Discovering Human Interactions with Large-Vocabulary Objects via Query and Multi-Scale Detection

Suchen Wang[†] Kim-Hui Yap[†] Henghui Ding [†] Jiyan Wu[†] Junsong Yuan[‡] Yap-Peng Tan[†]
[†]Nanyang Technological University, Singapore [‡]State University of New York at Buffalo
{wang.sc, ekhyap, ding0093, jiyan.wu, eyptan}@ntu.edu.sg, {jsyuan}@buffalo.edu

Abstract

*In this work, we study the problem of human-object interaction (HOI) detection with large vocabulary object categories. Previous HOI studies are mainly conducted in the regime of limit object categories (*e.g.*, 80 categories). Their solutions may face new difficulties in both object detection and interaction classification due to the increasing diversity of objects (*e.g.*, 1000 categories). Different from previous methods, we formulate the HOI detection as a query problem. We propose a unified model to jointly discover the target objects and predict the corresponding interactions based on the human queries, thereby eliminating the need of using generic object detectors, extra steps to associate human-object instances, and multi-stream interaction recognition. This is achieved by a repurposed Transformer unit and a novel cascade detection over multi-scale feature maps. We observe that such a highly-coupled solution brings benefits for both object detection and interaction classification in a large vocabulary setting. To study the new challenges of the large vocabulary HOI detection, we assemble two datasets from the publicly available SWiG and 100 Days of Hands datasets. Experiments on these datasets validate that our proposed method can achieve a notable mAP improvement on HOI detection with a faster inference speed than existing one-stage HOI detectors. Our code is available at* `https://github.com/scwangdyd/large_vocabulary_hoi_detection`.

1. Introduction

Discovering human interactions with objects plays a vital role in human-centric visual understanding and provides a means to understand human intentions, actions, and activities. The goal is to detect one or multiple tuples <`human`, `verb`, `object`> to indicate the positions of human and objects within the image, and through the verb predict how they interact with each other (*e.g.*, holding something, repairing something, *etc*.).

While recent studies on human-object interaction (HOI) detection [1, 23, 52, 46, 9, 16, 20, 43, 40, 27] have achieved great progress, they have largely focused on the regime with a limited variety of objects (*e.g.*, 80 COCO objects [26]). In reality, humans can interact with a large variety of objects in our visual world. We can see various human interactions with daily objects from YouTube videos [8] or Internet data [36]. Nevertheless, it remains an under-explored problem for HOI detection in the regime of large-vocabulary object categories, where there are a large number of interactions and only a few data samples available for most interactions. In this setting, existing approaches would face new difficulties due to the greater diversity of objects and contexts.

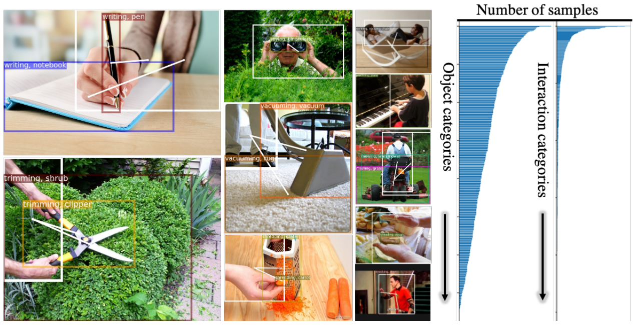

Figure 1: In this work, we aim to detect the human interactions with large-vocabulary object categories, where there are a large number of interactions and only a few data samples available for most categories.

The main goal of this work is to study the new challenges of discovering human interactions with large-vocabulary objects. For HOI detection, the common practice is to decompose the problem into two parts: (1) person and object instances detection; (2) instance matching and interaction classification. Depending on whether the two parts are conducted sequentially or in parallel, existing works can be further divided into the two-stage solutions [20, 19, 52, 40, 9] or end-to-end one-stage solutions [7, 23, 46, 17]. Regardless of which type of solutions, existing methods usually instantiate the object detection part as generic detectors, *e.g.*, Faster RCNN [34], CenterNet [5], RetineNet [25], etc. As previous HOI studies are mainly conducted in the same category space as COCO detection [26], it is viable to use the common architecture with the pre-training weights to ease the training and ensure the detection results. However, a new and large category space may pose challenges for HOI detection as the generic object detectors may perform poorly in the low-sample regime [12] and it is still an open problem to learn the effective detectors in large-vocabulary case. Different from the work of large-vocabulary object detection [12, 39, 45, 18, 37, 48], we would like to explore an effective HOI-specific solution to find the target objects by using the interaction clues.

Another challenge is the combinatorial explosion of the interactions when more object categories are involved. Given the combination nature of interactions, it is intuitively appealing to decouple the interaction detection into separate action and object predictions and then merge their scores to produce the final interaction score. Previous methods often approach this using completely separate or parallel branches. We observe that such an approach would become sub-optimal in the large-vocabulary scenario as the action and object predictions often conflict with each other and result in invalid combinations. Although the invalid combinations can be filtered out based on prior knowledge or external resources [50], we opine that invalid combinations can be effectively alleviated by coupling the interaction and object classification.

