# Supplementary Material for "Human-Object Interaction Detection with Missing Objects"

Kaen Kogashi, Yang Wu, Shohei Nobuhara, Ko Nishino
Kyoto University, Japan
https://vision.ist.i.kyoto-u.ac.jp

## 1 Supplementary Experimental Results

**Ablation Studies on HOI-MO-Net Components.** As shown in Table 1, every component of our HOI-MO-Net can bring a performance increase on Mixed-V-COCO test sets. For HOI-MO Data Augmentation, due to the difficulties of annotating HOI-MO activities, we design an HOI-MO data augmentation strategy to generate pseudo HOI-MO samples. We also use a semantic self-attention module that can help exclude background noises and extract finer self-attention based contextual representation. We add stuff context into the ablation study because our stuff module comes from the panoptic backbone and there are no other conventional methods considering the panoptic backbone's stuff feature. The experimental results show that stuff is useful for human-object interaction detection. Instance masks that contain geometric features can also benefit network performance.

**Ablation Studies on the Backbone.** As shown in Table 2, most conventional methods' backbones are Faster-RCNN, so they only generate instance representation at a bounding box level. We want to have more precise and detailed information, so we adopted the panoptic backbone in our proposed method. For the experimental results, with Faster R-CNN, bounding boxes are used as instance masks and stuff is unused. Results shown in Table 2 demonstrate the effectiveness of the panoptic backbone.

**Other visual examples.** In Figure 1 we show some qualitative results for comparing the performance of the proposed whole HOI-MO-Net model with its baseline version. The presented examples have great variations in object sizes, human sizes, and different HOI activity categories. The whole model performs significantly better than the baseline model on both HOI-MO test sets.

## 2 Details on HOI-MO Test Sets Construction

**Visual Examples of HOI-MO Categories.** As mentioned in the main text, the HOI-MO cases (which can easily lead to object detection failures) we observed are categorized into six types: occlusion, truncation, rare type, small scale, transparency, and gray image. Some representative examples of each type/category are shown in Figure 3 and Figure 5 of the main text.

Table 1. Ablation study on model components with Mixed-V-COCO.

| Method | mAP |
|---|---|
| Baseline (human, object) | 36.18 |
| Baseline + HOI-MO Data Augmentation (HDA) | 44.35 |
| Baseline + HDA + Semantic Self-attention (SS) | 46.13 |
| Baseline + HDA + SS + Stuff | 47.78 |
| Baseline + HDA + SS + Stuff + Instance Masks | **48.76** |

Table 2. Ablation study on backbones with Mixed-V-COCO.

| Method | mAP |
|---|---|
| Faster-RCNN (ResNet-50) [2] | 46.56 |
| Panoptic (ResNet-50) [1] | **48.76** |

In addition to that, more examples can also be seen in Figure 1 of this supplementary material.

**HOI-MO Category Annotation.** We found that some HOI-MO samples can be assigned to more than one HOI-MO category (e.g. occlusion and truncation) due to the co-existence of some factors. Though theoretically, one sample may belong to several HOI-MO categories, we found that it is very rare that the number of co-existing categories goes beyond 3. Though assigning multiple labels to the same sample can be more precise, choosing a single main category can be a better choice in practical applications. Therefore, we predefined a priority rank as shown in Table 3 for choosing the highest priority category label when there are coexisting candidates. The reasons for such a ranking result are as follows. Losing the color information usually influences the object detection performance very much and the significance is probably greater than any other factors, so we rank "gray image" the highest. Scale invariance is another great challenge, so "small scale"

Table 3. Label Assignment Priority Ranking of HOI-MO Categories.

| Priority Rank | 1 | 2 | 3 |
|---|---|---|---|
| Category | gray image | small scale | truncation |
| Priority Rank | 4 | 4 | 6 |
| Category | transparency | rare type | occlusion |

