South China University of Technology

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Harmonious Feature Learning for Interactive Hand-Object Pose Estimation

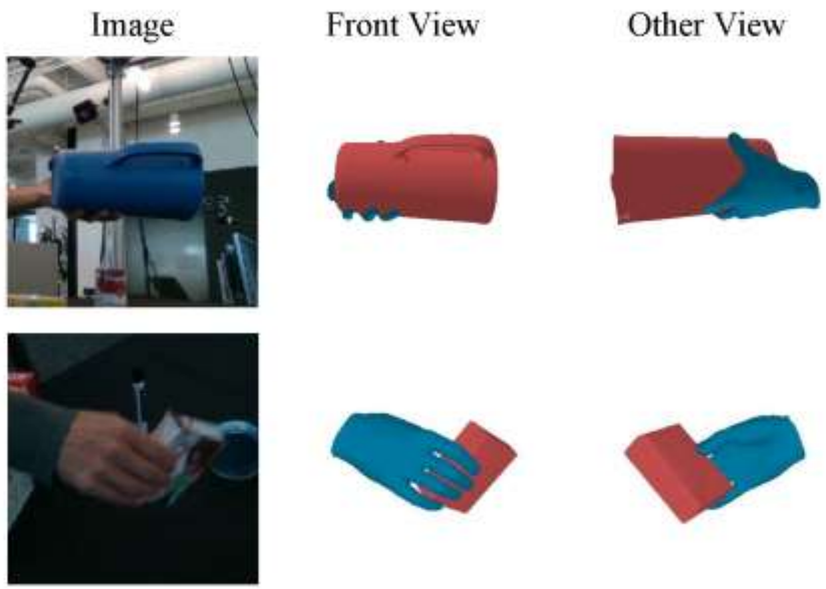
Zhifeng Lin[1] Changxing Ding[1,2*] Huan Yao[1] Zengsheng Kuang[1] Shaoli Huang[3]

[1] South China University of Technology [2] Pazhou Lab, Guangzhou [3] Tencent AI-Lab, Shenzhen

Paper Tag : WED-PM-060

Understanding hand-object interaction

- Estimating 3D hand and object pose from a single image

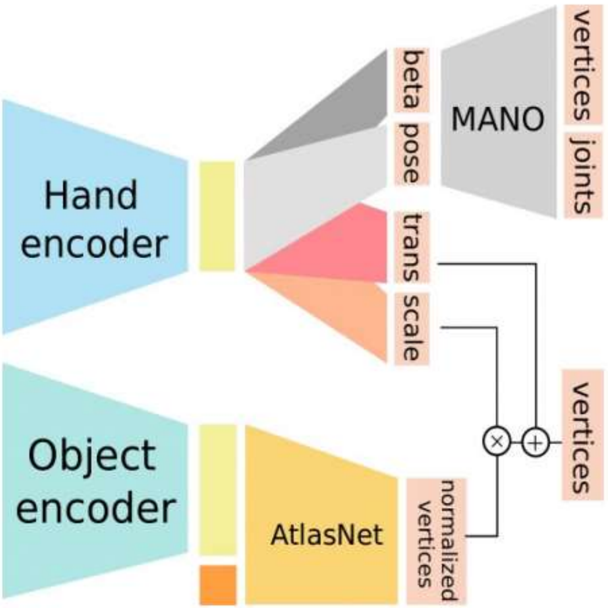


- Challenge:

