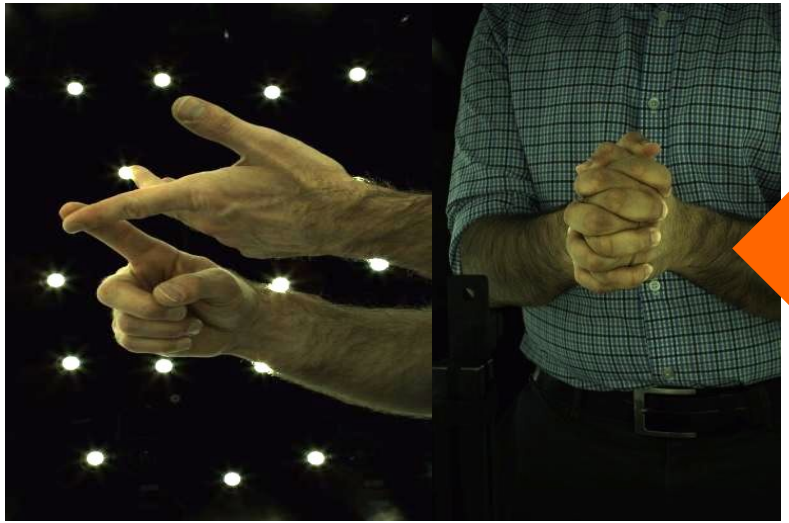
We often make two-hand interactions in our daily life



3D Interacting Hands Recovery in the Wild

- Most existing works only have focused on results on MoCap datasets, such as InterHand2.6M [1]
  - They have 3D data, but have ***severe appearance domain gap from in-the-wild images***

limited backgrounds/lightings with 3D GT



Images from InterHand2.6M

diverse backgrounds/lightings without 3D GT

In-the-wild images

[1] Moon, Gyeongsik, et al. "Interhand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image." ECCV, 2020.

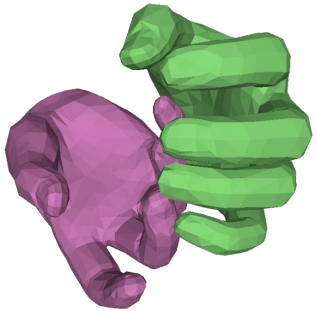
3D Interacting Hands Recovery in the Wild

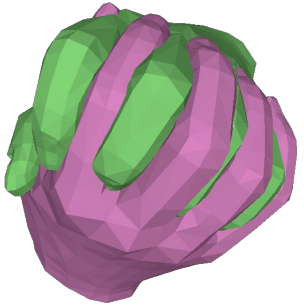
How can we recover 3D interacting hands from in-the-wild images without 3D GT from them?