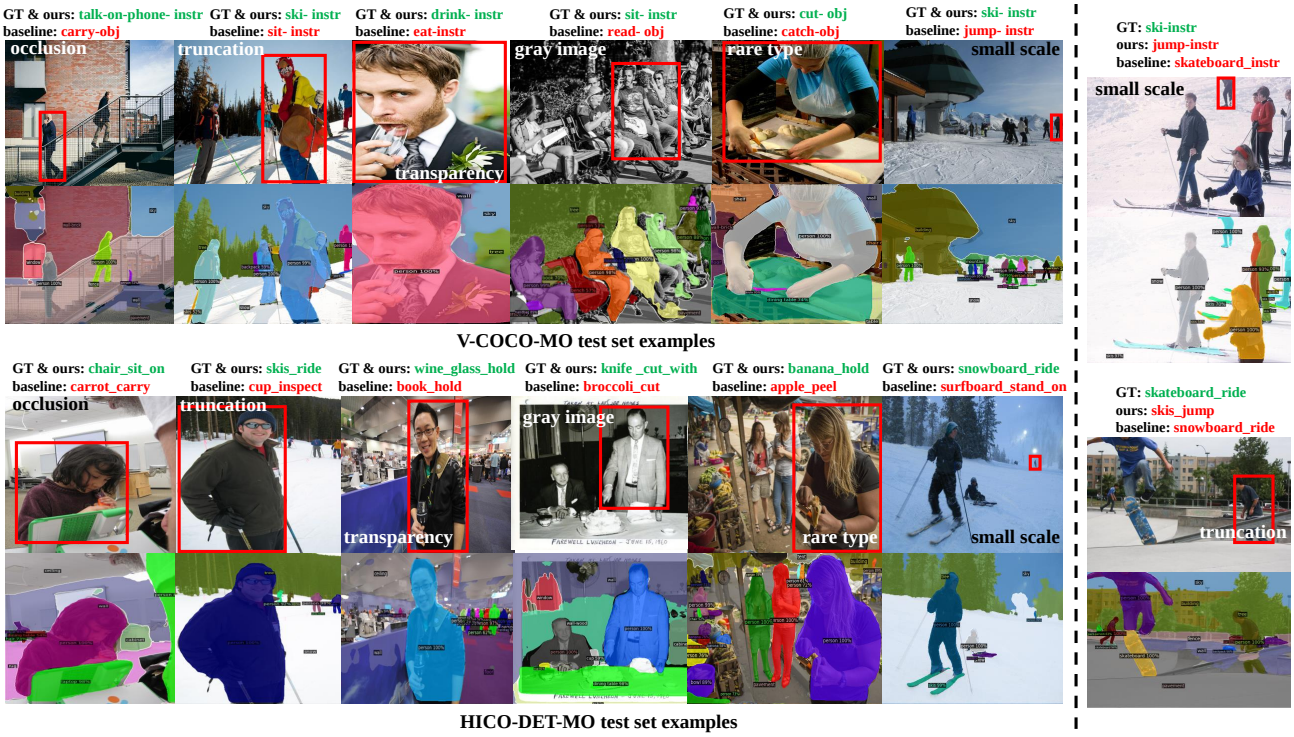


Figure 1. Visual examples of the HOI detection results of our whole HOI-MO-Net and our baseline model. For each example, the upper part shows the original image with red bounding boxes denoting the human detection result. The lower part shows the panoptic segmentation results overlaid on the original image. Green results are either the ground-truths or our predictions. The baseline model's results are shown in red. We can see that our whole HOI-MO-Net model's results are quantitatively and qualitatively much better. Input images are from V-COCO-MO and HICO-DET-MO test sets, showing the six HOI-MO categories from left to right. We also added some failure cases of our model on the right-most side, for which the ground-truths are shown in green and all prediction results are shown in red.

is ranked the next. Significant "truncation" happens at image borders and can also co-exist with a large scale, so it is the next challenging category. "Occlusion" means the object is still partially visible and in many cases the detector can still get assisted by the contextual information, so it is ranked the last, following "transparency" and "rare type", the two categories that can influence the appearance of the whole object. Though "transparency" may co-exist with "rare type" in the real world, we found that in V-COCO and HICO-DET datasets they never appear together, so we do not worry about their relative priority and just assign them the same rank.