In this paper, we propose a new strategy to address the HOI task and formulate it as a query problem. As the person is often the dominant class, conventional detectors can usually give a reliable result. We thus first detect humans in the image and use them as queries to search for corresponding interactions and target objects. We aim to develop a unified model powered by Transformers [41] to jointly detect the objects interacting with the given human query and predict their interactions. This alleviates the need of using generic object detectors, multi-stream interaction recognition, and additional human-object instance matching process. The main idea is to find the region that may include the target objects by comparing the human query feature with the image feature at all sliding window positions. Then, we update the human query feature by progressively aggregating the context information from regions with high attention scores. The updated feature will be used to predict both the action and object classes and regress the bounding box. To better improve the detection capability, we propose a new cascade detection pipeline and enable the use of multi-scale and high-resolution feature maps.

Our contributions are summarized below. (1) To the best of our knowledge, we are the first to study the HOI detection in the regime of large-vocabulary object categories. (2) To enable the study, we assemble an HOI dataset with 1000 object categories from SWiG [32]. Besides, we compose a subset from 100DOH dataset [36] and annotate them with ∼300 object categories for hand-object interaction. (3) We propose a new one-stage and end-to-end strategy by jointly detecting the target objects and interactions. This is achieved by using a repurposed Transformer unit and a new cascade detection framework over multi-scale feature maps. (4) Experimental results demonstrate that our method achieves a better result and faster inference speed than existing one-stage HOI detectors.

2. Related work

Previous HOI works [3, 10, 50, 49, 44, 30, 21] mainly performed with a limited variety of objects. Existing benchmarks [13, 3] focus on human interactions with 80 object categories. As they share the common category space as MS-COCO [26], existing approaches usually make use of generic object detectors [34, 24, 5] and their pre-trained weights to help with the instance detection, and focus on the subsequent interaction classification. As the standard object detection systems generate box candidates for almost all objects present in the image, it often introduces a lot of noises for the subsequent interaction prediction. Recent studies have gravitated towards solutions that can better discover the true interacting humans and objects among the detection results.

Gkioxari *et al.* [11] predict Gaussian heatmaps associated with humans to re-weight object candidates. Li *et al.* [22] propose to use a binary classifier to estimate the interactiveness of human-object pair and filter out low-confident ones before inference. Some recent works [31, 30, 50] utilize semantic features and word analogy to aid interaction predictions. As human body language often includes strong clues for the interaction, many recent studies [51, 42, 14, 6] model the human skeleton to improve the interaction prediction robustness. Another promising direction is to build graphs over all box candidates and then perform graph parsing to resolve the ambiguity caused by considering only single human-object pair [33, 9].

Different from the above two-stage pipeline, some recent works [46, 43] proposed one-stage solutions to jointly learn object detection and interaction classification. For example, PPDM [23] embeds CenterNet detector [5] to the whole framework and proposes an interaction branch to predict the interaction points and offsets to group the target instances. UnionDet [17] adopts RetineNet [25] to generate object boxes and predicts union regions in parallel to better produce interaction pairs and predictions. As the pure union regions of interactions may include unnecessary background information, DIRV [7] proposes a new one-stage approach to find the discriminative interaction regions. We observe that regardless of two-stage or one-stage methods, almost all previous methods use generic object detectors to find the target objects. In a large-vocabulary scenario, it is challenging to learn the object detectors to detect the diverse objects.

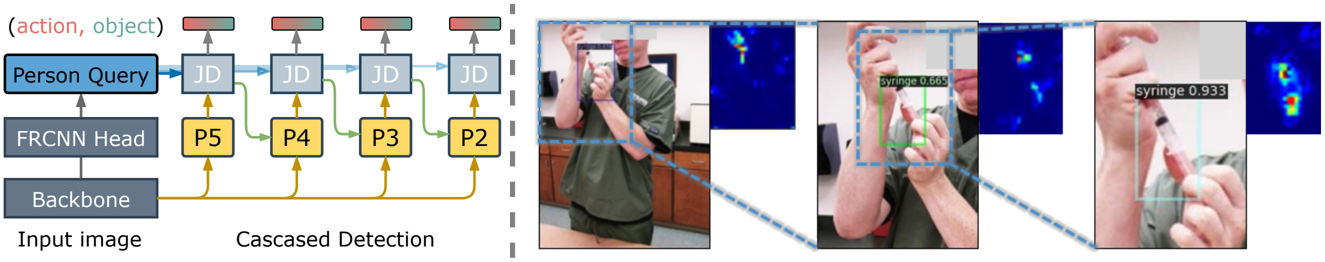

Figure 2: **Left**: The overall framework of our proposed method. We propose a cascaded detection to detect the HOIs over multi-scale feature maps (from low-resolution P5 to high-resolution P2). At each level, we conduct a joint detection (JD) for target objects and interactions based on the human query. **Right**: Illustration of the cascaded detection over multi-scale feature maps. For the small objects, we first estimate a rough position at the coarsest feature map. It then crops a sub-region on the higher-resolution feature map based on the initial guess and performs a second round of detection. In such a way, it progressively discovers the target object and refines the box.

3. Methodology

This work aims to address the problem of human-object interaction (HOI) detection with large vocabulary object categories. Given an image I, the goal is to generate multiple tuple (b_p, b_o, y_o, y_a) to represent HOIs, where the bounding box $b_p, b_o \in \mathbb{R}^4$ indicates the positions of the person and object, $y_a \in \mathcal{A} = \{1, \ldots, A\}$ represents the human actions, and $y_o \in \mathcal{O} = \{1, \ldots, C\}$ denotes the object categories. Different from existing HOI works, here we assume C is a large number (*e.g.*, $C = 1000$) instead of 80 MS-COCO object categories.