Human image

Hand image

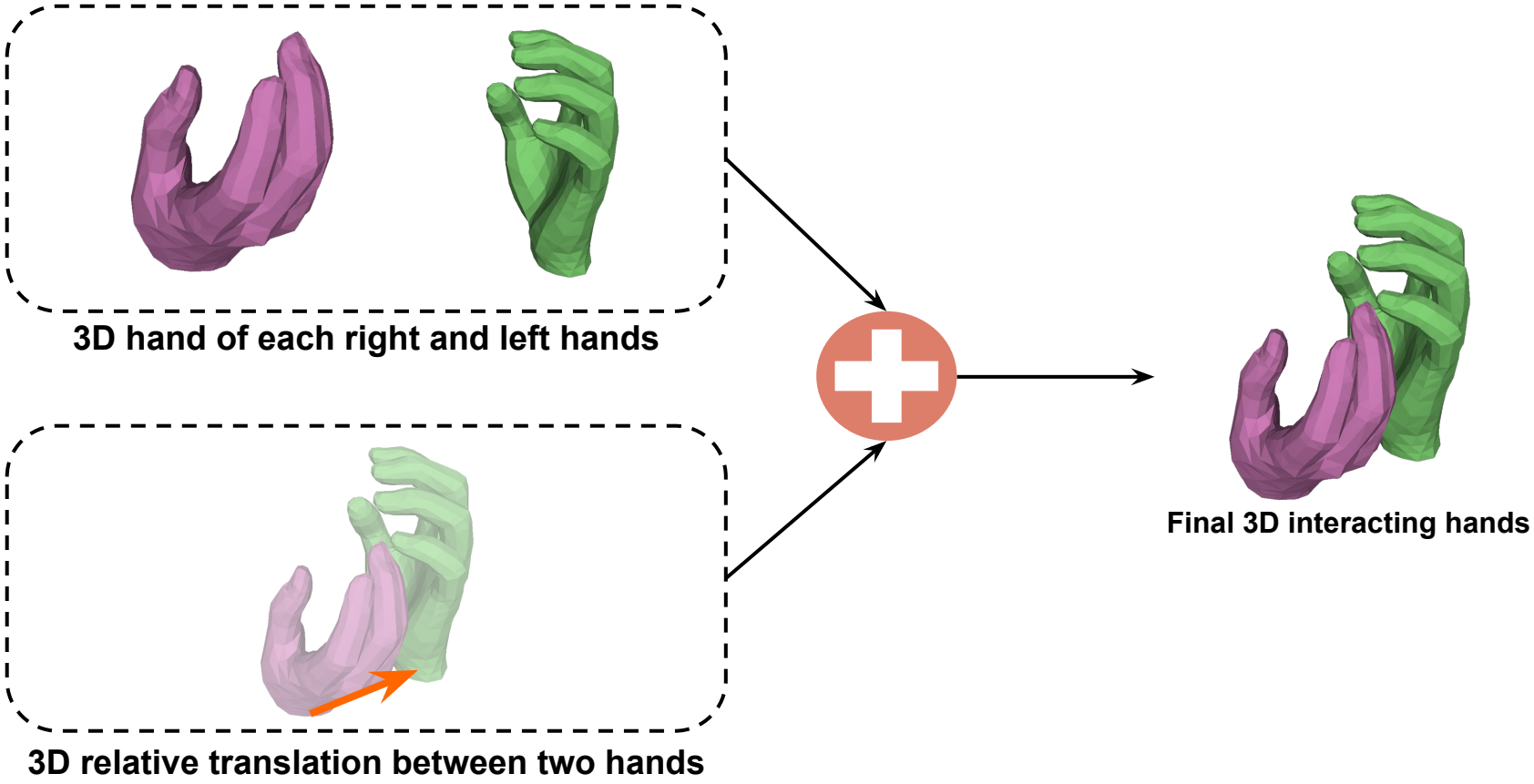
Front view



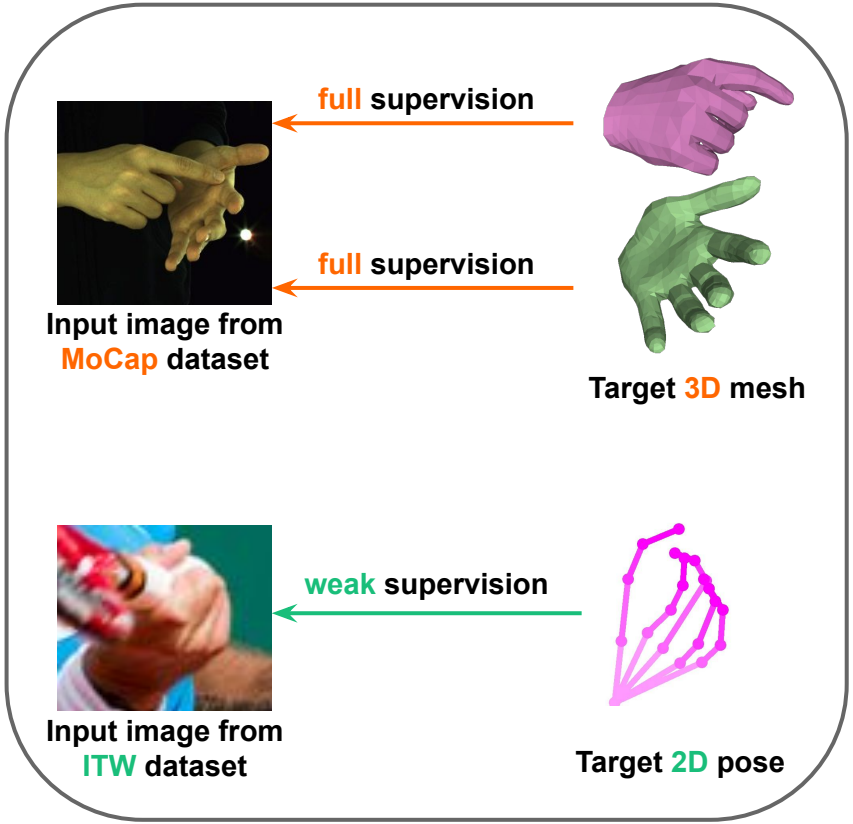


Other view

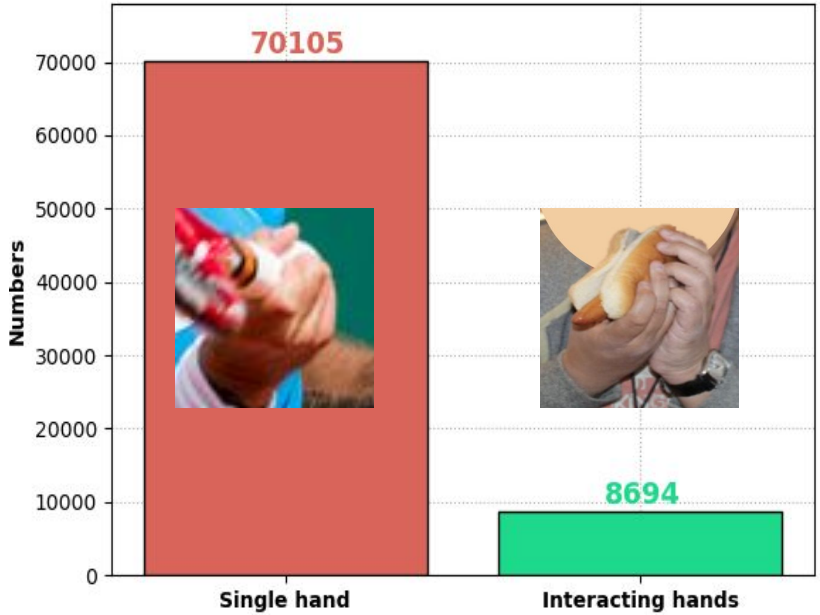
Two Sub-Problems of 3D Interacting Hands Recovery



