Discovering Human Interactions with Novel Objects via Zero-Shot Learning

Suchen Wang, Kim-Hui Yap, Junsong Yuan, Yap-Peng Tan

1Nanyang Technological University  2State University of New York at Buffalo

{wang.sc, ekhyap, eyptan}@ntu.edu.sg, {jsyuan}@buffalo.edu

Abstract

We aim to detect human interactions with novel objects through zero-shot learning. Different from previous works, we allow unseen object categories by using its semantic word embedding. To do so, we design a human-object region proposal network specifically for the human-object interaction detection task. The core idea is to leverage human visual clues to localize objects which are interacting with humans. We show that our proposed model can outperform existing methods on detecting interacting objects, and generalize well to novel objects. To recognize objects from unseen categories, we devise a zero-shot classification module upon the classifier of seen categories. It utilizes the classifier logits for seen categories to estimate a vector in the semantic space, and then performs nearest search to find the closest unseen category. We validate our method on V-COCO and HICO-DET datasets, and obtain superior results on detecting human interactions with both seen and unseen objects.

1. Introduction

Human-object interaction (HOI) detection [12, 5, 11, 27] is important for human-centric visual understanding. The goal is to detect interactions between humans and objects, and use verbs to describe their relationships (e.g., sit on bench, carry suitcase, etc.). Although recent studies [4, 21, 44, 48, 43, 40, 39] have achieved good progress, current HOI methods are limited to interactions with 80 object categories as defined in MS-COCO dataset [23].

Previous attempts [17, 37, 1, 30] to scale HOIs only focus on detecting human interactions with known objects. It aims to generalize the knowledge obtained from seen interactions (e.g., sit on chair, carry suitcase) to unseen interactions (e.g., sit on suitcase). In comparison, as shown in Figure 1, we aim to detect human interactions with unseen object categories (i.e., categories without any annotated visual samples in the training set).

Most existing HOI methods [4, 21, 44, 48, 43, 40, 39] applied an off-the-shelf object detector to first generate human and object candidates, and then applied an interaction model to predict their relationships. However, the commonly used object detectors [35, 22, 7] are designed for detecting all objects in the given image. As a result, many non-interacting human-object pairs are produced. Besides, those object detectors treat humans as an independent category like other object categories. In this way, the detection of objects cannot exploit the information of human appearance.

To alleviate the above limitations, we propose a detector (as shown in Figure 2) specifically for the HOI detection task. Our main idea is to leverage human visual clues to find interacting objects. The proposed detector follows the pipeline of Faster RCNN [35], while we replace the original region proposal network (RPN) with a novel human-object region proposal network (HO-RPN). It scores region proposals based on its interactiveness with the detected humans, and generalizes well to novel objects if they are interacting with humans (e.g., the red box in Figure 1). This enables us to detect human interactions with unseen object categories.
Our model consists of three main modules: (1) Our proposed human-object region proposal network (HO-RPN) first localizes humans and interacting objects. It estimates an interactiveness score for each object proposal. (2) The zero-shot classification module classifies the generated region proposals. It can recognize both seen and unseen object categories. (3) It then estimates the probability of verbs between every human-object pair. The score from three modules will be aggregated as the final score for the interaction.

To recognize novel objects, we assume each object category has a corresponding semantic embedding vector which is learned from text corpus [29, 26]. We build a zero-shot classification module upon the softmax classifier of seen categories. Based on the logit outputs of the classifier, we estimate a semantic embedding vector for the input region proposal and then search for the nearest unseen category in the semantic space.

Extensive experiments on V-COCO [12] and HICO-DET [4] datasets show that: (1) The proposed HO-RPN can learn to localize objects based on their interactiveness with humans; (2) By leveraging on human visual clues, our model significantly outperforms zero-shot object detection [2] on detecting novel objects; (3) Our model can outperform existing HOI methods on detecting human interactions with both seen and unseen objects. Besides, we construct a test set from Visual Genome [18] with 110 object categories (80 MS-COCO + 30 new) to show the ability of our model on detecting human-novel-object interactions.

Our contributes are summarized below. (1) This is the first attempt to detect human interactions with novel objects. (2) We propose a novel human-object region proposal network for the HOI detection task. (3) We design a zero-shot classification module to recognize novel objects. (4) Our approach\footnote{https://github.com/scwangdyd/zero_shot_hoi} achieves superior results of HOI detection on V-COCO [12] and HICO-DET [2] datasets.