3.1. Proposed method

Existing HOI methods often decompose the detection into two stages [10, 20, 52, 9] or parallel branches [23, 7, 46, 17], where one stage (or branch) uses generic object detectors (*e.g.*, Faster RCNN [34], CenterNet [23], etc.) to generate human and object boxes, while the other stage (or branch) is for constructing human-object pairs and predicting interactions. However, when detecting diverse objects, data will be inevitably scarce for certain categories. In this case, the deep learning methods for generic object detection may perform poorly [12]. Different from the work of large-vocabulary object detection [12, 37, 18, 48, 45], we aim to develop an HOI-specific detector to find the target objects by leveraging on the interaction clues. Our aim is a unified model that can jointly find the objects and corresponding interactions.

To achieve this goal, we reformulate the HOI detection as a query problem. We employ a standard Faster RCNN to find person boxes $\{b_p\}$ and then use them as queries to search for the interactions within the image. The main goal is to learn a model $\mathcal{F}(I, b_p)$ that takes the image and person query as input, and outputs a set of interaction predictions $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$, where each prediction P_k is either one interaction performed by the input person query or an empty element $\emptyset$ representing the no-interaction case. Fig 2 shows the overview of our proposed method. We propose a cascaded framework to detect HOIs over multi-scale feature maps. At each feature level, we propose a novel HOI detector (JD) to jointly predict action and object categories and regresss the bounding box of target objects. We will elaborate on each module in the following sections.

3.1.1 Person Query Generation

Compared with detecting large-vocabulary objects, person detection would be a relatively simple task and most existing detectors can give a reliable result. We employ Faster RCNN to first generate person boxes $\{b_p\}$ and extract the RoI-Pooled visual feature $\mathbf{f}_p \in \mathbb{R}^d$ for each person box b_p. Many works [11, 40, 52] have shown that the spatial information between the person and objects often provide strong priors of the interaction. Hence, in addition to the visual feature, we also compute the positional embedding $\mathbf{f}_b \in \mathbb{R}^d$ for each box [41]. Concretely, the box $[x, y, w, h]$ is converted to a concatenation of four $d/4$-dimensional embeddings, $\mathbf{f}_b = [\mathbf{e}^x; \mathbf{e}^y; \mathbf{e}^w; \mathbf{e}^h]$, where $\mathbf{e}^t$ is a combined sine and cosine vector of box variable $t \in \{x, y, w, h\}$. Interested readers can find the details of the computation in the supplementary material and [41]. Let us denote $\mathbf{q}$ as the person query vector and $\mathcal{F}_q : \mathbb{R}^d \to \mathbb{R}^d$ as a projection function. We obtain the query vector by

$$\mathbf{q} = \mathcal{F}_q(\mathbf{f}_p + \mathbf{f}_b) \in \mathbb{R}^d \tag{1}$$

3.1.2 Joint Detection of Objects and Interactions

Given the person queries, the next step is to search for the interactions and corresponding target objects within the image. Let $\mathbf{F} \in \mathbb{R}^{h \times w \times d}$ denote the feature representation of the image which integrates both the visual feature from the backbone network and the positional embedding. Our core idea is to compare the person query $\mathbf{q}$ with $\mathbf{F}$ at each sliding window position to gauge the existence of objects which may interact with the given person query. Then, we aggregate the features from high-response regions to predict the interactions and regress the bounding boxes.

We effectuate this idea by using re-purposed Transformer units. As each person query is processed independently and in parallel, for a concise presentation, we only discuss one single query $\mathbf{q}$ in the following. We aim to find potential regions (reflected by attention maps) that may include the target objects. Consider that one person query can interact with more than one object at the same time. We propose to use separate attention maps to disambiguate different target objects. Specifically, we duplicate the person query K times (*e.g.*, $K = 10$), $\mathbf{q}^1, \ldots, \mathbf{q}^K$, and introduce a set of learnable offset vectors, $\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^K$, to pertube them, *i.e.*, $\mathbf{q}^k + \boldsymbol{\delta}^k$. We expect that the offset vectors can force the duplicate queries to point to different target objects, if exist. Let $\mathbf{A}^k \in \mathbb{R}^{h \times w}$ denote the attention map with respect to the k-th pertubed query. Its weight at position (x, y) can be calculated as

$$A^k_{xy} \propto \exp\{\frac{(\mathbf{q}^k + \boldsymbol{\delta}^k)^T \mathbf{U}^T \mathbf{V} \mathbf{F}_{xy}}{\sqrt{d}}\} \tag{2}$$

and normalized as $\sum A^k_{xy} = 1$, in which $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times d}$ are both learnable projection weights and $\mathbf{F}_{xy} \in \mathbb{R}^d$ is the feature at position (x, y). Ideally, we expect that the attention weight can locate regions (in most case, the discriminative parts) of the targets objects. We then aggregate the features based on the attention weights as

$$\mathbf{c}^k = \mathbf{W}_2[\sum_{x,y} A^k_{xy} \cdot \mathbf{W}_1 \mathbf{F}_{xy}] \tag{3}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d \times d}$ are learnable weights. We can understand $\mathbf{c}^k$ as a context feature integrating the information of target objects. We then use it to update the original query vector similar as the original Transformer. That is,

$$\begin{aligned} \mathbf{q}^k &\leftarrow \text{LayerNorm}\big(\mathbf{q}_k + \text{Dropout}(\mathbf{c}^k)\big) \\ \mathbf{q}^k &\leftarrow \text{LayerNorm}\big(\mathbf{q}_k + \text{Dropout}(\text{MLP}(\mathbf{q}^k))\big) \end{aligned} \tag{4}$$

The above computations are conducted several times to progressively locate the target objects and aggregate the context features. The final updated query feature will pass through three different Feed Forward Network (FFN) to predict human actions, regress the bounding boxes of target objects, and recognize the object categories, respectively.