1st Sub-Problem: 3D Recovery of Each Hand

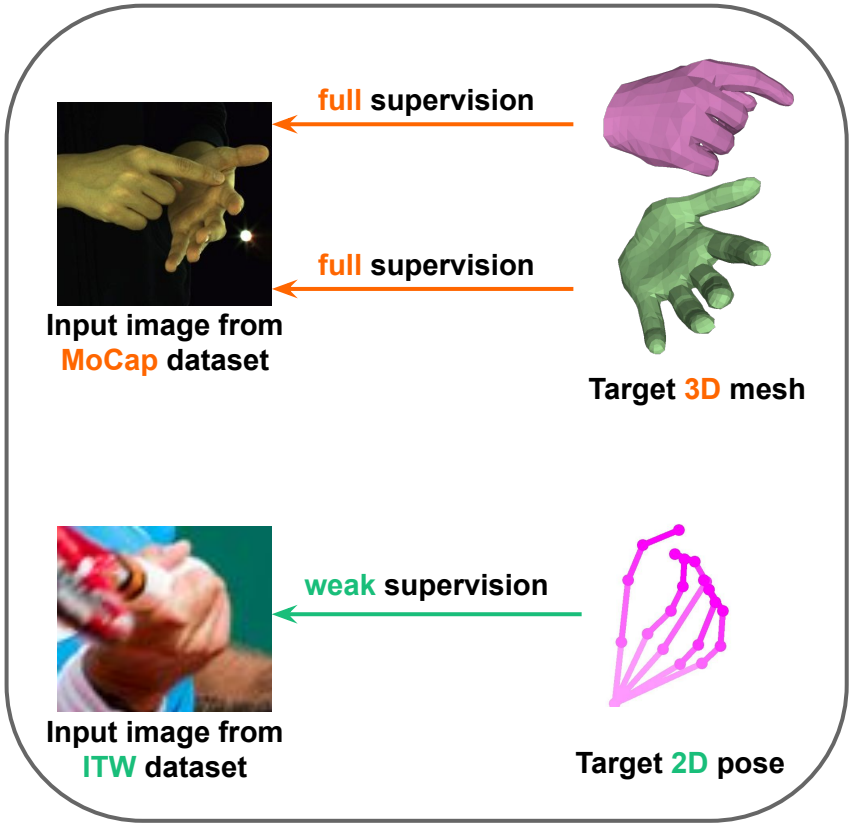


(a) Mini-batch of previous approaches with the mixed-batch training