Hands and objects are often self-occluded during interactions

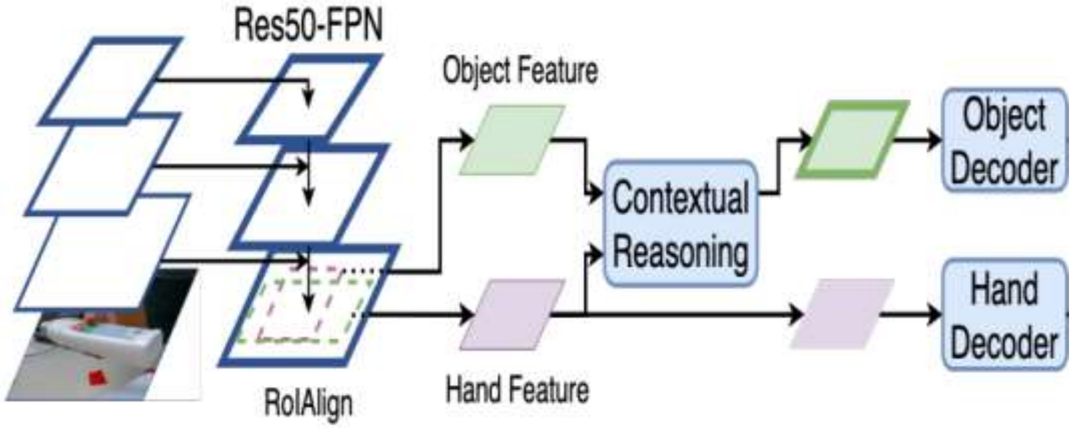
Related Work

Separate Encoders



Hasson et al., 2019

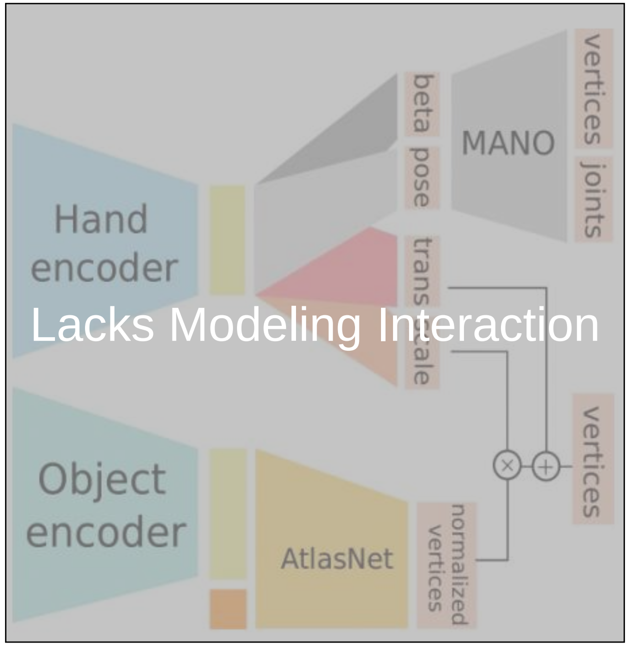
One Encoder and Feature Fusion



Liu et al., 2021

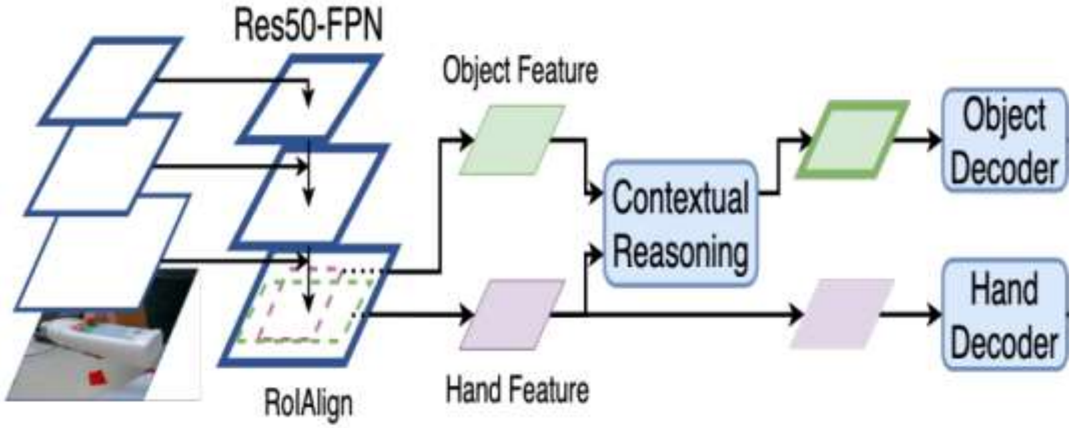
Related Work

Separate Encoders



Hasson et al., 2019

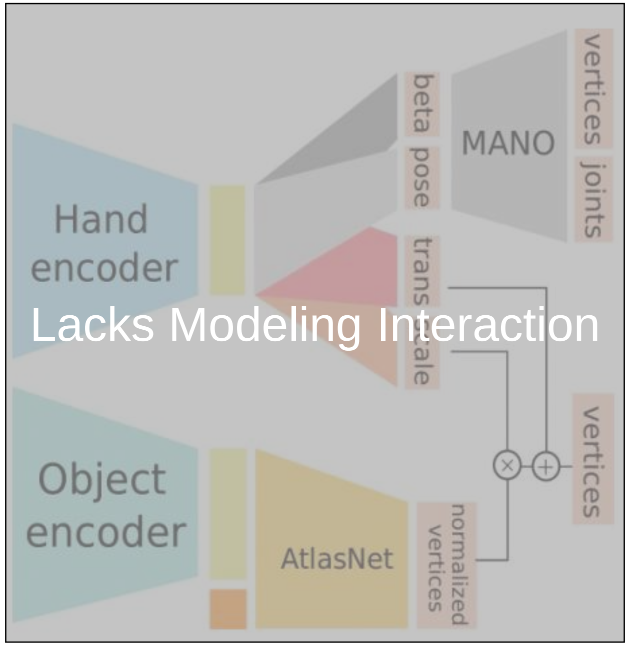
One Encoder and Feature Fusion



Liu et al., 2021

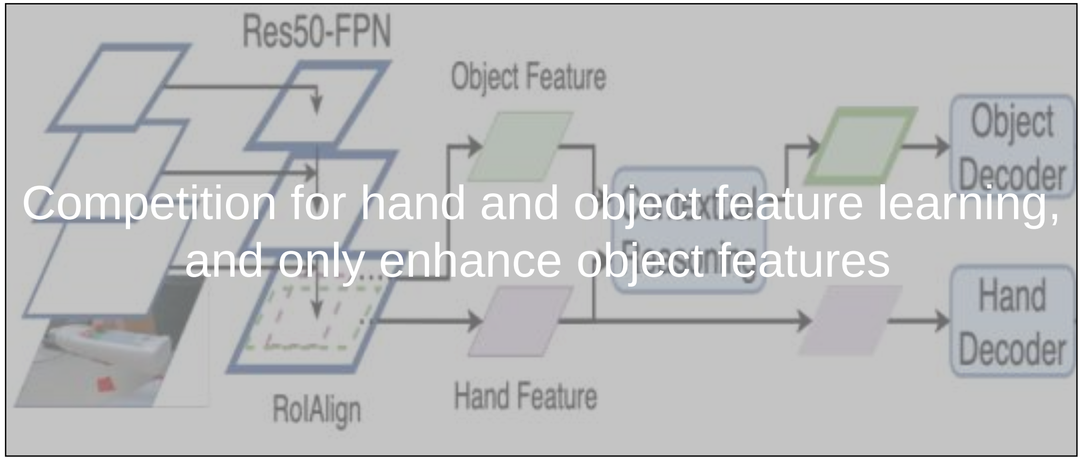
