Transformer-based Unified Recognition of Two Hands Manipulating Objects – Supplementary Material

Hoseong Cho Chanwoo Kim Jihyeon Kim Seongyeong Lee
Elkhan Ismayilzada Seungryul Baek

UNIST, South Korea

In this supplemental page, we detail the overall training procedure (Sec. 1), provide more results on the confusion matrix with and without the contact map representation (Sec. 2) and provide additional quantitative and qualitative results (Sec. 3). Additionally, more qualitative results are provided with the accompanied supplemental video. In the video, Kwon et al. [2]'s results are isolated from authors' video and compared to ours.

1. Overall training procedure

We present the details of out network architecture in Tables 1, 2, 3, and 4. The overall training procedure is summarized in Algorithm 1.

Layer	Operation	Dimensionality
	Input: $\mathbf{s}$	10845×256
1	Deformable self-attention	10845×256
2	Linear + ReLU+ Dropout	10845×1024
3	Linear + Dropout + LayerNorm	10845×256

Table 1. Encoder layer of pose estimation network f^{HOP}.

Layer	Operation	Dimensionality
	Input: $\mathbf{z}$	300×256
1	Self-attention	300×256
2	Deformable cross-attention	300×256
3	Linear + ReLU + Dropout	300×1024
4	Linear + Dropout + LayerNorm	300×256

Table 2. Decoder layer of pose estimation network f^{HOP}.

2. Effectiveness of using contact map

Fig. 1 shows the confusion matrices for interaction recognition according to the input modality. As we mentioned in Sec 3.2.1 of the main paper, *take out chips* is similar to *open chips* and *close chips* in the relative movement of hands and objects. Our model can better distinguish interactions if we use the contact map for interaction recognition.

Layer	Operation	Dimensionality
	Input: $\mathbf{z}'$	300×256
1	Linear + ReLU	300×256
2	Linear + ReLU	300×256
3	Linear	300×63

Table 3. Architecture of prediction head f_{hand}, f_{obj}.

Layer	Operation	Dimensionality
	Input: $\mathbf{v}$	$t \times 308$
1	LayerNorm + Self-attnetion	$(t+1) \times 308$
	Linear + GeLU + Dropout	$(t+1) \times 1232$
	Linear + Dropout	$(t+1) \times 308$
2	LayerNorm + Self-attnetion	$(t+1) \times 308$
	Linear + GeLU + Dropout	$(t+1) \times 1232$
	Linear + Dropout	$(t+1) \times 308$
3	LayerNorm + Self-attnetion	$(t+1) \times 308$
	Linear + GeLU + Dropout	$(t+1) \times 1232$
	Linear + Dropout	$(t+1) \times 308$
4	LayerNorm + Linear	1×36

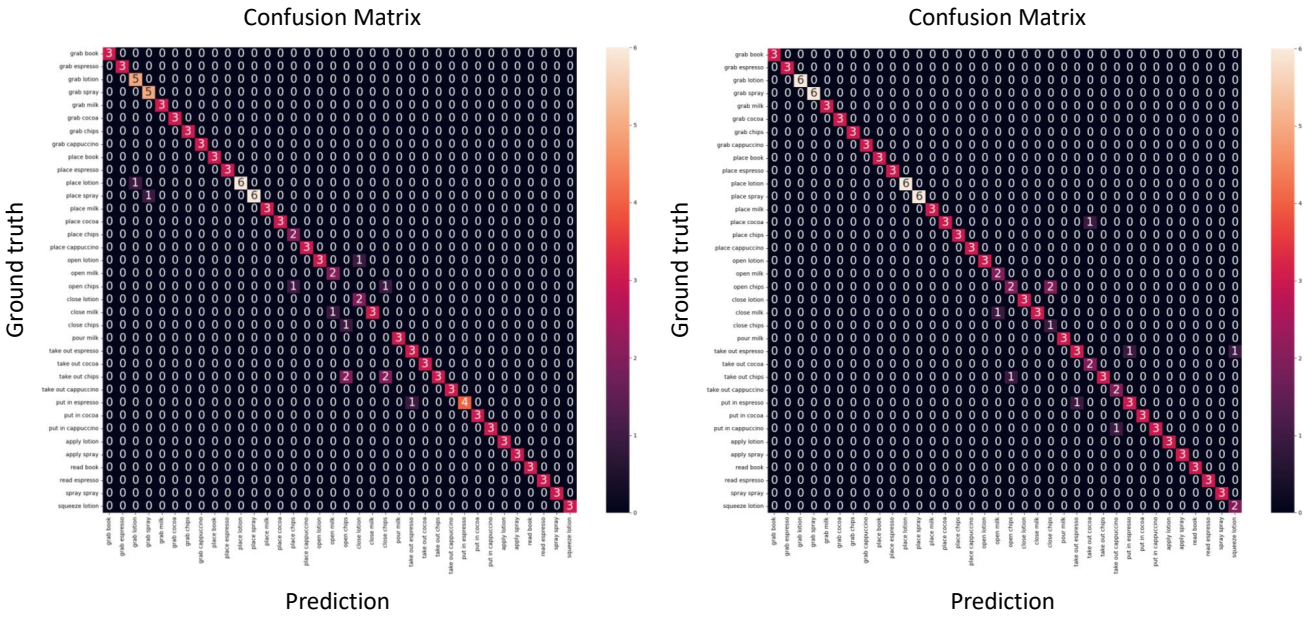
Table 4. Architecture of interaction recognition network f^{IA}.