3.1.3 Cascaded Multi-Scale Detection

Due to the computation complexities of Transformers, early works for object detection tasks [29, 2] usually use the high-level feature map with limited spatial resolution. It often attains a low performance on small objects. Recent works [53, 47] have proposed various variants to allow the use of high-resolution feature maps and achieved clear improvements. Inspired by this, we propose a cascaded detection framework for our HOI detection and also enable the use of multi-scale feature maps.

We employ the Feature Pyramid Network (FPN) backbone [24] and aim to search for interactions over P_2, P_3, P_4, and P_5 feature maps, which are $4\times$, $8\times$, $16\times$, $32\times$ smaller than the original image size, respectively. An intuitive way is to conduct separate joint detection (as introduced in Sec 3.1.2) at each feature level and then integrate their results. However, in this way, the computational complexity will increase with the feature resolution, as we compare the human query with each position in the map. Besides, we find that small objects still suffer from the poor bounding box regression even using the highest-resolution $P2$ feature map. To explore the behind reasons, we plot the maximal attention weight outside the ground truth boxes and the sum of attention weights within boxes (shown in Fig 3). We observe that, for small objects, the maximal attention out of the box is comparably low to large objects, while the sum of weights within the box is generally less than 0.5, which means that it aggregates a lot of noisy information. This is mainly due to the vast number of background positions and simply adopting a large feature map (*e.g.*, P_2) will not alleviate this issue.

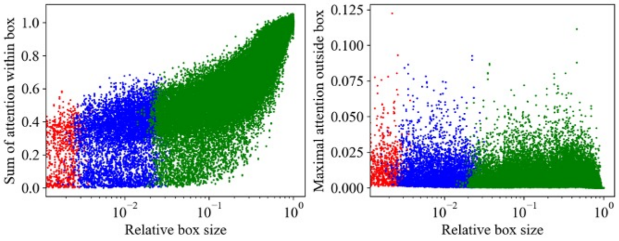

Figure 3: Left: Sum of attention weights within boxes. Right: Maximal attention weight out of boxes. Each point denotes one box prediction. Large, medium, and small objects are highlighted by green, blue and red colors, respectively. The bounding box regression for small objects generally suffers from more noises even though the attention module can precisely reveal their positions.

Hence, we propose a cascaded detection approach to progressively find the target objects. The assumption is that although the coarse feature map may not give a precise localization, it can roughly reveal its position within the image panel. Based on the initial estimation, we can crop a small region in a higher-resolution feature map and conduct a second round of prediction. Fig 2 shows an example of the proposed cascaded detection. We first perform a base detection on the coarsest P_5 feature map and obtain initial predictions for the target objects and corresponding interactions. Then, we conduct a second round of predictions on P_4 and so on.

Specifically, at P_i feature map, $i > 2$, if the relative area of the predicted box (compared with the input image size) is smaller than the pre-defined threshold τ_i^2 and both of the relative height and width are less than τ_i, we will continue to use the next higher-resolution feature map (*i.e.*, P_{i-1}) to refine the prediction. Instead of using the whole map, we crop a region of interest with a shape of (τ_i, τ_i) centered by the previously predicted box. We define the threshold τ_5, τ_4, τ_3 as $0.5, 0.25, 0.125$ respectively. We do this due to two considerations. First, we can use this way to control the computational complexity at high-resolution feature maps. Second, we can progressively eliminate the background regions and address the issue as shown in Fig 3.

3.1.4 Loss Function

In this subsection, we describe the loss functions of our proposed method. We use a standard Faster RCNN [34] to generate the person boxes. The first loss function $\mathcal{L}_{person}$ follows the conventional Faster RCNN including both the classification and regression loss of region proposal network and box head. For the joint object and interaction detector, we compute the losses at each feature level. At the i-th feature level, the loss function is similar to DETR [2] including the box regression loss $\mathcal{L}^i_{bbox}$, generalized IoU loss $\mathcal{L}^i_{giou}$ [35], and object classification loss $\mathcal{L}^i_o$, while we additionally consider the action classification loss $\mathcal{L}^i_a$. For each person query, the model will produce K different predictions. We assign them labels by finding the best bipartite matching between the predictions and ground truths. The final loss is calculated as

$$\mathcal{L}_{person} + \sum_{i=2}^{5} (\lambda_1 \mathcal{L}^i_{bbox} + \lambda_1 \mathcal{L}^i_{giou} + \mathcal{L}^i_o + \mathcal{L}^i_a) \qquad (5)$$

where $\lambda_1 = 5$ and $\lambda_2 = 2$.