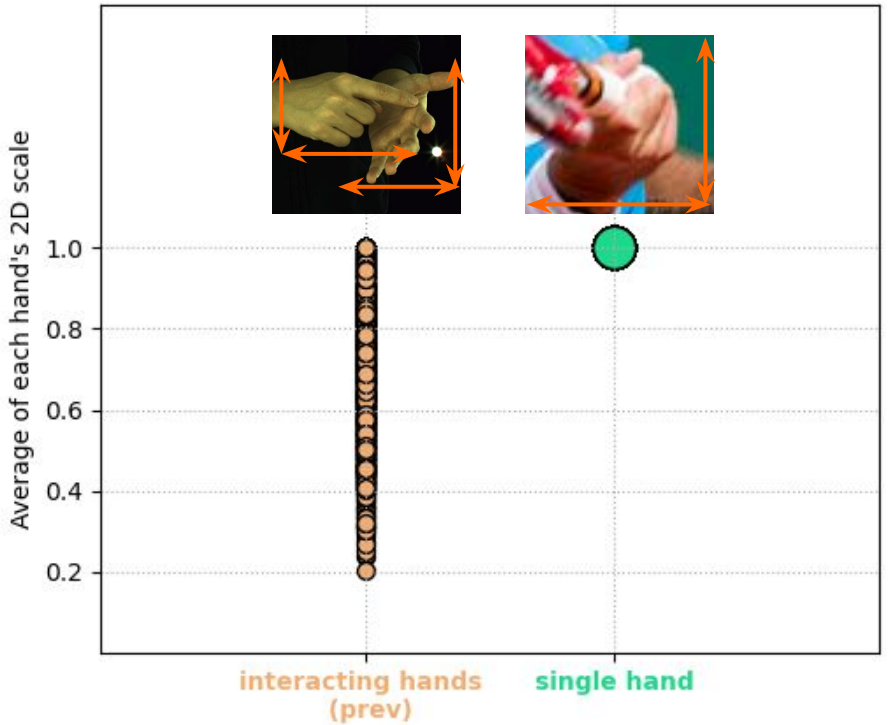


The number of single hand vs. interacting hands in MSCOCO

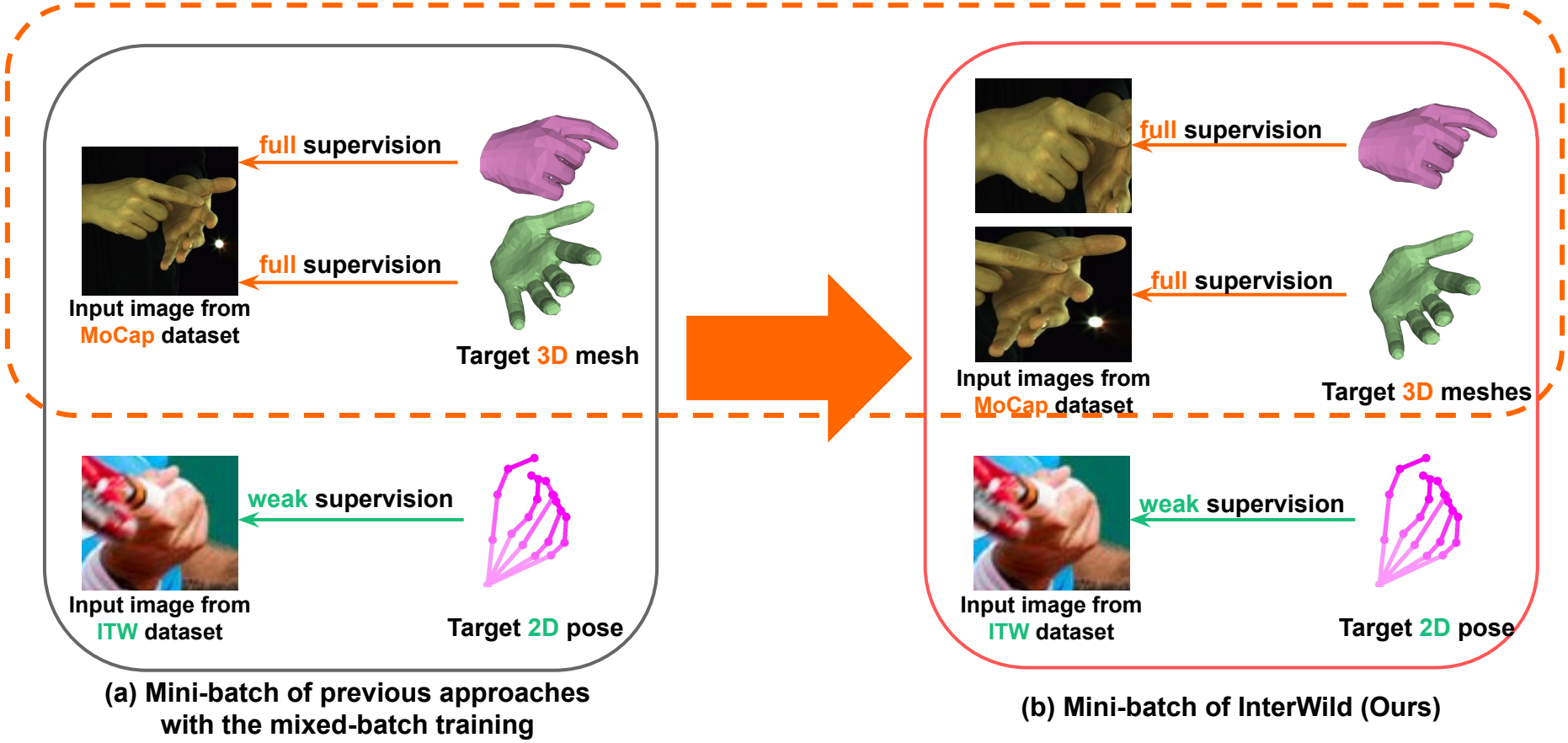
1st Sub-Problem: 3D Recovery of Each Hand