	Left.h	Right.h	Object	Acc.
HOI4D	25.8	19.4	51.3	91.1

Table 5. The quntitative result on HOI4D dataset.

3. Additional results

In Fig 2, 3 and 4, we propose more qualitative results on H2O [2] and FPHA [1] datasets. These qualitative results contain the visualization for pose estimation, interaction classification, contact map and 3D meshes. As a supplementary experiment, we tested our model on the HOI4D [3] dataset, which is a large-scale egocentric dataset that provides 3D hand pose, object 6D pose, object type, and inter-



(a) without contact map representation

(b) with contact map representation

Figure 1. Confusion matrix for interaction recognition.

action class. As the testing data of HOI4D is not publicly available, we split the available training data into the new 'train' and 'test' with a 8:2 ratio. Table 5 shows the reconstructed accuracy of left and right hands (mm), object poses (mm) and interaction recognition accuracy (%).

References

[1] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In CVPR, 2018. 1

[2] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In CVPR, 2021. 1

[3] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In CVPR, 2022. 1

Algorithm 1: The summary of our entire training process

Stage 1.
Input: Image I
Output: Hand, object poses and object type (h, o, c)
for $n = 1, \ldots, N$ **do**
  Extract the multi-scale feature map $\mathbf{s}$ from backbone.
  Feed them to encoder of f^{HOP} and obtain refined feature $\mathbf{s}''$.
  Feed the object queries $\mathbf{z}$ to decoder of f^{HOP} and obtain refined feature $\mathbf{z}'$ by deformable cross-attention with $\mathbf{s}''$.
  Feed them to f_{hand}, f_{obj}, f_{cls} to get hand pose $\mathbf{h}$, object pose $\mathbf{o}$, and object type $\mathbf{c}$.
  Caculate gradient ∇L_{H} and update f^{HOP}.
end

Stage 2.
Input: Hand, object poses and object type (h, o, c)
Output: Interaction class a
for $n = 1, \ldots, N'$ **do**
  for $t = 1, \ldots, T$ **do**
    Obtain MANO pose parameter θ by inverse kinematics and generate the left and right hand meshes $\mathbf{V}_t^{\text{Left}}, \mathbf{V}_t^{\text{Right}}$.
    Calculate the object 6D pose applying the rigid alignment between the predicted object pose and the GT object pose.
    Sample 2000 object vertices $\mathbf{V}_t^{\text{O}}$ and transform the vertices from object space to camera space.
    Transform left and right hand meshes from MANO space to camera space.
    Generate the contact map $\mathbf{m}_t^{\text{Left}}, \mathbf{m}_t^{\text{Right}}, \mathbf{m}_t^{\text{Obj}}$ based on distance.
    Generate $\mathbf{v}_t$ by concatenating the two hands and object mesh with contact map.
  end
  Concatenate $\mathbf{v}$ with action token α and feed them to f^{IA} to get the interaction class $\mathbf{a}$
  Caculate gradient ∇L_{action} and update f^{IA}
end

Figure 2. Examples of hand-object poses and interacting classes on H2O dataset predicted by our H2OTR. (Col 1) Input frame, (Col 2) Contact map in interaction space, (Col 3 - 5) Contact map in canonical poses, (Col 6) Estimated hand/object poses and interacting classes. (Col 7) Mesh in interaction space, (Col 8) Mesh in other viewpoint.