Related Work

Separate Encoders



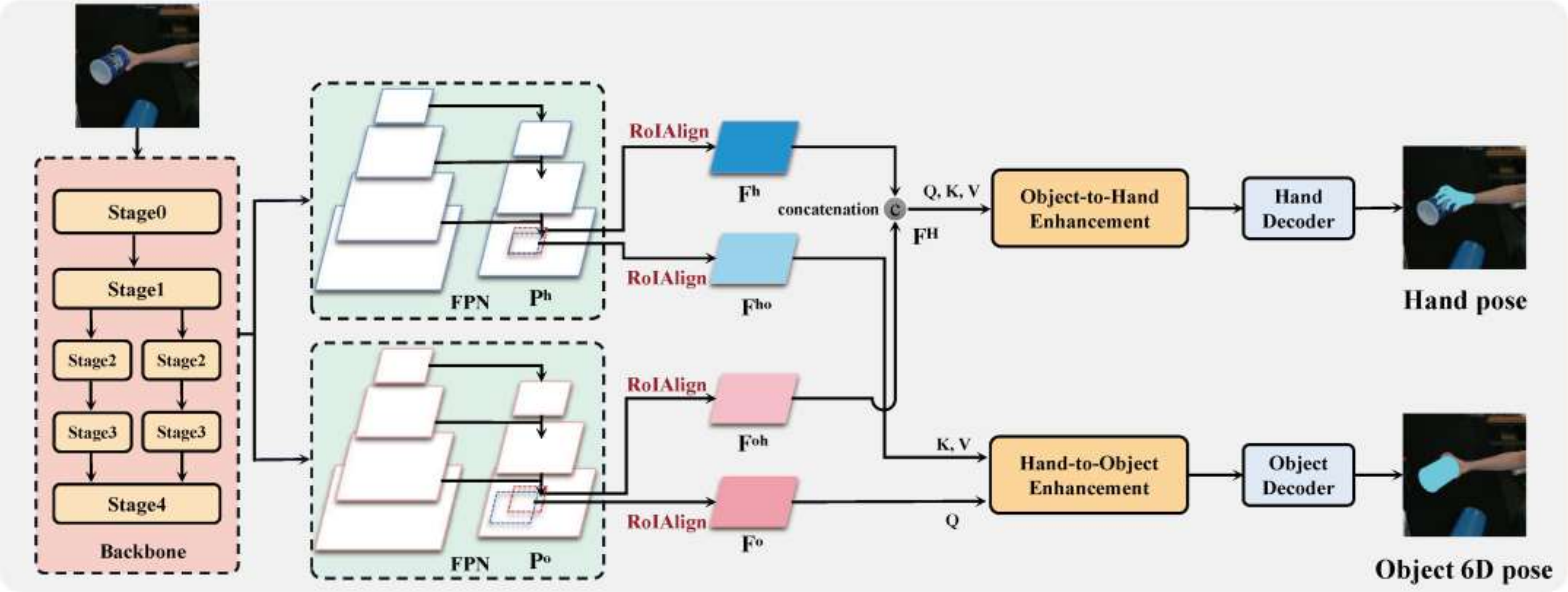
Hasson et al., 2019

One Encoder and Feature Fusion

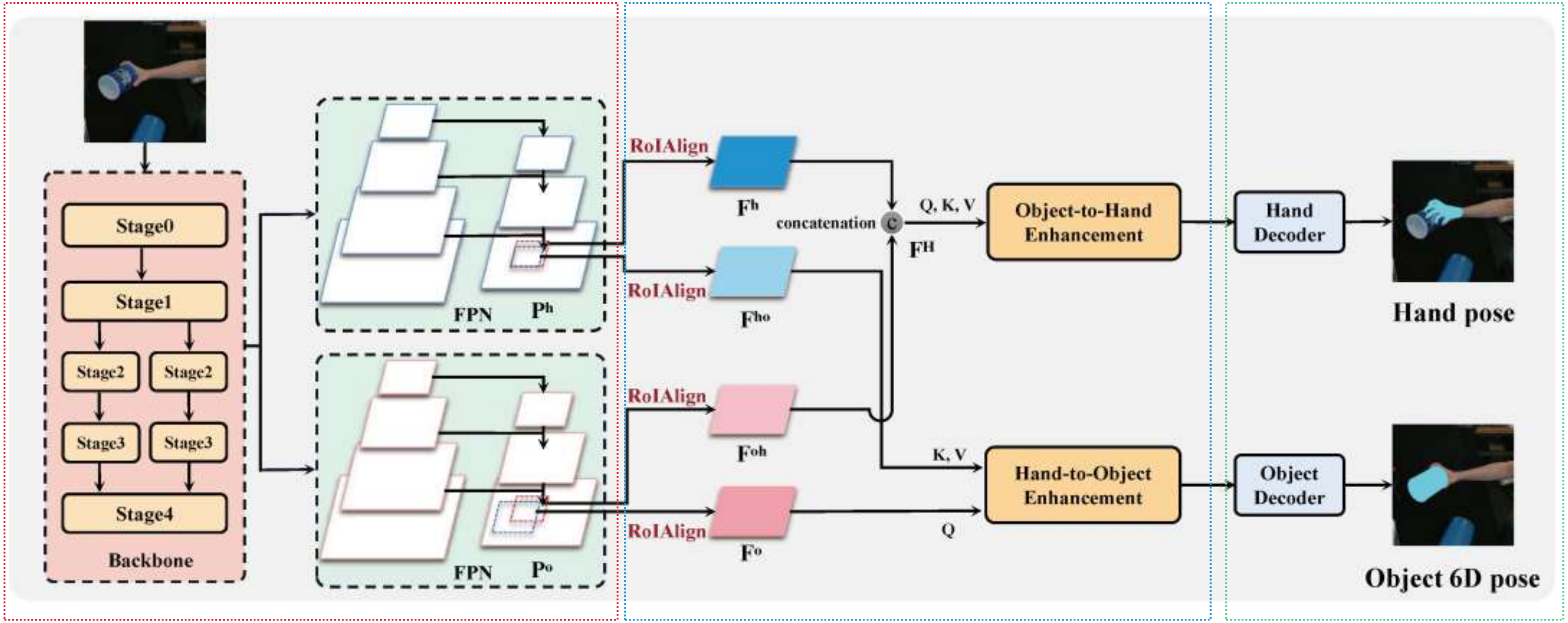


Liu et al., 2021

Proposed Method



Proposed Method

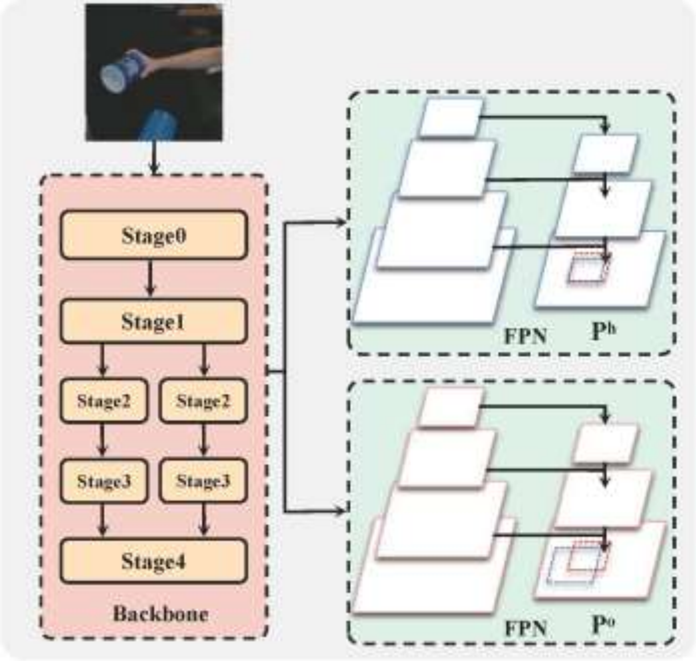


Feature Extraction Backbone

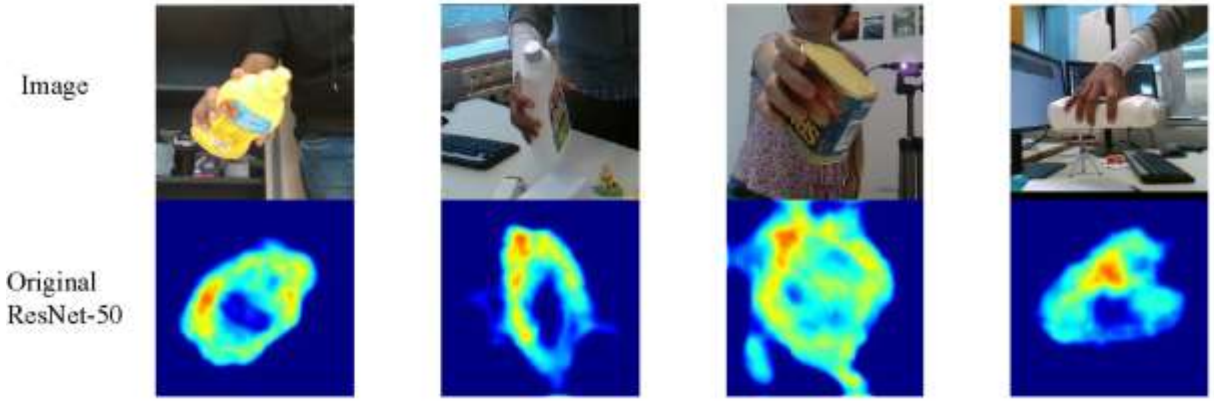
Interaction Modules

Two separate decoders

1.Feature Extraction Backbone

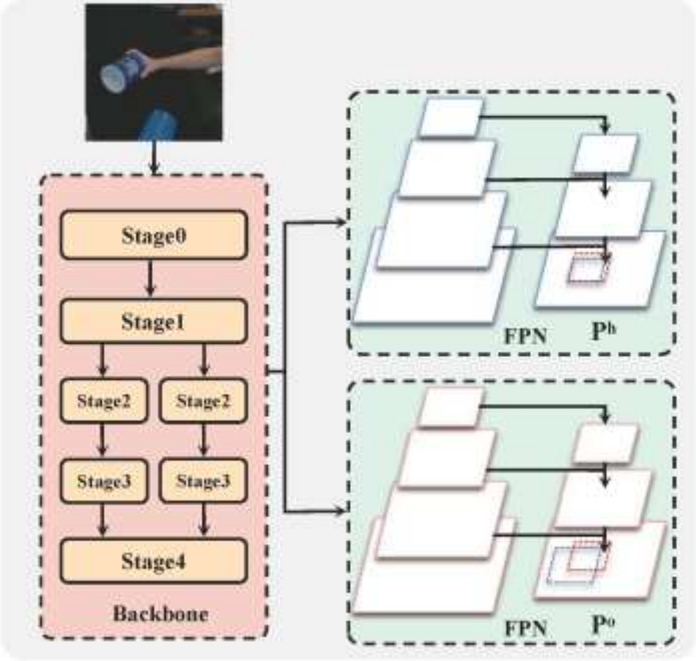