4. Experiments

Datasets We assemble two datasets from SWiG [32] and DOH datasets [36] to study human interactions with large-vocabulary objects. SWiG is originally collected for the task of grounded situation recognition. It provides 504 visually grounded verbs, ~10k noun categories and the corresponding bounding box annotations. We extract the top 1,000 frequent object categories that can interact with humans and obtain 406 human actions. In our extracted SWiG-HOI subset, there are ~45k train images and ~14k test images. To further investigate the effectiveness of our proposed method, we also compose a dataset from 100DOH for detecting hand interactions with diverse daily objects. As the original DOH dataset only annotates the bounding boxes of the target objects without specific categories, we annotate ~30k training images and ~5k test images with 300 object categories.

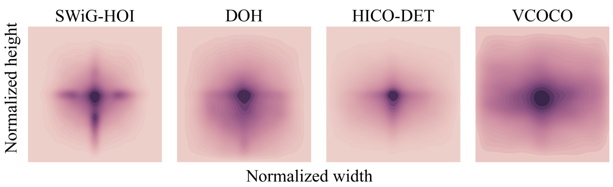

Figure 4: Distribution of object centers in normalized image coordinates for four datasets. VCOCO has the largest spatial diversity. Our composed SWiG-HOI and DOH have greater complexity than the commonly used HICO-DET dataset.

$n = 10$	AR	ARs	ARm	ARl	ARr	ARc	ARf
RFS	38.16	2.70	7.74	41.99	27.20	26.18	41.48
EQL	37.09	2.72	6.89	40.88	25.16	25.60	40.32
Ours	57.96	6.15	21.01	62.81	40.99	41.16	62.46
$n = 100$	AR	ARs	ARm	ARl	ARr	ARc	ARf
Ref.*	32.69	22.25	45.89	55.32	33.49	30.41	33.16
RFS	54.64	13.33	31.51	57.86	46.85	47.62	56.61
EQL	54.70	15.23	34.78	57.55	47.47	48.16	56.56
Ours	67.84	16.79	38.77	70.50	50.35	52.89	69.02

Table 1: Average recall of the top n class agnostic boxes on extracted SWiG-HOI. ARs, ARm and ARl represent the average recall of the small, medium and large objects. ARr, ARc and ARf represent the average recall of the rare, common and frequent objects. We also report a reference result (Ref.*) of the Faster RCNN with RFS on LVIS [12] v0.5 val set.

One important aspect to discuss is whether the objects are distributed throughout the image planes instead of being salient in the middle of the images. For this aim, Figure 4 shows the object-center density of the newly composed datasets and compares them against the HICO-DET [4] and VCOCO [13] benchmarks. It shows that all HOI datasets have center bias while the newly constructed datasets have greater complexity than HICO-DET. More details of the datasets are available in the supplementary material.

Implementation Details Our backbone is ResNet50 [15] with Feature Pyramid Network (FPN) [24], which is initialized using the pre-trained COCO detection weights. The joint detection model uses a stack of 4 Transformer decoder units (without the cross-attention computations among input queries) at each feature level and set the feature dimension to 256. We adopt AdamW [28] to train the model with 40 epochs and set the initial learning rate as 0.0001 and a weight decay of 0.0001.

4.1. Target Object Detection

HOI detection involves the localization and classification of humans and interacting objects. To better understand the challenges of the task with large vocabulary objects, we start by discussing the performance of object detection. We use the COCO-style Average Recall (AR) to reflect the localization quality and Average Precision (AP) for the object detection performance. As most previous two-stage HOI methods rely on generic detectors (*e.g.*, Faster RCNN) to find the target objects, in the following we mainly explore the performance of Faster RCNN with large-vocabulary techniques [37, 18, 45, 39] as well as the state-of-the-art one-stage HOI detectors.

	E2E	Action	AP	APr	APc	APf
FRCNN+RFS [12]	✓		6.12	2.45	5.61	10.87
EQL [37]	✓		5.74	1.51	5.21	10.95
BAGS [18]			6.62	2.47	6.05	11.98
SimCal [45]			5.61	1.27	5.08	10.89
De-confound [39]			5.96	1.62	5.67	10.21
JSL [32]	✓	✓	6.81	3.26	6.42	10.80
Ours baseline	✓		5.70	1.39	5.12	11.15
Ours	✓	✓	**7.31**	**3.83**	**6.75**	**12.14**

Table 2: Average precision of large-vocabulary object detection methods on SWiG-HOI. We highlight the end-to-end (E2E) approaches and methods using the action clues to assist the object detection.

Target object coverage Among large-vocabulary object detectors [37, 18, 45, 32, 39] with Faster RCNN or Mask RCNN implementation, one common assumption is that the class-agnostic proposals usually have a reliable recall on the target objects. In Table 1, we explore if this assumption still holds for the HOI task on SWiG and report the results of the proposals generated by Faster RCNN with repeated factor sampler (RFS) [12] and equalization loss [37] (EQL). To give a better sense, we also report the AR of RFS on the LVIS v0.5 val set as the reference. Consider that there are often a few target objects (compared with object detection tasks) interacting with the humans in the image. We limit the number of proposals and expect the top confident ones can well capture the targets.

