O1-2-1

17th International Conference on Machine Vision Applications (MVA)
Fully Online, July 25–27, 2021.

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

Abstract

Existing studies on human-object interaction (HOI) assume that human and object instances can be detected. This paper proposes a more practical HOI detection method for when object instances are not necessarily easily detectable. To our knowledge, we introduce the first method for such challenging HOI detection that incorporates global scene information. The two most widely used public HOI benchmark datasets are shown to contain many cases of HOI with missing objects (HOI-MO). We label these to compose new test sets for the proposed method. The effectiveness and superiority of the proposed method are demonstrated through extensive experiments and comparisons.

1 Introduction

Human-object interaction (HOI) is an important task which finds applications in a wide range of fields including human-robot collaboration and smart environments. Past methods mainly follow the conventional process of first detecting human and object instances and then recognizing their interactions. As a result, HOI activities can only be detected for cases where human and object instances are accurately detected. HOI activities that do not necessarily have clear views of the target objects, which we refer to HOI with missing objects (HOI-MO) cases, are common in the real world. We found that even in widely adopted public HOI detection datasets (V-COCO and HICO-DET) which are carefully constructed to avoid HOI-MO cases, there still is a large amount of HOI-MO instances (more than 10%). Humans can easily interpret these HOI-MO cases and it is of paramount importance to make computers detect them accurately.

Why can then humans easily recognize human-object interactions even when the people or objects are not discernible in the image? This is likely because our perception leverages global scene context in the image. That is, we can extrapolate the constituents of the interaction from the overall scene context and correctly infer their relation.

In this paper, we propose a novel HOI detection method that realizes this ability. To our knowledge, our method is the first solution for *human-object interaction with missing objects (HOI-MO)* detection. Figure 1 shows two concrete examples of HOI-MO detection. The key idea is to fully leverage background scene information. The method builds upon a panoptic segmentation backbone whose middle-level representation for human and object instances and background stuff are leveraged to detect HOI with challenging (*e.g.*, occluded) and missing objects. Experimental results on two new HOI-MO test sets validate its effectiveness. Each test set includes six HOI-MO categories as shown in Figure 3: (a) occlusion, (b) truncation, (c) rare type, (d) small scale, (e) transparency, and (f) gray image. We believe that our work can largely expand HOI's applications in real-world scenarios.

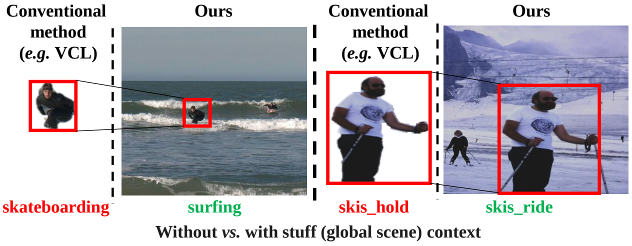
Conventional method (*e.g.* VCL)
Ours
Conventional method (*e.g.* VCL)
Ours
skateboarding
surfing
skis_hold
skis_ride
Without *vs.* with stuff (global scene) context

Figure 1. Existing HOI detection solutions can fail when the target objects cannot be detected. Our method leverages global scene context to reliably recognize such instances of human-object interaction with missing objects (HOI-MO).

2 Related Work

Context Modeling in HOI Detection. Contextual information plays a crucial role in improving the performance of many computer vision tasks such as object detection [13, 2], segmentation [4], and HOI detection. Conventional methods for HOI analysis use context in their models but implicitly [16, 18]. In contrast, we leverage panoptic segmentation to explicitly integrated per-pixel scene context.

Visual Cues for HOI Detection. In HOI detection, primary visual cues come from object and human detection results. Recent works have also explored the use of human pose [11, 15]. Our key idea is to leverage panoptic segmentation, for the first time for HOI detection, so that both stuff segmentation and instance segmentation can be integrated into context modeling and self-attention based instance representation.