- single stream backbone -> treats the hand and object both as foreground, competitive in feature learning



- double stream backbone -> large number of parameters, the different feature spaces between backbones

1.Feature Extraction Backbone



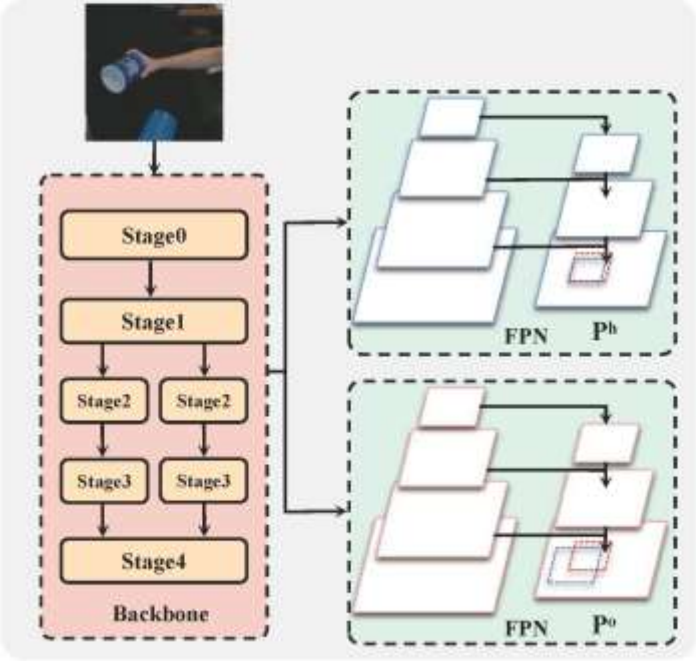
- Our backbone keeps the structure of the stage-0, stage-1, and stage-4 layers of the ResNet-50 model unchanged, but adopts independent stage-2 and stage-3 layers for the hand and object.