2. Related Work

Human-Object Interaction (HOI) The goal of HOI is to detect interactions between humans and objects. It is closely related to visual relationship detection [46, 6, 25, 31, 45], while HOI detection requires more fine-grained verbs to describe the relationships. Existing HOI works [4, 10, 44, 43, 40, 30] usually take advantage of pre-trained object detectors [35, 22], and focus their attention on improving the verb predictions. In their frameworks, the verb prediction model needs to differentiate noninteracting human-object pairs since noninteracting objects are also detected by object detectors. Gkioxari et al. [11] proposed to use human appearance to predict the potential location of interacting objects and then re-weight object candidates based on their distances to the prediction. Li et al. [21] learned a binary classifier to estimate the interactiveness of human-object pairs and filter out noninteracting ones. Qi et al. [32] proposed to build a graph network among all human and object candidates and then parse their relationships. Human body language often includes strong clues for the interaction. Many recent works [48, 39, 13, 8] leveraged on human pose to improve the robustness of verb predictions such that they can reduce the false positive predictions on noninteracting human-object pairs.

Instead of improving the verb prediction, our main goal is to detect more interactions by extending novel objects. Shen et al. [37] have scaled HOIs to 600 classes by increasing the number of verbs. However, current HOIs are still limited to the 80 MS-COCO object categories [23].

Most ZSL works focus on the zero-shot image recognition problem [41, 47, 42, 49], which aims to recognize unseen classes by generalizing the knowledge learned from seen classes. For ZSL, additional side information is required. Early works [9, 19, 24] utilize attributes to link various classes. This requires much human labor to design attributes, especially for large-scale datasets. An easier way is to use the semantic word embedding of category names [3, 28, 38]. Based on the distance in the semantic space, we can implicitly gauge the relationships among classes.

Apart from zero-shot image recognition, zero-shot object detection [33, 2, 34] has received much attention recently. It aims to detect unseen objects from the given image using annotations of seen object categories. Existing methods [33, 2, 34] usually choose semantic embedding vectors as side information. This task is different from our focus in this paper since it aims to detect all novel objects in the images, while we focus on the novel objects in the interaction with humans.
3. Approach

3.1. Problem Statement

The goal of human-object interaction (HOI) detection is to find one or multiple tuples \(<\text{human, verb, object}>\) from the given image. Formally, human-object interaction can be defined as \(<b_h, v, b_o, y>\), where the bounding box \(b_h, b_o \in \mathbb{R}^d\) indicates the location of humans and objects, verb \(v \in \mathcal{V} = \{V_1, \ldots, V_m\}\) denotes the action performed by the human, and \(y \in \mathcal{Y}\) denotes the object category.

In this paper, our main focus is to scale HOIs by extending the object category space via zero-shot learning. Let \(\mathcal{Y}_S = \{1, \ldots, c_1\}\) and \(\mathcal{Y}_U = \{c_1 + 1, \ldots, c_1 + c_2\}\) denote the seen and unseen object category space respectively. We aim to develop a model which can detect human interactions \(<b_h, v, b_o, y>\) with both seen and unseen object categories, i.e., \(y \in \mathcal{Y} = \mathcal{Y}_S \cup \mathcal{Y}_U\). To make this a feasible task, we assume that each object category \(y \in \mathcal{Y}\) has a semantic embedding vector \(q_y \in \mathbb{R}^p\) such that we can leverage on the relations in the semantic space to detect human interactions with unseen objects.

3.2. Formulation

Existing HOI detection methods [12, 11, 44, 13] consist of two main components, i.e., an off-the-shelf object detector [35, 7, 20, 22] followed by an interaction model. Given an input image \(x\), the object detector first detects \(b_h\) and \(b_o\), and predicts box score \(p(b_h, y|x)\) and \(p(b_o, y|x)\). Without the loss of clarity, we replace \(p(b_h, y = \text{“person”}|x)\) with \(p(b_h|x)\) in the following discussion for concise representation. Given \(b_h\) and \(b_o\), the interaction model then estimates the probability of verbs denoted by \(p(v|b_h, b_o, y, x)\). The final score for the interaction \(<b_h, v, b_o, y>\) is given as

\[
p(b_h, v, b_o, y|x) = p(v|b_h, b_o, y)p(b_o, y|x)p(b_h|x) \tag{1}
\]

In this way, the original problem is reduced to 3 tasks, i.e., human detection, object detection, and verb prediction. Notice that the framework in Eq.(1) treats the detection of humans and objects as two independent processes, i.e., \(p(b_h, b_o, y|x) = p(b_h, y|x)p(b_o|x)\). This assumption leads to the following limitations.