Vision in the Wild. Recently, various studies have looked into more challenging (wilder) data for traditional vision tasks including face recognition [14], pose estimation [7], and object detection [3]. Some others

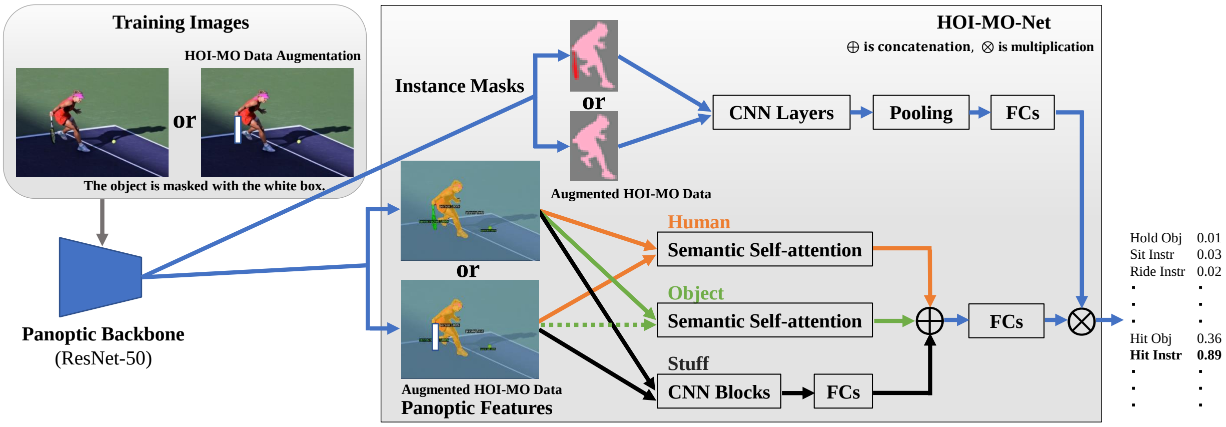

Figure 2. The overall architecture of the proposed HOI-MO-Net.

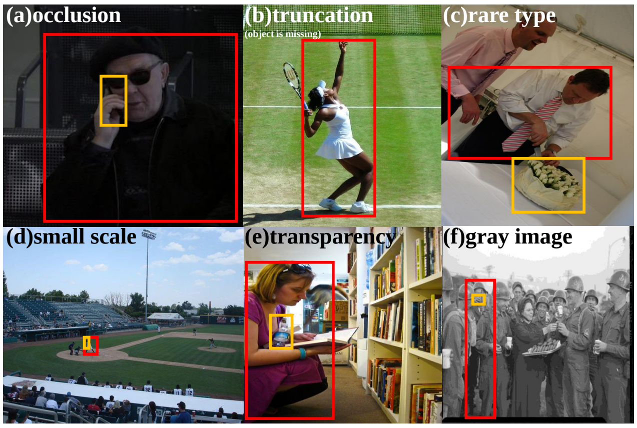

Figure 3. Six HOI-MO categories chosen in a data-driven manner are shown. The red bounding boxes indicate ground-truth human instances and the yellow bounding boxes depict object instances typically missed by object detectors, respectively.

even consider multiple tasks with the same wild data [19]. Existing data for evaluating HOI detection models still do not truly reflect instances in the wild. In this paper, we investigate this issue by proposing new HOI-MO test sets in addition to introducing a novel method for HOI detection that can handle HOI instances with target objects missed by object detectors.

3 Method

Panoptic Backbone. The novel HOI detection network is named **HOI-MO-Net** whose overall architecture is illustrated in Figure 2. The proposed HOI-MO-Net is based on a panoptic segmentation backbone. The backbone provides both semantic segmentation results for stuff regions and instance segmentation results for human and object instances. We choose the backbone from a simple yet effective panoptic segmentation model [9] pre-trained on MS COCO dataset [12], which shares the same original data source with V-COCO [6] and HICO-DET [1], the two most commonly used standard benchmarks for HOI detection.