(a) Mini-batch of previous approaches with the mixed-batch training



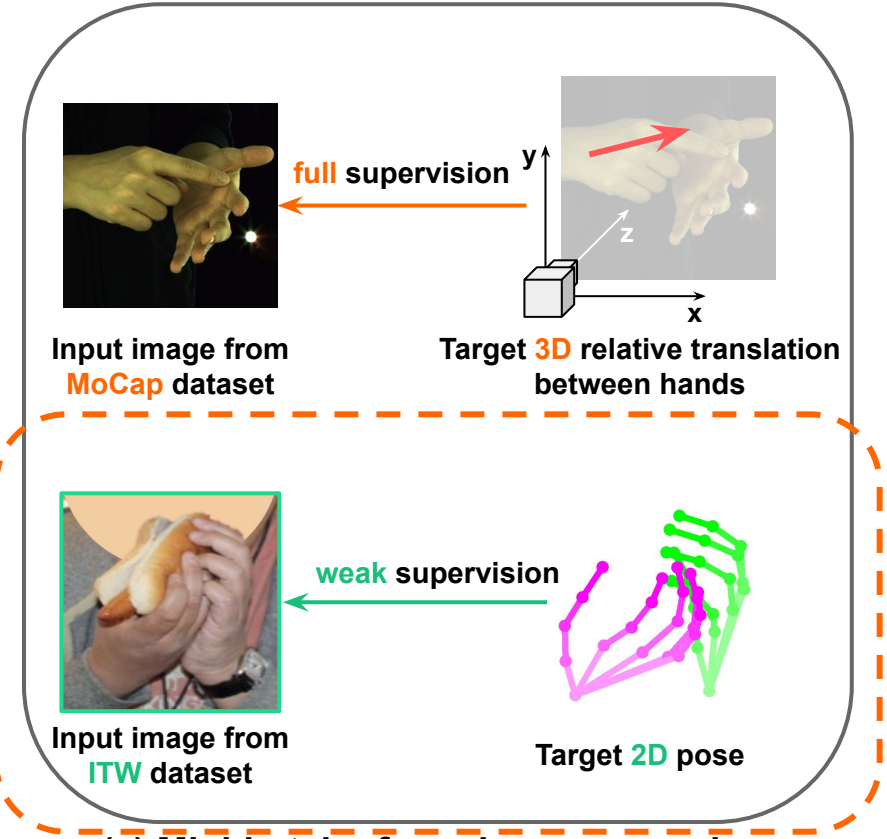
We bring inputs to a *shared 2D scale* space!