- The feature maps output by the stage-1 layers are fed into the two sets of stage-2 and stage-3 layers.

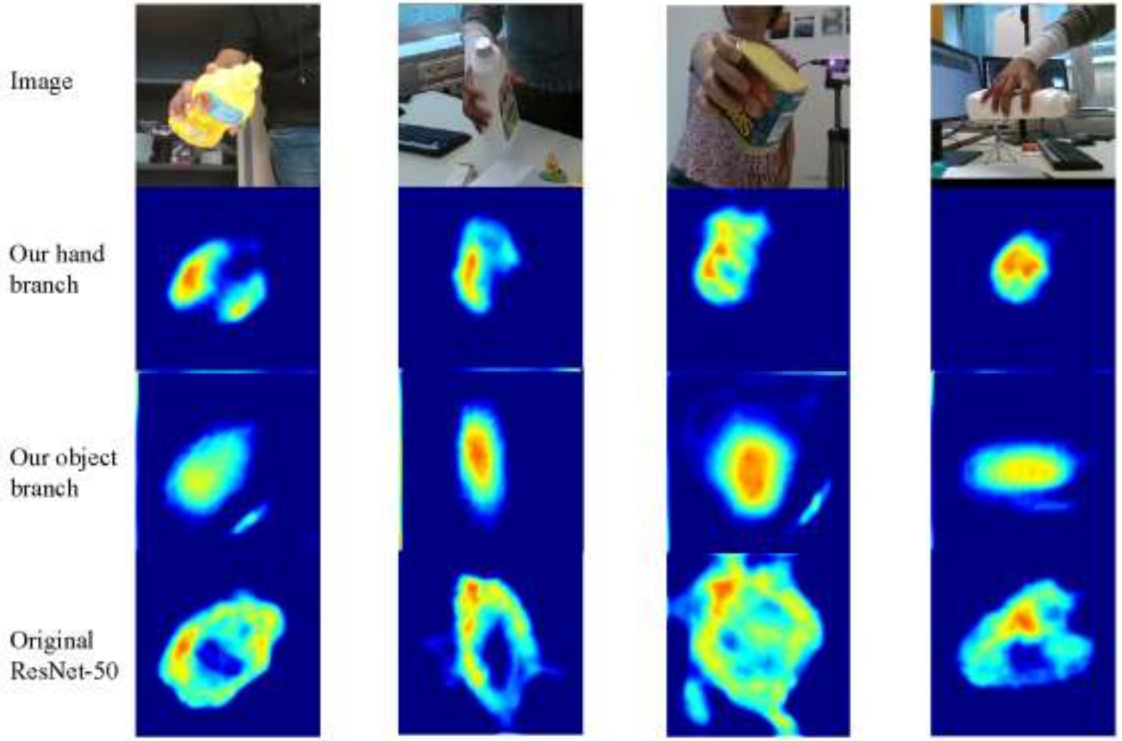
- The two sets of feature maps output by the stage-3 layers are fed into the same stage-4 layers.

- Finally, we adopt Feature Pyramid Network (FPN) to combine the features in different scales.