HOI-MO-Net has the following four novel components that collectively achieve accurate HOI-MO detection.

HOI-MO Data Augmentation. Due to the difficulties of annotating HOI-MO activities which greatly limit the amount of labeled training data, we design an HOI-MO data augmentation strategy to generate pseudo HOI-MO samples. Since the key difference between normal data and HOI-MO data is whether the object instances are easily detectable or not, our strategy is to use the ground-truth bounding box of each object instance to mask out the object with constant white color, change the ground-truth segmentation maps, and assign null values to the corresponding bounding boxes of masked object instances. Although such HOI-MO data augmentation is only a crude approximation of real cases where objects are missing or undetected and does not directly correspond to the six HOI-MO types we are particularly interested in modeling (see Figure 3), the model can still learn how to explore other information for HOI detection in the absence of object instances in the image.

Semantic Self-attention. Unlike existing self-attention based models (*e.g.*, iCAN [5]), our semantic self-attention uses instance segmentation masks rather than instance bounding boxes for pooling the query instance features. Figure 4 shows a schematic of this self-attention module, where the "masked features" denote the instance features within its spatial segmentation mask rather than the bounding boxes. Such a semantic self-attention module can help exclude background noises and extract finer self-attention based contextual representation. The dotted line in Figure 2 heading to "Object Semantic Self-attention" corresponds to a special operation of filling a region with zeros when

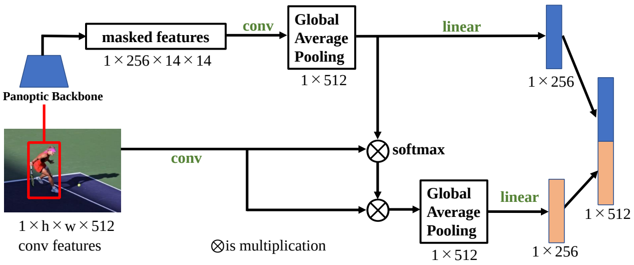

Figure 4. Semantic Self-attention module. Conv is a convolution layer. Linear applies a linear transformation to the input data.

object detection fails. A convolution layer and a fully connected layer are used to construct the semantic self-attention modules. A linear transformation is adopted to concatenate and fuse human and object features.

Stuff Context. Scene context, especially regarding stuff (*i.e.*, none-object regions), can be an informative cue for determining the human-object interaction in the image (see Figure 2 for an example). We use the semantic segmentation head described in [9] for modeling the stuff context. It takes the Feature Pyramid Network (FPN) features as input and merges information from all levels of the FPN into a single output. We extract stuff features with a simple CNN-based network branch for all the 54 MS COCO stuff categories.

Precise Spatial Configuration based on Instance Masks. The Spatial configuration of human and object instances has been proven to be useful for enhancing HOI detection in many existing works (*e.g.*, iCAN [5]). However, these works have only used the bounding boxes of the instances for extracting such information. HOI-MO-Net uses instance segmentation results from the panoptic backbone, which can lead to much higher precision.

4 Experimental Results

4.1 Datasets and Experimental Settings

Original Datasets. V-COCO [6] is a subset of the MS COCO dataset [12] with HOI annotations. V-COCO consists of a total of 10,346 images, of which 5,400 images are for training and validation, and the remaining 4,946 are for testing. Each person is annotated with 26 different actions. Each person can perform multiple actions at the same time, for example, jumping while snowing or surfing. HICO-DET [1] is the other dataset larger than V-COCO. It contains 600 HOI categories and over 80 object categories [12], with a total of 38,118 images for training and 9,658 images for testing.

HOI-MO Test Sets. Our HOI-MO test sets and V-COCO/HICO-DET share the same input images. We, however, add HOI instances with missing objects whenever applicable. For instance, for the middle images on the top row of Figure 3, the original annotation of HICO-DET consists of the racket as the object, the player as the person, and "tennis racket swing" as the HOI category. We add "sports ball hit" as the HOI category even though the ball is missing as the object. As a result, HOI-MO represents particularly hard cases as an additional test set.