(a) Mini-batch of previous approaches with the mixed-batch training

(b) Mini-batch of InterWild (Ours)

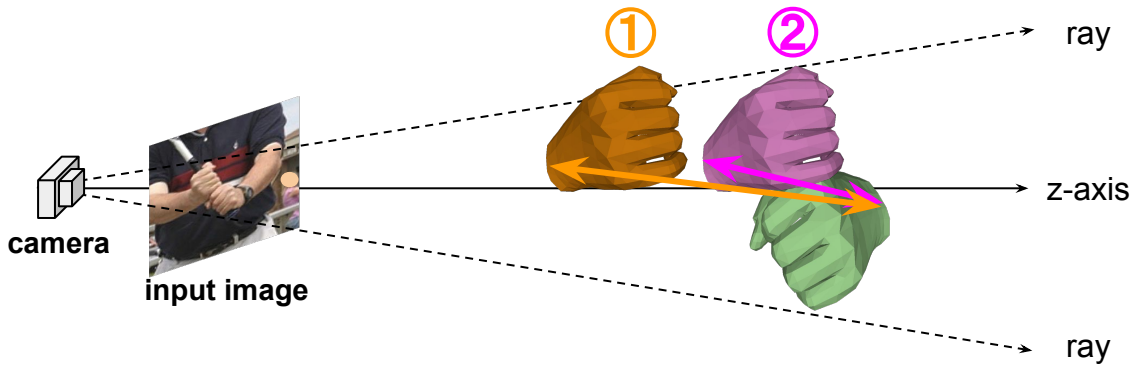
2nd Sub-Problem: 3D Relative Translation between Two Hands



(a) Mini-batch of previous approaches with the mixed-batch training

3D relative translation between two hands - Hard to be weakly supervised

- 3D relative translation between two hands are not restricted at all
- On the other hand, each 3D hand is restricted by a 3D hand model (e.g., size is about 15 cm)

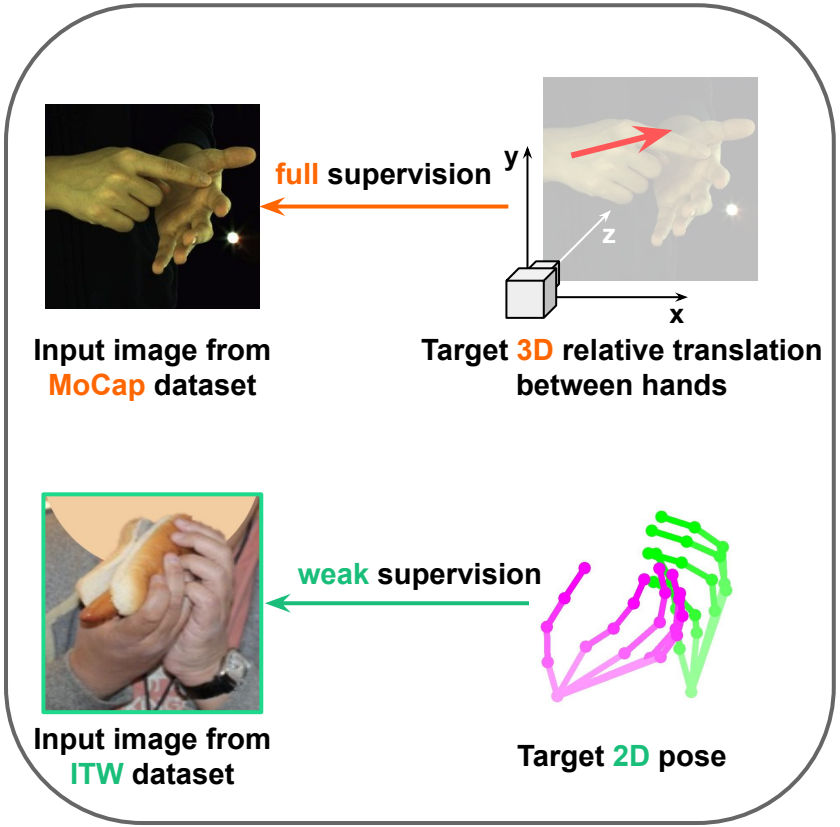


① wrong global translation (right hand) ⟷ wrong relative translation

② correct global translation (right hand) ⟷ correct relative translation

The failure case of the 2D-based weak supervision for the 3D relative translation between two hands