Compared with the reference result in LVIS, we observe that class-agnostic boxes produced on SWiG generally give a higher AR except for the small (ARs) and medium (ARm) objects. The higher result is mainly due to that there are fewer target objects (only those interacting with humans) required to be detected than LVIS. We hypothesize that the low ARs and ARm are due to that small/medium objects often suffer from occlusions when they interacting with humans, making it more challenging to find them.

Table 1 also shows that our proposed method can better find the target objects. Especially at the top 10 boxes, our method achieves ∼20 AR improvements than the object detection counterpart. We believe that the gap is mainly because generic detectors produce many non-interacting objects present in the image as they are incapable of differentiating the meaning of interaction. In contrast, our HOI-specific detection method can find the true target by comparing them with the human queries; thus, it is less likely to be influenced by other objects in the background.

Large-vocabulary detection In addition to the box localization, another challenging problem is to classify the large-vocabulary categories. Some pioneer works [37, 18, 45, 39] have been proposed to address the long-tailed instance segmentation and object detection problem. In Table 2, we explore their results on the composed SWiG-HOI. Besides, we compare against the Joint Situation Localizer (JSL) from SWiG [32] which modifies the RetineNet to recurrently find targets based on the action prediction and previously detected objects. For a fair comparison, all methods use the ResNet-50 backbone and limit the number of class-specific detections per image to 100 with the minimum score threshold of 0.001.

n=10	AR	ARs	ARm	ARl	ARr	ARc	ARf
PPDM	57.89	6.17	20.00	62.82	40.76	40.41	62.40
DIRV	55.65	4.74	17.57	60.58	38.45	37.70	60.26
Ours	**57.96**	**6.15**	**21.01**	**62.81**	**40.99**	**41.16**	**62.46**
n=100	AP	APs	APm	APl	APr	APc	APf
PPDM	3.47	1.90	2.31	3.74	0.49	2.64	9.38
DIRV	3.08	1.52	1.94	3.34	0.46	2.35	8.30
Ours	**7.31**	**2.52**	**4.14**	**7.62**	**3.83**	**6.75**	**12.14**

Table 3: Comparison with state-of-the-art one-stage HOI detectors. We report the average recall at the top 10 boxes and the average precision of the top 100 predictions on SWiG-HOI.

Table 2 shows that the task-specific detectors (JSL and ours) generally achieve a better result than generic object detectors. For the HOI task, objects present in the image are usually sparsely annotated as only interacted objects are treated as foreground, while other objects will become background. This does not match with the objective of generic object detectors, and it is difficult to train generic detectors, resulting in worse results as shown in Table 2. Besides, we observe that coupling the action prediction can assist the object classification. One evidence is that JSL, without special designs for tail categories, achieves an AP of 6.81 and beats all generic detectors. To delineate this point, we report a baseline model that ablates the action prediction from the framework and this leads to a 1.61 AP drop. We hypothesize that objects and actions often have obvious dependency relationships. We opine that, in addition to the generic large-vocabulary techniques, it would also be possible to leverage on unique characters of HOIs (*i.e.*, the compositional relationship, the dependency between objects and actions) to handle the large-vocabulary object categories.

End-to-end HOI detectors In addition to ours, some prior works have also proposed end-to-end solutions to detect interacted objects rather than using off-the-shelf object detectors. As they are originally developed under the regime of limited object categories (*e.g.*, 80 MS-COCO categories), we test their performance on the large vocabulary case. Table 3 reports the results of two state-of-the-art end-to-end HOI detectors, PPDM [23] and DIRV [7], whose codes have been made available. Their common idea is to predict interaction points or regions, working with the embedded object detector, to better detect interacted humans and objects. From Table 3, we can see that they can achieve a good recall on capturing the target boxes, while the performance for AP is leaves much to be desired. The object scores are produced alone by the embedded object detector

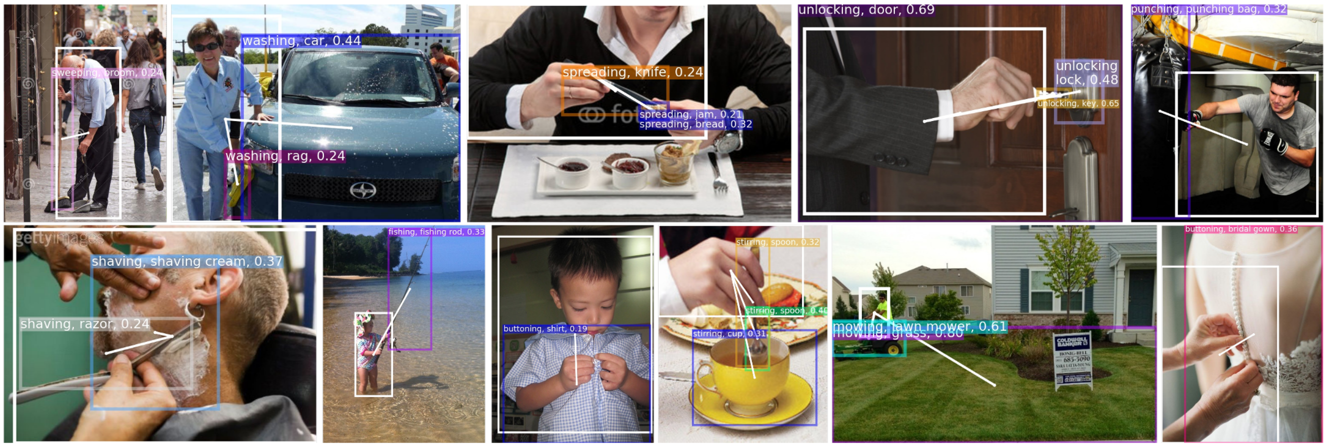

Figure 5: Examples of human interactions with diverse objects detected by our method.