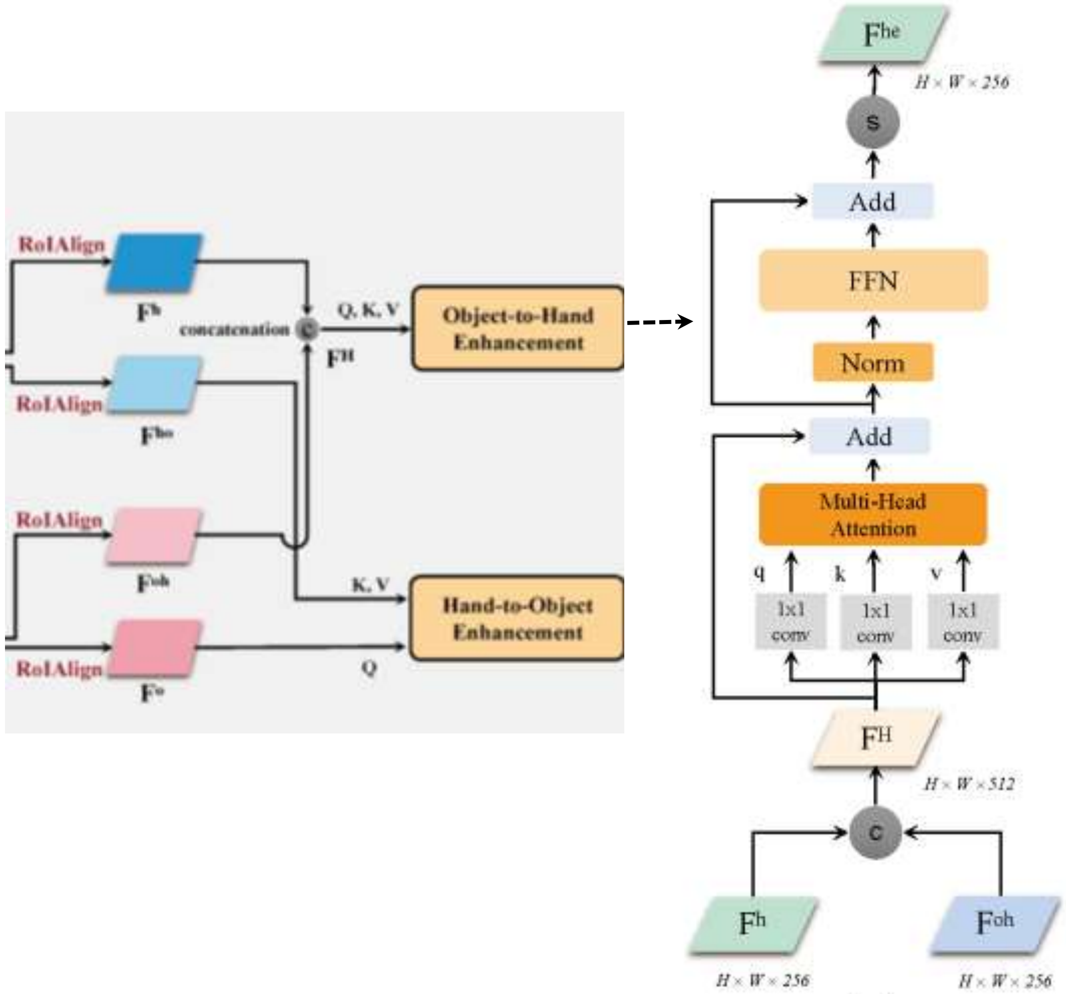
1.Feature Extraction Backbone



- independent stage-2 and stage-3 layers -> regard the hand and object respectively as the sole foreground target
- shared stage-4 layers -> the hand and object features are forced to be in similar feature spaces



2.Interaction Modules

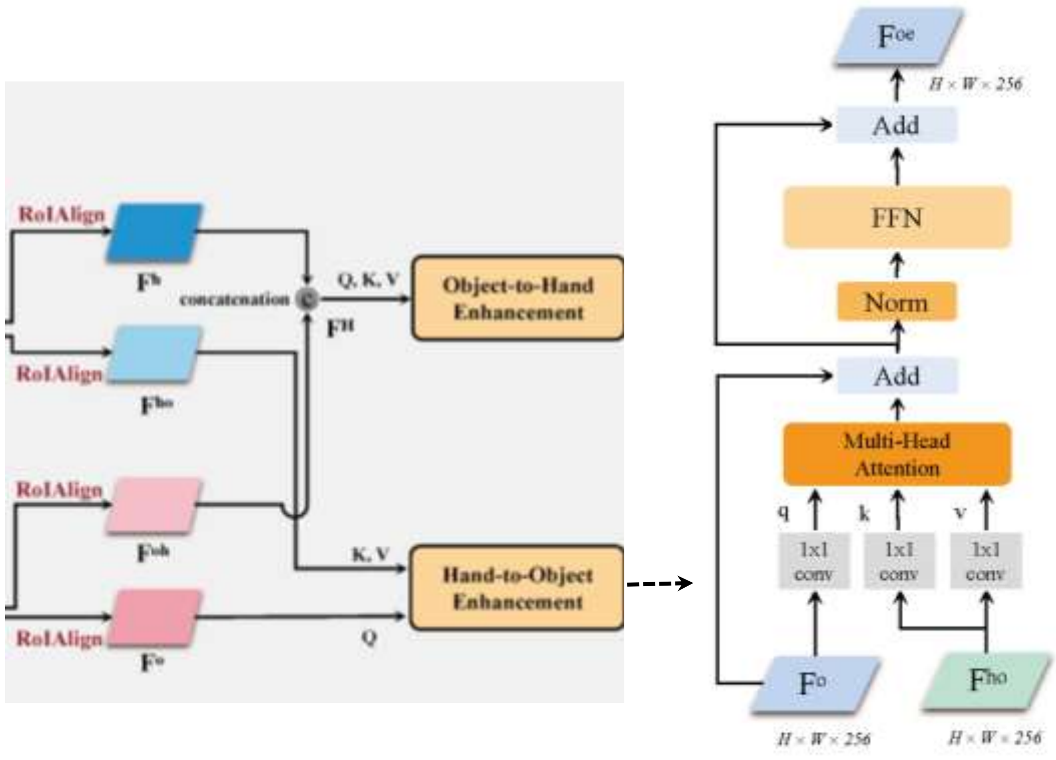


(a)

- hand-> non-rigid, flexible, high degree of freedom

- We use the ROIAlign to obtain F^h and F^{oh} from P^h and P^o, according to the hand bounding box.
- And concatenating them along the channel dimension to get F^H.
- Finally, We feed F^H into the Object-to-Hand Enhancement module.

2.Interaction Modules

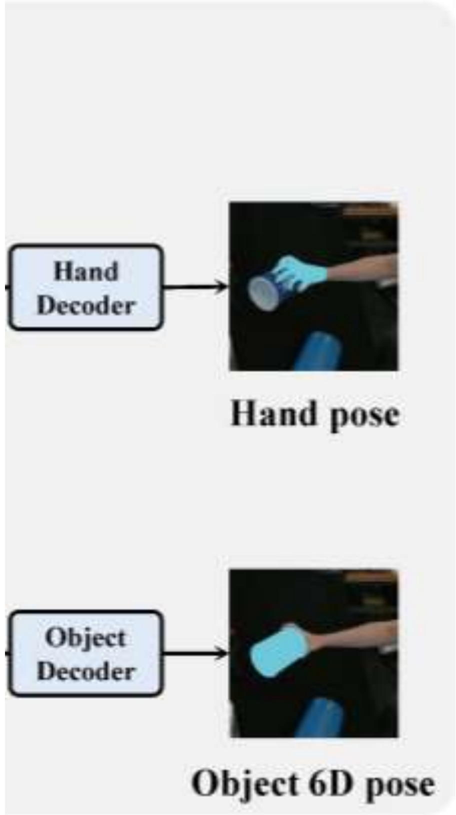


(b)

- object-> rigid, and less flexible

- We use the ROIAlign to obtain F^{O} from P^{o}, according to the object bounding box, and obtain F^{ho} from P^{h}, according to the overlapped area between the hand and object bounding boxes.
- Finally, We feed F^{o} and F^{ho} into the Hand-to-Object Enhancement module.