First, isolating the detection of objects with humans makes it impossible to detect only interacting objects. In this framework, the object detection will detect both interacting and noninteracting objects as candidates. Previous methods [11, 32, 21] often perform post-processing to suppress noninteracting ones. In comparison, we propose a more efficient way, i.e., producing only interacting objects at the detection stage. Second, the framework in Eq.(1) cannot handle novel objects well, even with zero-shot object detection [2, 34], since it is difficult to differentiate unseen objects with the background using only its visual feature.

However, the human body language usually implies the location of interacting objects. For instance, as shown in Figure 1, we can rely on human visual clues to localize the unseen novel object “rose”.

To alleviate the above limitations, we propose to score the object box based on its interactiveness with the detected human boxes, i.e., \(p(b_o, y|b_h, x)\). Then the score of interaction \(p(b_h, v, b_o, y|x)\) can be expressed as

\[
p(v|b_h, b_o, y, x)p(b_o, y|b_h, x)p(b_h|x) \tag{2}
\]

Our model is shown in Figure 2. It is composed of a novel human-object region proposal network, a zero-shot classification module, and a verb prediction module. We will elaborate them in Sec.3.3, Sec.3.4 and Sec.3.5 respectively.

3.3. Human-Object Region Proposal Network

We have two separate branches for generating human and object region proposals. We use the original region proposal network (RPN) [35] to generate human region proposals. For object region proposals, we design a novel human-object region proposal network (HO-RPN).

The original RPN estimates the score of anchor boxes based on the objectness. Intuitively, a high score will be assigned to an anchor box if it well covers an object. However, this criterion does not well match the goal of HOI detection, since we are only interested in interacting objects. In our case, we expect that a high score is assigned if the anchor box well covers an object and, more importantly, the object is interacting with humans. To achieve this goal, we propose a human-object region proposal network (HO-RPN). It scores the anchor boxes based on its interactiveness with humans.

The architecture of HO-RPN is shown in Figure 3. In addition to the convolutional feature maps from the backbone network, HO-RPN also takes as input the feature of the top \(K\) detected human anchor boxes from the RPN hidden layer. HO-RPN first performs 2D convolution on the backbone feature maps to obtain a hidden feature.
map. Then, at each sliding window position, we apply a relational network (RN) [36] to reason about the interactivity of $n$ anchor boxes with different shape, $s_r = [s_r^{(1)}, \ldots, s_r^{(n)}] \in \mathbb{R}^n$, based on its visual features and relationships with detected humans. Let $x_o^{(j)} \in \mathbb{R}^d$ denote the feature of $j$-th sliding window position in HO-RPN, and $x_h^{(k)} \in \mathbb{R}^d$ denote the feature of $k$-th human anchor box from human RPN. Specifically, the relational network computes the score by

$$s_r = \sigma \left( f \left( \sum_{k=1}^{K} g(x_o^{(j)}, x_h^{(k)}) \right) \right)$$  \hspace{1cm} (3)$$

where $\sigma(\cdot)$ is a sigmoid function. Here $g(\cdot) : \mathbb{R}^{2d} \rightarrow \mathbb{R}^d$ is a simple multi-layer perceptron (MLP), which processes the concatenated feature of $x_o^{(j)}$ and $x_h^{(k)}$. The role of $g(\cdot)$ is to infer if the object and human are interacting. Here $f(\cdot) : \mathbb{R}^{d} \rightarrow \mathbb{R}^n$ is another MLP, which aggregates the relations with $K$ human boxes. This predicted score will be used to generate object region proposals.

### 3.4. Zero-Shot Object Classification

Once human and object region proposals are generated by HO-RPN, a head network is used to predict $p(b_h | x)$ and $p(b_o, y | x)$, and regress the box $b_h$ and $b_o$. To be compatible with novel object categories, here we use a class agnostic bounding box regressor instead of class-specific regressors. To recognize unseen categories, we follow the idea from ConSE [28] and design a zero-shot classification module for our task. The objective behind our module is that it should not change the architecture of detectors such that we can take advantage of the well-trained weights on seen categories. The benefit of doing so is that if we can add new object categories into the object space, we do not need to re-train the network.