The two HOI-MO test sets are named V-COCO-MO and HICO-DET-MO, which covers 22 V-COCO categories and 155 HICO-DET categories, respectively. In addition to these test sets, we also evaluate our method on mixed test sets that combine the original and the new HOI-MO test sets (named Mixed-V-COCO and Mixed-HICO-DET).

Evaluation metrics. We adopt the commonly used role mean average precision (role mAP) [10] for evaluation. If the predicted bounding boxes for human and objects both have IoUs ≥ 0.5 compared with the ground truth and the human-object interaction prediction score per action is also correct, then the prediction is considered as a true positive.

Implementation details. We use the panoptic backbone from Detectron2 [17] to generate human and object bounding boxes. We keep human and object instances whose box scores are higher than 0.5. We train our network for 30 epochs on each dataset with a learning rate of 0.001 and batch size 4. Training our network on V-COCO takes 12 hours on one Quadro RTX 8000 48G card. For HICO-DET, training the network on the train set takes 52 hours with the same GPU.

4.2 Results on HOI-MO Test Sets

For HICO-DET, we follow the settings in [1]: Full (600 HOIs), Rare (138 HOIs), Non-Rare (462 HOIs) in Default. For V-COCO, we evaluate the commonly used 24 actions. Our experiment results on mixed test sets are generally better than conventional methods. Table 1 shows that iCAN's results are better than TIN-net [11] and VCL [8] on HICO-DET-MO's Full and Non-Rare settings. This is because HICO-DET has 600 verbs. TIN-net and VCL are larger in model size and they are also more complex than iCAN, so they are easier to overfit on small data like HOI-MO test sets. Since V-COCO only has 24 verbs, performing HOI detection on it is simpler compared with HICO-DET. We used iCAN's code from its Github website. We deployed TIN-net and VCL's pre-trained models from its Github websites for evaluation.

4.3 Performance on Each HOI-MO Category

We present the per-category performance in Table 2. Our experimental results on HOI-MO categories are generally better than conventional methods. Conventional methods don't use the HOI-MO data augmentation strategy, moreover, transparency (e) and gray image (f) only have few samples, so their results are close

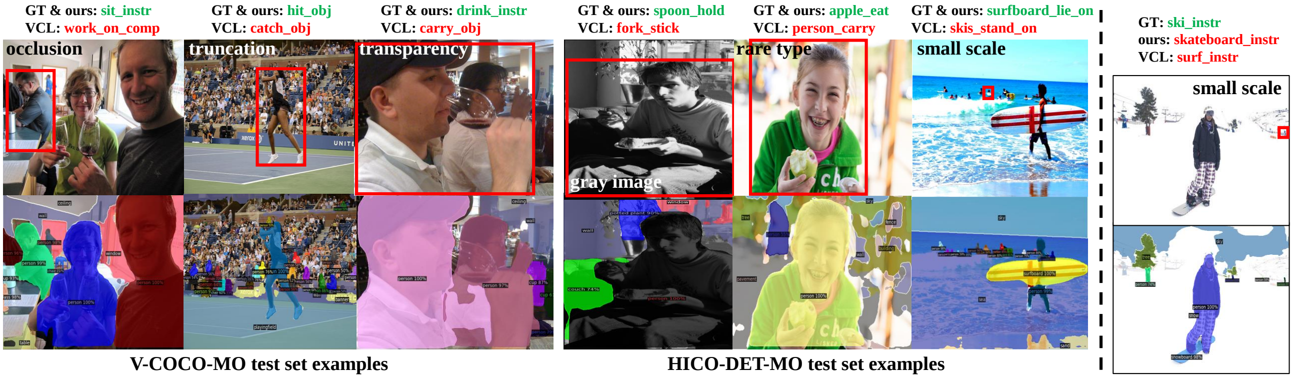

Figure 5. Examples HOI detection results of our HOI-MO-Net and VCL. For each example, the first row shows the original image with red bounding boxes denoting the human detection result. The second row shows the panoptic segmentation results overlaid on the original image. Green results are either the ground-truths or our predictions. VCL's results are shown in red. We can see that our model's results are more accurate. 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 show 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.