## 3 Detailed Statistics of HOI-MO Test Sets

The sample and HOI activity category distributions of the two HOI-MO test sets are shown in Table 4. A more detailed sample distribution of the larger dataset HICO-DET-MO w.r.t. each of its HOI activities is shown in Figure 2, which is a long tail, similar to its superset HICO-DET dataset. Detailed HOI activity (verb or object-verb combination) names and their corresponding sample sizes are listed in Table 5.

Table 4. The number of samples (NS) and the number of original HOI Activity Categories (NC) for each HOI-MO type/category.

| HOI-MO Category | NS-VCOCO | NC | NS-HICO | NC |
|---|---|---|---|---|
| occlusion | 1360 | 18 | 2080 | 86 |
| truncation | 198 | 7 | 331 | 31 |
| transparency | 16 | 2 | 22 | 12 |
| gray image | 30 | 5 | 293 | 30 |
| rare type | 258 | 20 | 618 | 127 |
| small scale | 514 | 9 | 228 | 34 |
| Total | 2376 | 22 | 3572 | 155 |

Table 5. Detailed HOI activity category (verb or object-verb combination) names. "NS" denotes the number of samples.

| V-COCO-MO | | HICO-DET-MO | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Verb | NS | Object | Verb | NS | Object | Verb | NS | Object | Verb | NS |
| sit instr | 606 | chair | sit on | 1287 | surfboard | hold | 9 | apple | cut | 2 |
| ski instr | 576 | sports ball | hit | 142 | surfboard | lie on | 9 | sandwich | cut | 2 |
| snowboard instr | 519 | bench | sit on | 120 | tie | wear | 9 | apple | eat | 2 |
| surf instr | 126 | bicycle | sit on | 104 | sports ball | block | 7 | banana | eat | 2 |
| hit obj | 125 | bicycle | ride | 103 | bed | lie on | 7 | donut | eat | 2 |
| hold obj | 98 | bicycle | straddle | 103 | cow | walk | 7 | pizza | eat | 2 |
| ride instr | 67 | knife | cut with | 95 | suitcase | carry | 6 | banana | hold | 2 |
| carry obj | 55 | sports ball | catch | 89 | suitcase | hold | 6 | bicycle | hold | 2 |
| skateboard instr | 52 | knife | hold | 79 | tennis racket | hold | 6 | bowl | hold | 2 |
| eat obj | 28 | baseball bat | swing | 73 | sheep | pet | 6 | donut | hold | 2 |
| catch obj | 24 | skis | stand on | 66 | tennis racket | swing | 6 | surfboard | load | 2 |
| eat instr | 21 | skis | ride | 62 | suitcase | drag | 5 | cake | make | 2 |
| cut instr | 17 | skis | wear | 56 | sheep | feed | 5 | oven | operate | 2 |
| cut obj | 15 | couch | sit on | 53 | bed | sit on | 5 | orange | pick | 2 |
| jump instr | 14 | bird | feed | 51 | surfboard | sit on | 5 | pizza | pick up | 2 |
| drink instr | 13 | bird | watch | 50 | knife | stick | 5 | suitcase | pick up | 2 |
| lay instr | 8 | motorcycle | ride | 43 | bottle | carry | 4 | sports ball | serve | 2 |
| talk on phone instr | 3 | motorcycle | sit on | 43 | donut | carry | 4 | wine glass | sip | 2 |