Our zero-shot classification module is shown in Figure 4. It uses the logit output of the softmax classifier for seen categories to estimate the probability of unseen categories. Suppose $f_y(b_o) \in \mathbb{R}$ is the probability of region proposal $b_o$ belonging to seen category $y \in \mathcal{Y}_S$ predicted by a classifier. Given $\{f_y(b_o)\}_{y \in \mathcal{Y}_S}$, we denote the most likely seen category as

$$\hat{y}_t := \arg \max_{y \in \mathcal{Y}_S} f_y(b_o)$$  \hspace{1cm} (4)$$

Similarly, let $\hat{y}_j$ denote the $j$-th most likely seen category, i.e., the category with the $j$-th largest value among $\{f_y(b_o)\}_{y \in \mathcal{Y}_S}$. Based on the top $K$ predicted seen categories and their semantic word embedding $\{q_{\hat{y}_1}, \ldots, q_{\hat{y}_K}\}$, we estimate a semantic embedding vector $e \in \mathbb{R}^p$ for the input region proposal $b_o$ by

$$e = \frac{1}{Z} \sum_{j=1}^{K} f_{\hat{y}_j}(b_o) \cdot q_{\hat{y}_j}$$  \hspace{1cm} (5)$$

where $Z = \sum_{j=1}^{K} f_{\hat{y}_j}(b_o)$, which is a normalization factor.

Besides the categories in $\mathcal{Y}_S$, the background is also a category in the softmax classifier. If the softmax classifier is very confident in its prediction for $y = \text{"background"}$, it will not be a seen object. Then, we estimate if it belongs to unseen categories. To do so, we compute the similarity of $e$ to the semantic embedding vector of unseen categories as

$$s_y = \cos(e, q_y), \quad y \in \mathcal{Y}_U$$  \hspace{1cm} (6)$$

where $\cos(\cdot, \cdot)$ is the cosine similarity. If $s_y < \tau$ for any $y \in \mathcal{Y}_U$, the region proposal $b_o$ will be predicted as background.

For seen categories, we take the output of the softmax classifier as the prediction, i.e., $s_y = f_y(b_o), y \in \mathcal{Y}_S$. By doing so, our proposed zero-shot classification module enables the detector to detect unseen categories without the loss of performance on seen categories. Finally, the score of object boxes is obtained as

$$p(b_o, y | b_h, x) := s_r \times s_y$$  \hspace{1cm} (7)$$

### 3.5. Verb Prediction

Once humans and objects are detected, verb branch will predict the probability of verbs for every human-object pair. Except for the visual appearance within box $b_o$ and $b_h$, we also take as input the union region of $b_o$ and $b_h$, since there may be additional context information. Consider that each human can simultaneously perform multiple actions for an object, e.g., looking at and holding, we formulate the verb prediction as multiple binary classification problem. For a verb category $v \in \mathcal{V}$, we predict its probability by

$$p(v | b_h, b_o, y, x) := \sigma \left( h_v(x_o, x_h, x_{h,o}) \right)$$  \hspace{1cm} (8)$$

where $x_o, x_h, x_{h,o} \in \mathbb{R}^d$ are the visual features extracted from box $b_o, b_h$, and their union region using RoIAlign [14]. Here, $h_v(\cdot, \cdot, \cdot) : \mathbb{R}^{3d} \rightarrow \mathbb{R}$ is a MLP which processes the concatenated feature and $\sigma(\cdot)$ is a sigmoid function.

Figure 4: The pipeline of zero-shot classification. The softmax classifier for seen categories (including the background) first classifies the region proposal. If the background category has a high response, we then identify if it belongs to unseen categories. We estimate a semantic embedding using Eq. (5). The nearest unseen category in the semantic space will be chosen as the prediction.
4. Experiments

We mainly evaluate our method using V-COCO [12] and HICO-DET [4] datasets, which have exhaustive annotations of human interactions with 80 MS-COCO categories [23]. There are 25 and 117 verbs on V-COCO and HICO-DET respectively to describe human-object interactions. We also construct a test set from Visual Genome [18] with 30 novel categories apart from the 80 MS-COCO categories to evaluate our method.