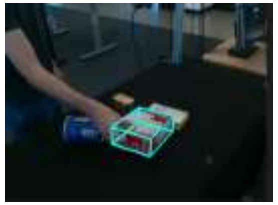
3.Two separate decoders



- Hand decoder output 2D joints, 3D mesh

3D hand mesh parameterized by MANO model

- Object decoder output 2D control points



21 control points pre-defined on object mesh 6D object pose computed by PNP algorithm

Experiments

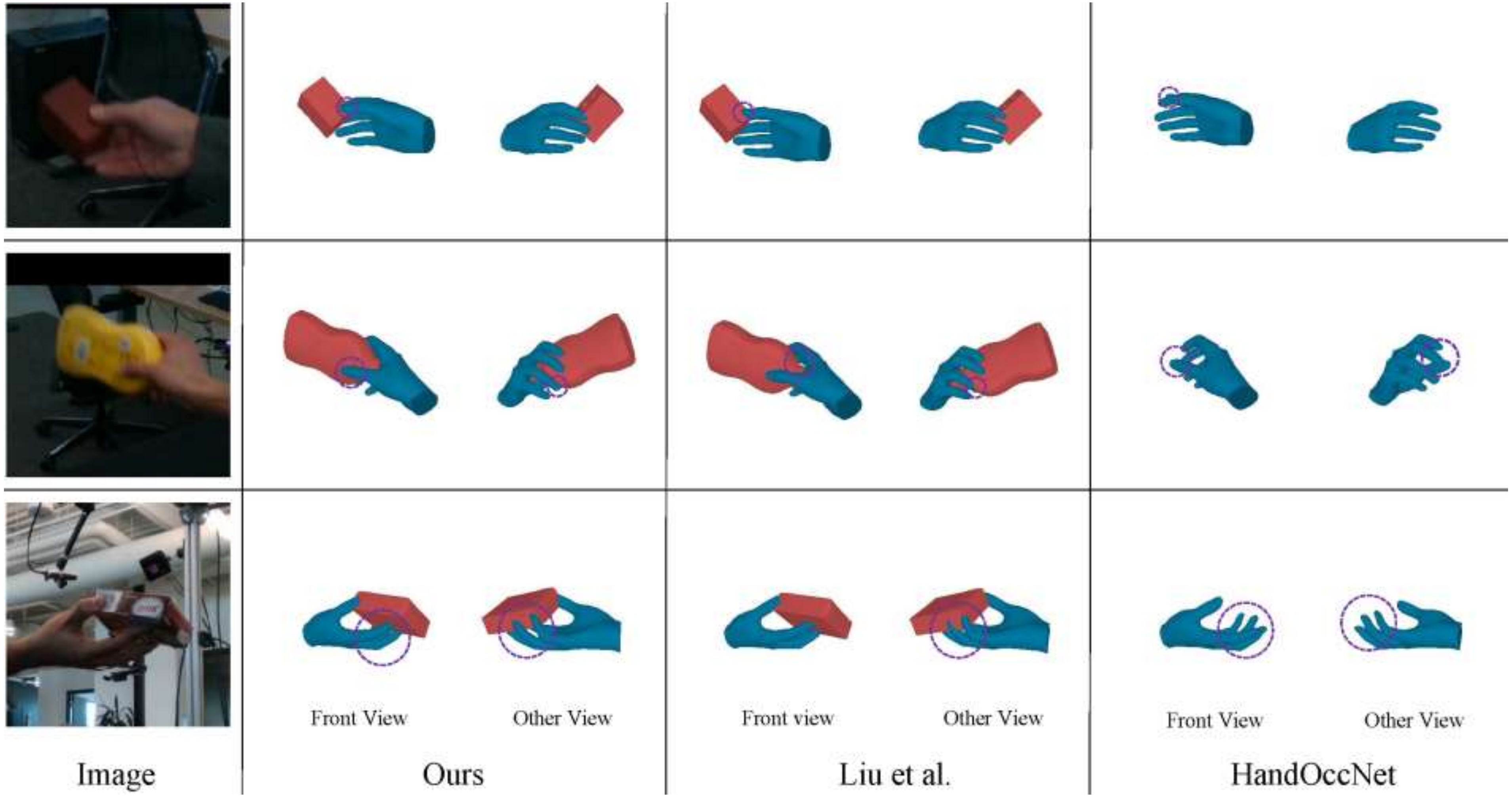
Methods	Joint↓	Mesh↓	cleanser↑	bottle↑	can↑	average↑
Single-Stream	10.4	10.3	80.1	55.3	46.2	60.5
Double-Stream	9.7	9.6	82.2	74.1	49.4	68.6
Ours	9.8	9.7	84.1	70.3	48.2	67.5