| throw obj | 3 | motorcycle | straddle | 43 | giraffe | feed | 4 | knife | wield | 2 |
| work on computer instr | 3 | handbag | carry | 33 | sheep | herd | 4 | banana | buy | 1 |
| hit instr | 2 | handbag | hold | 31 | bottle | hold | 4 | bicycle | carry | 1 |
| read obj | 1 | bird | chase | 30 | cup | hold | 4 | person | carry | 1 |
| | | skateboard | ride | 30 | person | hug | 4 | pizza | carry | 1 |
| | | train | ride | 29 | apple | inspect | 4 | sandwich | cook | 1 |
| | | train | sit on | 29 | orange | inspect | 4 | carrot | cut | 1 |
| | | skateboard | stand on | 29 | skateboard | jump | 4 | broccoli | eat | 1 |
| | | cell phone | hold | 25 | cake | pick up | 4 | wine glass | fill | 1 |
| | | snowboard | stand on | 25 | spoon | sip | 4 | skateboard | flip | 1 |
| | | fork | hold | 24 | sports ball | throw | 4 | motorcycle | hold | 1 |
| | | snowboard | ride | 21 | cup | carry | 3 | oven | hold | 1 |
| | | snowboard | wear | 21 | skateboard | carry | 3 | sports ball | hold | 1 |
| | | cell phone | carry | 19 | scissors | cut with | 3 | dog | hug | 1 |
| | | fork | lift | 19 | bottle | drink with | 3 | knife | lick | 1 |
| | | couch | lie on | 17 | sandwich | eat | 3 | suitcase | load | 1 |
| | | backpack | carry | 16 | orange | hold | 3 | book | open | 1 |
| | | baseball bat | hold | 16 | pizza | hold | 3 | bottle | open | 1 |
| | | spoon | hold | 16 | sandwich | hold | 3 | oven | open | 1 |
| | | backpack | wear | 16 | scissors | hold | 3 | scissors | open | 1 |
| | | backpack | hold | 15 | wine glass | hold | 3 | motorcycle | park | 1 |
| | | cake | hold | 15 | banana | inspect | 3 | bird | pet | 1 |
| | | cell phone | talk on | 15 | sports ball | kick | 3 | cow | pet | 1 |
| | | bus | ride | 14 | sandwich | make | 3 | dog | pet | 1 |
| | | bus | sit on | 14 | cake | no interaction | 3 | apple | pick | 1 |
| | | cup | drink with | 13 | donut | pick up | 3 | banana | pick | 1 |
| | | banana | carry | 12 | boat | ride | 3 | bottle | pour | 1 |
| | | cake | eat | 12 | surfboard | ride | 3 | motorcycle | push | 1 |
| | | cow | herd | 12 | boat | row | 3 | book | read | 1 |
| | | sports ball | carry | 11 | boat | sit on | 3 | car | ride | 1 |
| | | cake | carry | 10 | sports ball | spin | 3 | wine glass | toast | 1 |
| | | sports ball | dribble | 10 | surfboard | stand on | 3 | horse | train | 1 |
| | | cell phone | text on | 10 | apple | buy | 2 | horse | walk | 1 |
| | | cake | cut | 9 | orange | buy | 2 | | | |

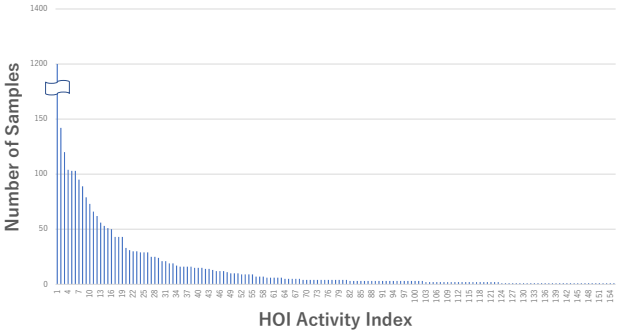


Figure 2. HICO-DET-MO's long tail sample distribution of the HOI activity categories.

## References

[1] A. Kirillov, R. Girshick, K. He, and P. Dollár. Panoptic feature pyramid networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6392–6401, 2019.

[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.