4.1. Evaluation Protocol

**Seen/unseen split** We simulate the zero-shot scenario on V-COCO [12] and HICO-DET [4] by partitioning the 80 object categories into seen and unseen sets. All images including unseen object categories will be removed from the training set. Existing seen/unseen splits [2, 33, 34] are mainly constructed based on the statistics of COCO, which cannot be applied to the above HOI datasets. For example, in previous work [33], “vase” is a seen class with 4,623 instances in the COCO train2014 set, while no HOI with “vase” exists in the V-COCO train set. For this reason, we propose a new seen/unseen split following the steps in [33]. We sort the classes per supercategory in ascending order based on the total number of instances in V-COCO train set and HICO-DET train set. For each supercategory, we pick 20% rare classes as unseen classes, which results in 43 seen classes and 37 unseen classes (see supplementary materials for details). In this way, we construct a training set of 1,878 images in V-COCO [12] and 30,854 images in HICO-DET [4] with only seen object categories.

**Implementation details** Our model is built upon the ResNet50 [15] with Feature Pyramid Network (FPN) [22]. We adopt synchronized SGD training on 4 GPUs with 2 images per GPU. The learning rate is 0.005 with a weight decay of 0.0001 and a momentum of 0.9. We first search the best hyper-parameters on V-COCO val set. Then we train our model for 6k iterations on V-COCO trainval set.
<table>
<thead>
<tr>
<th>Methods</th>
<th>Train set</th>
<th>k = 100</th>
<th>k = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>COCO</td>
<td>VCOCO</td>
<td>all seen &amp; unseen</td>
</tr>
<tr>
<td>RPN w/ FPN</td>
<td>✓</td>
<td>✓</td>
<td>83.6</td>
</tr>
<tr>
<td>RPN w/ FPN</td>
<td>✓</td>
<td>✓</td>
<td>87.1</td>
</tr>
<tr>
<td>HO-RPN (ours)</td>
<td>✓</td>
<td>✓</td>
<td><strong>98.9</strong></td>
</tr>
</tbody>
</table>

(a) **Generated region proposals.** Results are evaluated on V-COCO val set. We train network with only annotations of seen categories. A true positive is considered if the generated region proposals have an IoU > 0.5 with ground truth bounding boxes.

<table>
<thead>
<tr>
<th>#boxes</th>
<th>VCOCO</th>
<th>HICO-DET</th>
<th>Time (s)</th>
<th>Methods</th>
<th>VCOCO</th>
<th>HICO-DET</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>36.2</td>
<td>43.7</td>
<td>0.214</td>
<td>ZSOD [2]</td>
<td>3.3</td>
<td>5.1</td>
</tr>
<tr>
<td>8</td>
<td>42.7</td>
<td>45.7</td>
<td>0.287</td>
<td>ZS-HOID w/o ( s_r )</td>
<td>3.4</td>
<td>5.3</td>
</tr>
<tr>
<td>12</td>
<td>42.8</td>
<td>45.9</td>
<td>0.356</td>
<td>ZS-HOID (ours)</td>
<td><strong>11.5</strong></td>
<td><strong>11.3</strong></td>
</tr>
</tbody>
</table>

(c) **Ablation on #human boxes.** Results are AP@IoU=0.5 and the average inference time per image.

<table>
<thead>
<tr>
<th>Methods</th>
<th>VCOCO</th>
<th>HICO-DET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster RCNN</td>
<td>28.2</td>
<td>33.2</td>
</tr>
<tr>
<td>Interactiveness [21]</td>
<td>36.6</td>
<td>41.4</td>
</tr>
<tr>
<td>HOID w/o ( s_r )</td>
<td>29.4</td>
<td>33.6</td>
</tr>
<tr>
<td>HOID (ours)</td>
<td>42.7</td>
<td>45.7</td>
</tr>
</tbody>
</table>

(b) **Interacting object detection.** Results are AP@IoU=0.5 among 79 object categories (except “person”).

<table>
<thead>
<tr>
<th>Embeddings</th>
<th>VCOCO</th>
<th>HICO-DET</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe [29]</td>
<td>10.4</td>
<td>9.1</td>
</tr>
<tr>
<td>FastText [16]</td>
<td>10.2</td>
<td>9.9</td>
</tr>
<tr>
<td>GoogleNews [26]</td>
<td><strong>11.5</strong></td>
<td><strong>11.3</strong></td>
</tr>
</tbody>
</table>

(d) **Novel object detection.** Results are AP@IoU=0.5 over all unseen categories based on our seen/unseen split.