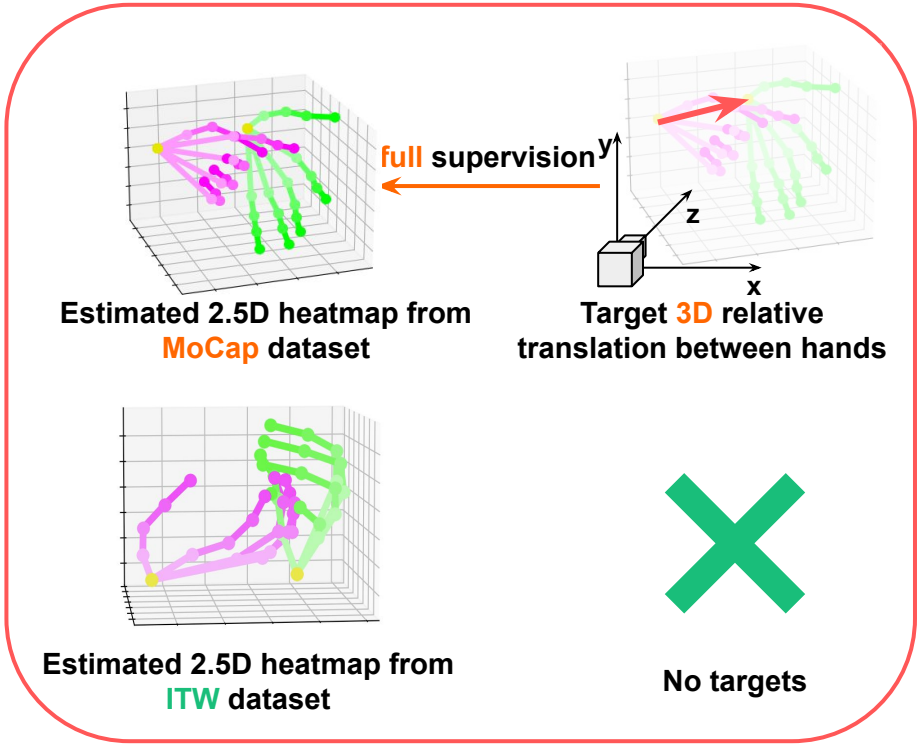
Just remove the weak supervision?



(a) Mini-batch of previous approaches with the mixed-batch training

- Without the weak supervision, in-the-wild samples are not exposed to the regressor
- A regressor trained only on MoCap images would not generalize to in-the-wild images!

We bring inputs to a *shared appearance-invariant* space!

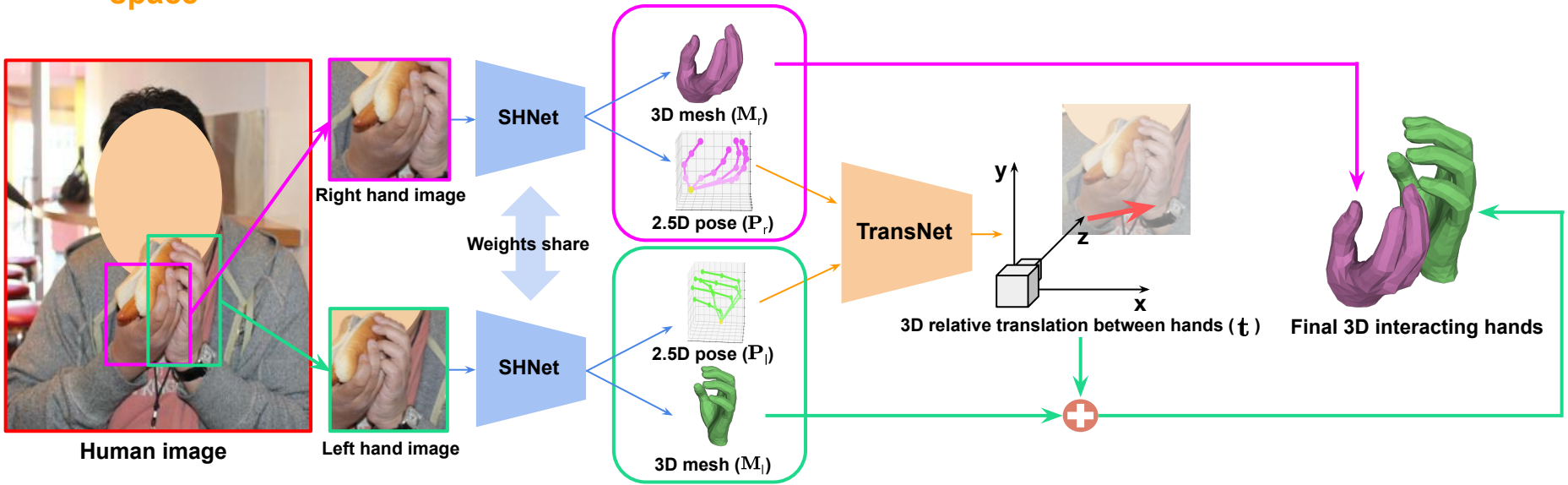


(b) Mini-batch of InterWild (Ours)