Table 1. Results on Mixed-V-COCO/HICO-DET and V-COCO/HICO-DET-MO test sets in mAP.

Method	Mixed-V-COCO	Mixed-HICO-DET			V-COCO-MO	HICO-DET-MO		
		Full	Rare	Non-Rare		Full	Rare	Non-Rare
iCAN [5]	38.06	13.93	10.35	15.11	5.37	5.42	4.30	5.67
TIN-net [11]	40.88	16.04	13.36	17.02	5.60	4.97	4.31	5.12
VCL [8]	41.07	18.00	16.36	18.75	8.45	4.70	4.55	4.74
HOI-MO-Net	**48.76**	**19.83**	**16.39**	**20.93**	**31.23**	**13.21**	**12.45**	**13.38**

Table 2. Performance on V-COCO-MO by individual HOI-MO categories(HC) in mAP, with '(a)'–'(f)' denoting the HOI-MO categories shown in Figure 3. 'NS' denotes the number of samples. 'Ours' denotes HOI-MO-Net.

HC	NS	iCAN [5]	TIN-net [11]	VCL [8]	Ours
(a)	1360	5.99	6.35	8.54	33.37
(b)	198	0.72	0.85	1.06	60.37
(c)	258	8.46	9.77	15.9	46.40
(d)	514	1.87	1.66	3.67	31.67
(e)	16	0.00	0.00	0.00	50.00
(f)	30	0.00	0.00	0.00	36.38

to zero. HOI-MO-Net seems to be especially good at "truncation," because in most truncation images the human body shape is clear, and the network can still judge the HOI activities. Compared with "truncation," performance on "occlusion" is lower because in most occlusion cases, not only the objects but also some human bodies are occluded. The performance on "small scale" is also low because in most cases human body shapes are ambiguous. "Transparency" and "gray images" are rare cases, but can still be detected by our proposed method.

4.4 Qualitative Results

Figure 5 shows qualitative results, comparing the proposed HOI-MO-Net with the VCL model. The images show the variance in object size, human size, and different interaction classes.

5 Conclusion and Future Work

In this work, we introduced HOI detection with missing objects, the task of HOI detection for challenging images in which objects are hard to detect, and derived a novel HOI method that leverages scene context. We built two test sets by using commonly used public benchmark datasets for HOI detection. Extensive experimental results demonstrate the effectiveness of our method. Possible straightforward future work is to explore a way for jointly optimizing the panoptic backbone and the HOI-MO-Net, so that these two may better cooperate and generate interpretable new middle-level representations.

Acknowledgments

This work was in part supported by JSPS 20H05951, JSPS 21H04893, and JST JPMJCR20G7.

References

[1] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2018.

[2] X. Chen and A. Gupta. Spatial memory for context reasoning in object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4106–4116, 2017.

[3] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018.

[4] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2393–2402, 2018.

[5] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018.

[6] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. In *arXiv preprint arXiv:1505.04474, 2015.*

[7] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.

[8] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *arXiv preprint arXiv:2007.12407, 2020.*

[9] 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.

[10] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari. Detecting visual relationships using box attention. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1749–1753, 2019.

[11] Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, and C. Lu. Transferable interactiveness knowledge for human-object interaction detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3580–3589, 2019.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *In ECCV*, 2014.

[13] Y. Liu, R. Wang, S. Shan, and X. Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2018.

[14] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[15] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He. Pose-aware multi-level feature network for human object interaction detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9468–9477, 2019.

[16] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, and J. Laaksonen. Deep contextual attention for human-object interaction detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5693–5701, 2019.

[17] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. 2019.

[18] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2010.

[19] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2886, 2012.