<table>
<thead>
<tr>
<th>Methods</th>
<th>VCOCO</th>
<th>HICO-DET</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZSOD</td>
<td>3.3</td>
<td>5.1</td>
</tr>
<tr>
<td>ZS-HOID w/o ( s_r )</td>
<td>3.4</td>
<td>5.3</td>
</tr>
<tr>
<td>ZS-HOID (ours)</td>
<td><strong>11.5</strong></td>
<td><strong>11.3</strong></td>
</tr>
</tbody>
</table>

(e) **Ablation on semantic embedding.** Results are AP@IoU=0.5 over all unseen object categories.

Table 1: Ablation studies on various components in our proposed model.

set and 20k iterations on HICO-DET train set, and report the results on their test set.

**Evaluation metrics** Our central interest is to detect tuples \(<b_h, v, b_o, y>\). We evaluate the detection performance using mean Average Precision (mAP) following previous works [11, 44, 37, 21]. The Average Precision is first calculated per interaction class \(\{v, y\} \in \mathcal{V}, y \in \mathcal{Y}\) and then we take the mean. Formally, a detected tuple is considered as true positive if (1) the predicted human and object bounding box have an Intersection-over-Union (IoU) of 0.5 or higher with the ground truth, (2) the verb prediction is correct, and (3) the object category is correct. In the ablation studies, we use the COCO-style Average Precision at IoU = 0.5 (AP@IoU=0.5) to evaluate the performance of interacting object detection.

### 4.2. Ablation Analysis

In the following experiments, we call our HOI detection model using HO-RPN as HOID for short. When the proposed zero-shot classification module is used to detect novel objects, we call it as ZS-HOID.

**Generated region proposals** In Table 1a, we evaluate the quality of region proposals generated by our proposed HO-RPN using the recall at the top \( k \) proposals. We compare against the original RPN [35] built with FPN [22], where no human clues are used to generate object proposals. Here the ground truth only includes the objects interacting with humans. A true positive is considered if the proposal has an IoU > 0.5 with ground truth.

The first row in Table 1a refers to the model trained on COCO trainval set (~37k images excluding V-COCO val and test set and images with unseen classes). It shows the worst performance since it captures many noninteracting objects. For a fair comparison, we also fine-tune it on V-COCO train set with seen interacting objects. As reported, our HO-RPN can cover 1.5% more seen objects and 10.8% more unseen objects than RPN with FPN at the top 100 proposals. This result suggests the benefits of using human visual clues to localize objects. Figure 5 depicts some qualitative results on unseen object categories. As shown, our HO-RPN can better capture novel objects than the RPN.

**Interacting object detection** In Table 1b, we investigate the performance of our model, HOID, on detecting interacting objects. We compare against Faster RCNN [35] and two competitive baselines, InteractNet [11] and Interactiveness [21]. Given the boxes detected by Faster RCNN, InteractNet and InteractivenessNet perform post-processing to suppress non-interacting objects. In this experiment, all 80 MS-COCO object categories are used for training models. To study the impact of interactiveness score \( s_r \), we ablate it from the object box score in Eq.(7).

We evaluate the detected boxes on V-COCO test set and HICO-DET test set using AP@IoU=0.5. Notice that here detections of non-interacting objects will be seen as false positives. As shown, our method can outperform the best baseline by 6.6 and 4.3 points on V-COCO and HICO-DET. We observe that the ablation of \( s_r \) drops the AP by 13.3 and 12.1 points. It suggests that the interactiveness score predicted by HO-RPN makes the main contribution to our improvement. In Figure 6, we show some detection results of our method.

**Number of human boxes in Eq.(3)** In Table 1c, we change the number of human boxes in HO-RPN (i.e., \( K \) in Eq.(3)). We observe that as involving more human boxes, the performance on detecting interacting objects increases. However, it will cost more inference time. As a trade-off between performance and speed, we choose 8 human boxes in our experiments.
### Novel object detection

In Table 1d, we evaluate the performance of our model, ZS-HOID, on detecting unseen novel objects. As this task is closely related to zero-shot object detection, we compare against a state-of-the-art zero-shot object detector (ZSOD) [2]. Here we use word2vec [26] trained on GoogleNews corpus as the semantic word embedding. As shown, our method can achieve 8.2 and 6.2 points improvement over ZSOD. This improvement is mainly due to the interactiveness score produced by HoRPN, since the AP drops to 3.4 and 5.3 when we ablate it.