- Remove the weak supervision
- Generalize well to in-the-wild data

Overall Pipeline

- **SHNet**: takes a single hand image for the 3D mesh of each hand - **shared 2D scale space**
- **TransNet**: takes 2.5D pose of two hands for the 3D relative translation - **shared appearance-invariant space**



Ablation Studies

Taking a single-hand image gives lower 3D errors of SHNet

Inputs of SHNet	HIC [33]	IH2.6M [23]
Two-hand image	29.80 / 35.86	11.36 / 13.20
Single-hand image (Ours)	**15.65 / 15.70**	**11.12 / 13.01**

Taking a 2.5D pose gives better generalization power
Weak supervision is always bad for any input types of TransNet

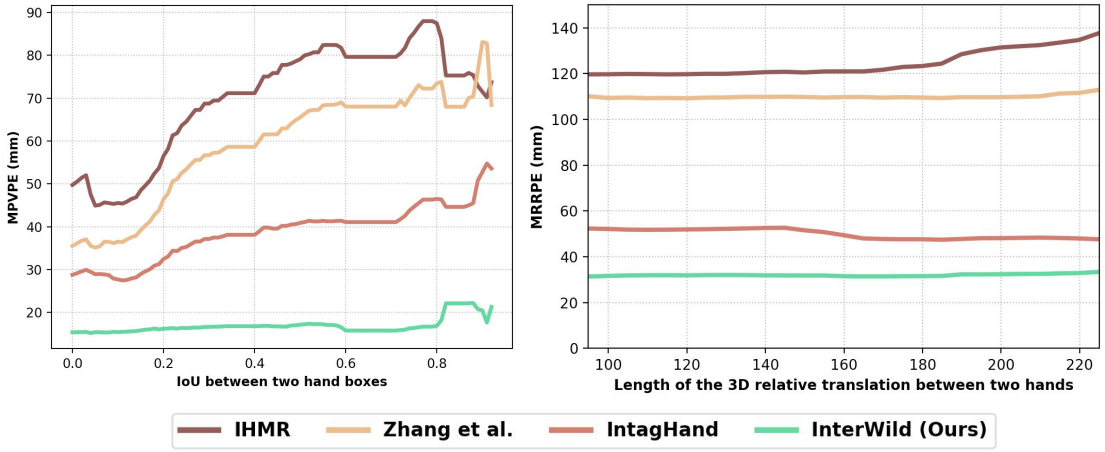
Inputs of TransNet	weak sup.	HIC [33]	IH2.6M [23]
Img.	✗	206.83	27.67
	✓	215.35	35.72
Img. + 2.5D hm.	✗	54.36	**27.19**
	✓	58.53	33.15
2D hm.	✗	38.64	31.51
	✓	51.19	35.51
2.5D hm.	✗**(Ours)**	**31.35**	29.29
	✓	61.05	33.91

Comparison to State-Of-The-Art Methods

HIC: indoor images with more realistic image appearances than IH2.6M, but with small scale and limited poses

Better generalization to in-the-wild images

Methods	HIC [33]		IH2.6M [23]	
	MPVPE	MRRPE	MPVPE	MRRPE
IHMR [29]	30.76 / 46.38	119.64	15.35 / 18.53	33.39
Zhang *et al.* [36]	23.53 / 31.79	110.25	11.76 / 14.17	31.56
IntagHand [16]	18.83 / 27.31	52.46	11.18 / 13.49	29.31
InterWild (Ours)	**15.65 / 15.70**	**31.35**	**11.12 / 13.01**	**29.29**




Human image
Hand image
Front view
Other view
InterWild (Ours)
Front view
Other view
IntagHand


Human image
Hand image
Front view
Other view
InterWild (Ours)
Front view
Other view
IntagHand