interaction modules

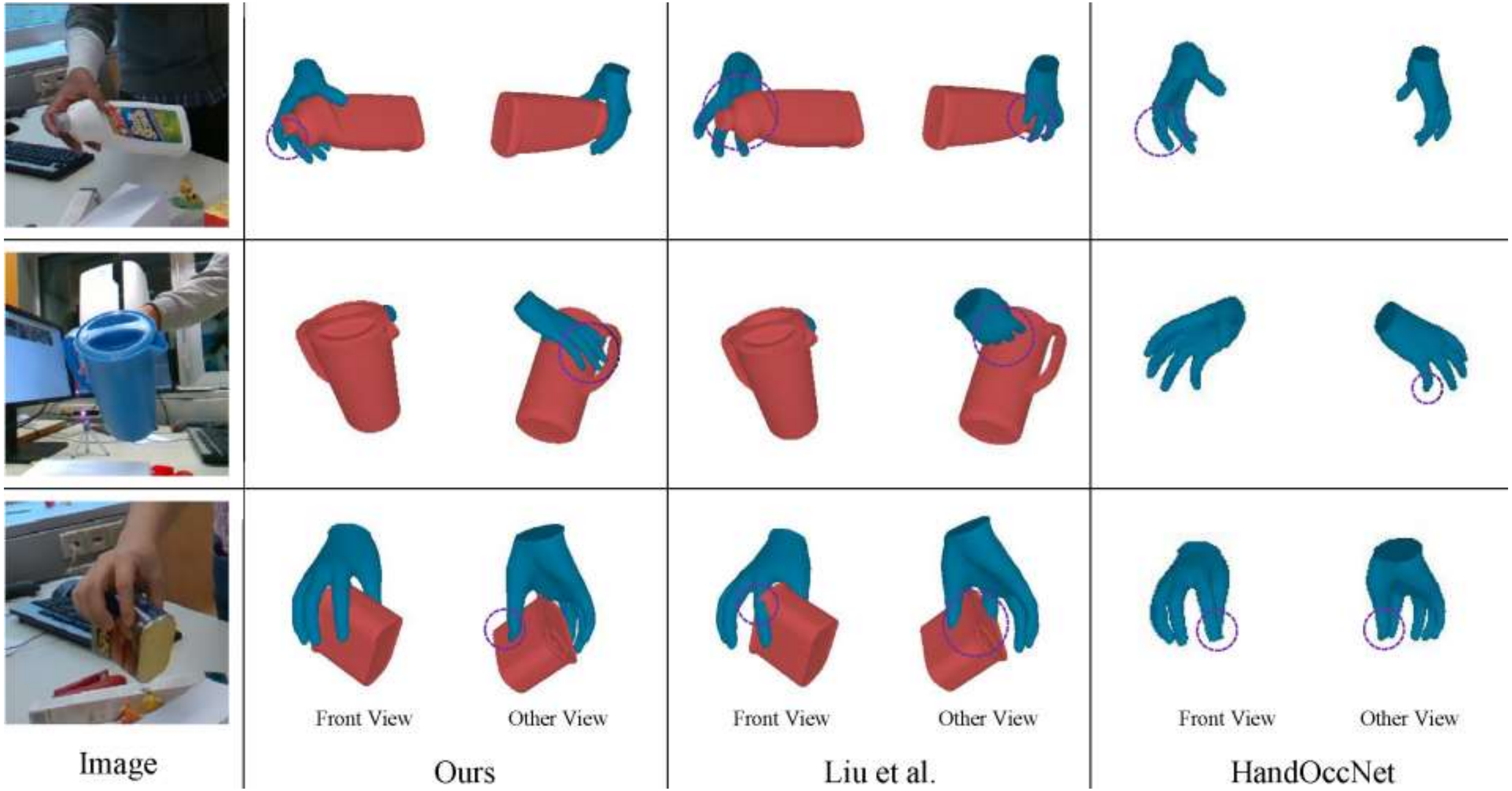
Methods	Joint↓	Mesh↓	cleanser↑	bottle↑	can↑	average↑
Single-Stream	10.2	10.0	86.2	62.1	42.3	63.5
Double-Stream	9.5	9.4	**91.2**	73.3	46.8	70.4
Ours	**8.9**	**8.7**	81.4	**87.5**	**52.2**	**73.3**

- The double-stream backbone works better than the single-stream without adding interaction modules, while our approach achieves close to the double-stream effect by adding only a small number of parameters.
- The performance gain of the double-stream backbone after adopting the interaction modules are quite small, while our approach has a larger improvement.

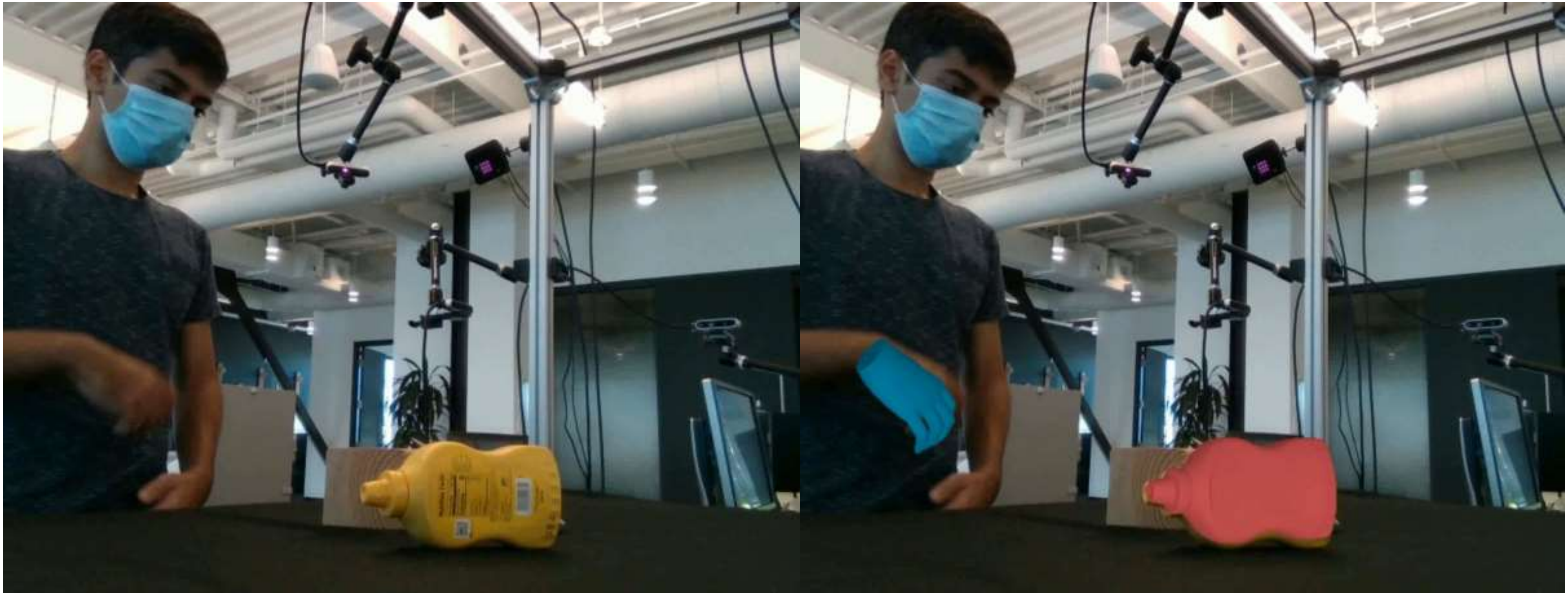
Qualitative examples:



Qualitative examples:



Qualitative examples:



Qualitative examples:

Thanks for listening

code:https://github.com/lzfff12/HFL-Net