#### Semantic embeddings

In Table 1e, we test three semantic embeddings, GloVe [29] trained on Wikipedia 2014 and Gigaword 5 corpus, word2vec [26] on GoogleNews corpus, and FastText [16] on Wikipedia 2017 and UMBC corpus. It shows that word2vec on GoogleNews achieves the overall best performance. Thus, in the rest of experiments with zero-shot classification, we use the word2vec on GoogleNews as the semantic embedding.

#### Comparison of various seen/unseen splits

In our experiments, we observe that the successes of zero-shot classification strongly rely on the related seen categories. For instance, the novel category “surfboard” can be recognized if the classifier for seen categories “skateboard”, “snowboard” and “boat” has a relatively higher response than other seen categories. To investigate the influences of seen/unseen splits, we perform two different random splits: Split-A and Split-B. Split-A conducts splits per supercategory (e.g., animals, sports, vehicles). It ensures that each unseen category can have at least one seen category from the same supercategory. In comparison, Split-B selects an entire supercategory as seen or unseen. In this way, there is no common supercategory between seen and unseen. Figure 7 depicts the experimental results. As shown, the results of Split-B are generally worse than Split-A, even though it has more training samples. It suggests that having related seen categories is a key factor for detecting novel objects.

### HOI detection

In Table 2, we compare against state-of-the-art HOI detection methods. Here all annotated samples of 80 MS-COCO categories are used to train our model. We evaluate on HICO-DET dataset using the provided evaluation protocol. It shows that our method can achieve state-of-the-art performance on full and non-rare interactions. We believe that the promotion is due to the good performance of our detector, which produces only interacting objects and indirectly reduces the false positive interaction detections. It is worth noting that we do not utilize extra pose estimators to extract the human skeleton. But we believe that using the pose estimator can further improve our performance.

#### 4.3. Human-Novel-Object Interaction Detection

In this section, we evaluate the performance of human-novel-object interaction detection on V-COCO and HICO-DET datasets using our seen/unseen split. For comparison with existing HOI detection methods, we replace their object detectors with a zero-shot object detector (ZSOD) [2] which is trained with seen categories. Here we choose InteractNet [11] and Interactiveness [21] as competitors since their models can utilize human visual information to suppress noninteracting objects produced by the object detector. We re-train their interaction models using annotations of seen categories and evaluate on the full test set with...
both seen and unseen categories.

The results on V-COCO and HICO-DET datasets are shown in Table 3 and Table 4. We separately report the mean Average Precision (mAP) over (1) interactions with seen categories and (2) interactions with unseen categories. Our main focus is on detecting human-novel-object interactions. For V-COCO dataset, as shown in Table 3, our method outperforms the best baseline by 6.21 points on unseen categories. For the challenging HICO-DET dataset, we increase the mAP by nearly 1.86× over the best baseline on interactions with unseen categories.

4.4. Scaling HOIs by Extending Object Space

To further demonstrate the ability of our method on detecting HOIs with novel object categories, we collect more images from Visual Genome [18] to test our model (see more details in supplementary materials). Besides the 80 MS-COCO categories, we collect 30 novel object categories (e.g., camera, fishing pole, guitar, pen, etc.). Here we use the model trained on HICO-DET train set with the annotations of all 80 MS-COCO categories to detect human interactions with novel objects. In this experiment, we mainly use qualitative evaluation. Some results of our model are shown in Figure 8.

5. Conclusion

In this paper, we propose a novel human-object region proposal network for the human-object interaction detection task. It leverages human visual clues to find objects. We show that our proposed model can well detect interacting objects, even though they belong to unseen object categories. Furthermore, we design a zero-shot classification module to recognize novel objects. These contributions allow us to detect human interactions with unseen object categories.

Acknowledgement This work was conducted within the DeltaNTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc. and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme.
References

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kaneltis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 2, 5, 8
[21] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactivity knowledge for human-object interaction detection. In CVPR, 2019. 1, 2, 3, 6, 7, 8


[37] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Fei Fei Li. Scaling human-object interaction recognition through zero-shot learning. In WACV, 2018. 1, 2, 6

[38] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In NIPS, 2013. 2


[44] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S. Kankanhalli. Learning to detect human-object interactions with knowledge. In CVPR, 2019. 1, 2, 3, 6, 7