	mAP	mAP-r	mAP-nr	mRec	mRec-r	mRec-nr
PPDM	3.17	1.62	6.53	14.17	7.77	28.13
DIRV	2.83	1.46	5.82	12.50	6.90	24.69
JSR	7.33	6.10	10.01	18.20	14.32	26.67
Ours	**7.98**	**6.63**	**10.93**	**20.17**	**16.03**	**29.21**

Table 4: Experimental results of HOI detection. We report the average precision and recall of rare interactions (mAP-r, mRec-r) and non-rare interactions (mAP-nr, mRec-nr).

	mAP	mAP-r	mAP-nr	mRec	mRec-r	mRec-nr
PPDM	7.67	5.19	13.08	56.24	50.71	68.30
DIRV	7.49	5.03	12.87	52.60	46.34	66.26
JSR	19.46	15.32	28.48	50.14	40.85	70.43
Ours	**20.96**	**16.48**	**29.88**	**59.64**	**48.69**	**83.54**

Table 5: Results of HOI recognition without box grounding requirements.

which is isolated with their interaction branches. Specifically, PPDM uses CenterNet [5] and DIRV uses EfficientDet [38] to produce the object scores. To better handle the rare and common categories, additional techniques for addressing the low-shot samples are required to work with their detectors. Besides, during the re-training, we notice that DIRV conducts a 1000-way classification for all anchors in the feature map, resulting in huge memory costs and computations.

4.2. HOI Detection

In this section, we discuss the full interaction detection with large-vocabulary object categories. We follow the existing HOI benchmark [3] and use the mean average precision (mAP) and mean recall (mRec) evaluation metrics. Specifically, we first calculate the AP and Rec per interaction category and then report the mean. An interaction detection is considered as positive only if the following conditions are satisfied: (1) the person and object bounding boxes have an IoU ≥ 0.5 with the ground truth; (2) the interaction prediction is correct, including both the correct action and object prediction. Following [3], we treat interaction as a rare case if it has at least one but less than 10 training samples. Next, we discuss the result of the interaction detection.

HOI detection with box grounding. We test two state-of-the-art one-stage HOI detectors which are re-trained in SWiG-HOI and the well-trained model JSR from the original SWiG work [32]. Due to the compositional nature of HOIs, the number of possible interactions will grow quickly as more object categories are involved. In our composed test set, there is a total of 4,745 seen human-object interactions. Among them, 1,491 belongs to the non-rare case and 3,254 are rare interactions. Table 4 reports the result of HOI detection using the standard HOI evaluate metrics. Our proposed method outperforms state-of-the-art one-stage HOI detectors by a clear mAP margin ($\sim$4.81 mAP). Among all methods, we also observe that our method can achieve a relatively higher recall on the interactions.

HOI recognition without box grounding. Here we relax the requirements of localization and only consider the correctness of interaction prediction. In this case, a prediction will be treated as true positive if the interaction category is correctly predicted. Table 5 reports the evaluation results without box requirements. As shown, our method still gives the best performance on the interaction prediction. We also observe that JSR misses 1,635 out of 3,254 (50.25%) rare interactions. In comparison, our method misses 1,507 (46.31%) rare interactions. Although our result is slightly better than the baseline, it is still an initial result and more advanced solutions along this direction can be expected in the future.

Unseen interactions between actions and objects. The above evaluation is mainly conducted on the seen interactions. As more object categories are included, there will be more chances to see the novel interactions. In the composed test set, there are about $\sim$1.8k novel interactions that

	mAP-novel	mRec-novel
PPDM	0.78	2.73
DIRV	0.75	2.62
JSR	2.34	4.30
Ours	**2.64**	**8.55**

	Inference Speed (ms)
PPDM	237
DIRV	214
JSR	353
Ours	**93**

Table 6: Left: Experimental results on novel interactions. Right: Inference speed of one-stage HOI detectors.

do not appear in the training set. Table 6 (left) reports the results (with box requirements) on these novel combinations between the actions and objects. Compared with previous methods, our proposed method can capture more novel interactions. While the results are expectedly low in both mAP and mRec, this presents another promising research direction to study how to handle the novel combinations in the regime of large-vocabulary object categories.

Inference speed Table 6 (right) shows the inference speed of various methods. Existing one-stage HOI detectors often include a matching process to associate the generated person and object instances for each predicted action class. Their computation complexity will increase as more action and object classes are required to be considered. In comparison, our model jointly detects the target objects and interactions without extra matching processes. Due to this advantage, it can achieve a faster speed than the baseline methods.

4.3. Hand-Object Interaction

In addition, we investigate the effectiveness of our proposed model on another closely related task - hand and object interaction. We conduct the experiments on the 100DOH dataset as it provides human hand interactions with diverse daily objects (including ∼300 categories based on our annotations). To keep it consistent with the above studies and our formulation, we treat each hand as a query and search for the target objects. In this experiment, the action is defined as the contact state (*e.g.*, self-contact/person-to-person/contact-to-portable/non-portable). More details about this dataset are available in the supplementary materia.