Figure 3. Examples of hand-object poses and interacting classes on H2O dataset predicted by our H2OTR. (Col 1) Input frame, (Col 2) Contact map in interaction space, (Col 3 - 5) Contact map in canonical poses, (Col 6) Estimated hand/object poses and interacting classes. (Col 7) Mesh in interaction space, (Col 8) Mesh in other viewpoint

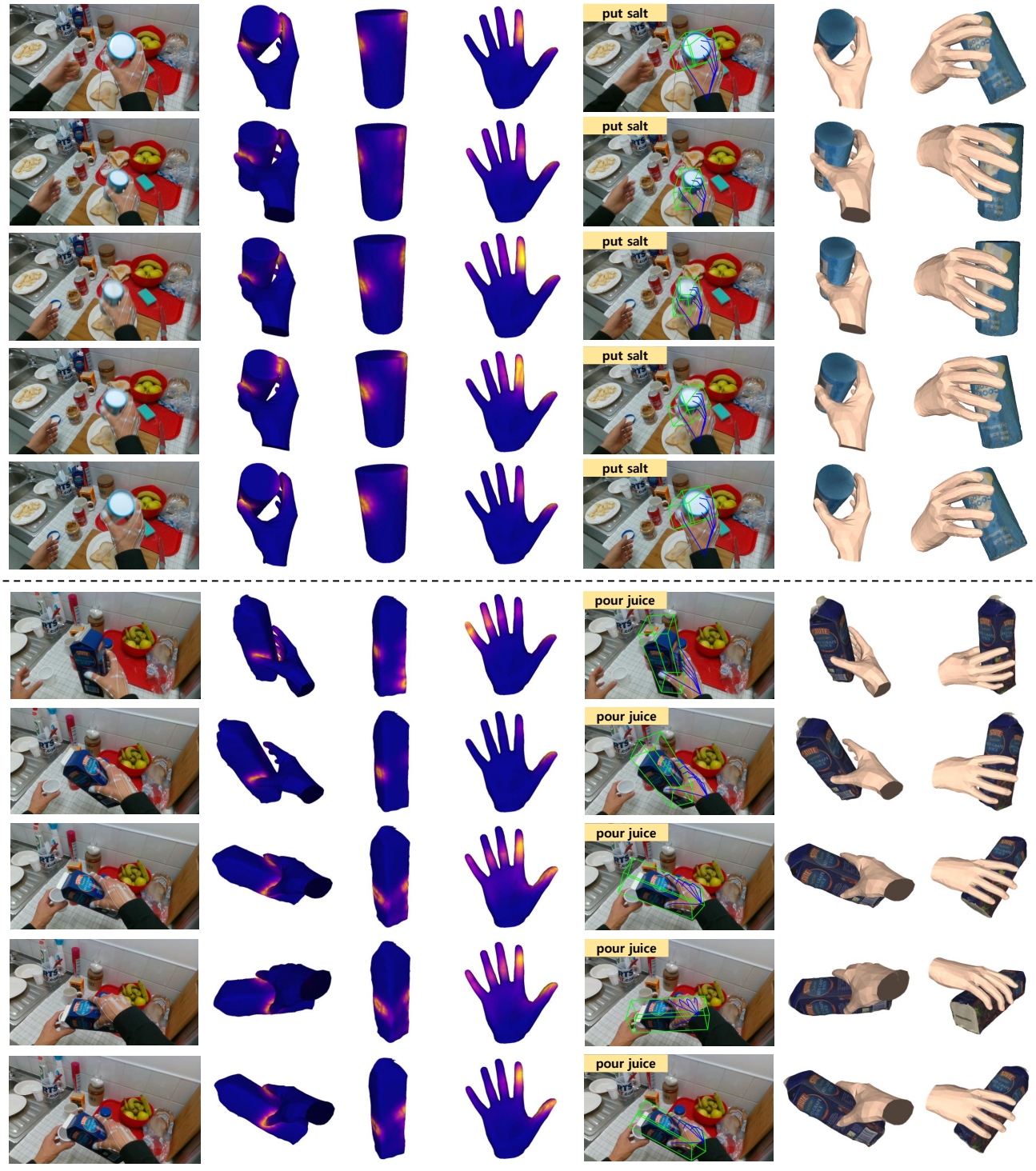


Figure 4. Examples of hand-object poses and interacting classes on FPHA dataset predicted by our H2OTR. (Col 1) Input frame, (Col 2) Contact map in interaction space, (Col 3, 4) Contact map in canonical poses, (Col 5) Estimated hand/object poses and interacting classes. (Col 6) Mesh in interaction space, (Col 7) Mesh in other viewpoint