Table 7 reports the experimental result of both object detection and interaction detection. Similar to above, we use the COCO-style AP to evaluate the object detection and mAP metric to reflect the performance of interaction prediction. We compare against the DOH [36] baseline method and one-stage HOI detectors, PPDM and DIRV. Although in this experiment, the simple contact state definition may not provide additional clues to ease the object recognition, we still see that our proposed method can effectively detect target objects. As shown, our method achieves 2.5 AP improvement on detecting the interacted objects. For the full interaction detection, our method can obtain a 1.9 mAP boost compared to selected one-stage HOI baselines.

	AP	APr	APc	APf	mAP	mAP-r	mAP-nr
DOH	20.8	10.3	18.9	27.8	22.6	12.1	25.6
PPDM	25.5	13.4	23.5	30.8	24.5	15.3	27.2
DIRV	25.1	14.2	22.1	31.5	24.0	15.2	26.6
Ours	**28.0**	**15.8**	**24.3**	**32.7**	**26.4**	**16.5**	**29.2**

Table 7: Experimental results on the composed DOH dataset. We evaluate the object detection performance using AP, APr(are), APc(ommon), APf(reqeunt). The interaction detection is evaluated using mAP metric similar as existing benchmarks.

	multi-scale	cascade	action	AP	APs	APm	APf
Baseline 1				5.16	1.24	3.03	5.78
Baseline 2	✓			5.28	2.11	3.48	5.87
Baseline 3	✓	✓		5.70	2.34	3.92	6.29
Full model	✓	✓	✓	7.31	2.52	4.14	7.62

Table 8: Ablation studies of the proposed method. We ablate the multi-scale feature maps, cascaded detection framework, and coupled object detection and action prediction.

4.4. Ablation Studies

In this section, we conduct some ablation studies for our proposed method. Table 8 reports the result of our baseline models. We first investigate the effectiveness of using multi-scale and high-resolution feature maps. The first baseline model can simply be interpreted as a variant of the vanilla DETR [2] model (*baseline 1*). It only uses the coarse P5 feature maps to find the target objects. As shown, such a basic model has a low result on APs and APm. Then we incorporate multi-scale and higher-resolution feature maps. An intuitive way is to parallelly detect objects at P2, P3, P4, P5 feature maps from the FPN backbone (*baseline 2*) and then merge the results. We see that such a simple modification can bring about 0.87 AP improvement in APs. We then introduce our proposed cascade detection framework (*baseline 3*) which further boosts the object detection performance. The biggest challenge is the box classification in the large vocabulary case. When we couple the object classification with the action classification, we see another obvious boost for AP performance.

5. Conclusion

In this paper, we propose a novel model to address the problem of HOI detection in a regime of large-vocabulary object categories. It jointly discovers the target objects and interactions with human queries in a cascaded framework over multi-scale feature maps. It does not rely on any front-stage object detectors and can be end-to-end trained. We assemble two datasets from SWiG and DOH datasets to study the new challenges of HOI detection in the large-vocabulary setting and investigate the effectiveness of our method. We observe that the coupled object detection and interaction prediction not only helps detect the target objects but also delivers a notable improvement on interaction prediction. These contributions allow us to detect human interactions with diverse objects in our daily life.

References

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *The European Conference on Computer Vision (ECCV)*, 2020. 4, 5, 8

[3] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 2, 7

[4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 5

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 7

[6] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *The European Conference on Computer Vision (ECCV)*, 2018. 2

[7] Hao-Shu Fang, Yichen Xie, Dian Shao, and Cewu Lu. Dirv: Dense interaction region voting for end-to-end human-object interaction detection. In *The AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 1, 2, 3, 6

[8] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4991–5000, 2018. 1

[9] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, 2020. 1, 2, 3

[10] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference*, 2018. 2, 3

[11] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 2, 3

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3, 5, 6

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 5

[14] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[16] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 1

[17] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, 2020. 1, 2, 3

[18] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 5, 6

[19] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[20] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *Advances in Neural Information Processing Systems*, 2020. 1, 3

[21] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[22] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 2

[23] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3, 6

[24] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 4, 5

[25] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1, 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2

[27] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 1

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[29] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. 4

[30] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[31] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *The IEEE*

International Conference on Computer Vision (ICCV), 2017. 2

[32] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020. 2, 5, 6, 7

[33] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 3, 5

[35] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 5

[36] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 1, 2, 5, 8

[37] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 5, 6

[38] Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, 2020. 7

[39] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 2, 5, 6

[40] Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2, 3

[42] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[43] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with novel objects via zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2

[44] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[45] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, 2020. 2, 3, 5, 6

[46] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 3

[47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 4

[48] Jialian Wu, Liangchen Song, Tiancai Wang, Qian Zhang, and Junsong Yuan. Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1570–1578, 2020. 2, 3

[49] Tete Xiao, Quanfu Fan, Dan Gutfreund, Mathew Monfort, Aude Oliva, and Bolei Zhou. Reasoning about human-object interactions through dual attention networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[50] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 2

[51] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[52] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Haibin Ling, and Jianbing Shen. Cascaded human-object interaction recognition. In *CVPR*, 2020. 1, 3

[53] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4