Discovering Human Interactions with Novel Objects via Zero-Shot Learning

Suchen Wang[1] Kim-Hui Yap[1] Junsong Yuan[2] Yap-Peng Tan[1]
[1]Nanyang Technological University [2]State 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.

Figure 1: We aim to scale the object space in HOI detection via zero-shot learning. This figure depicts the output of our model. Apart from the 80 object categories (green boxes) as defined in MS-COCO, our model can detect human interactions with unseen object categories, *e.g.*, rose (red box) in this image.

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.

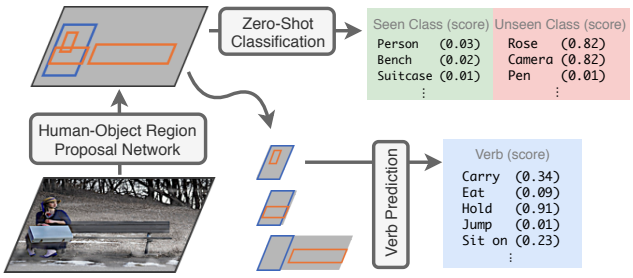

Figure 2: Overview of our model. It consists of three main modules. (1) Our proposed human-object region proposal network (HO-RPN) first localizes humans and interacting objects. It estimates an interactiveness score for each object proposal. (2) The zero-shot classification module classifies the generated 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[1] 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]. Previous attempts to scale HOIs using zero-shot learning [17, 1, 37] mainly focus on the unseen interactions with known objects. When scaling HOIs by adding new object categories, they require the bounding box annotations of new object categories to re-train their detector. In comparison, we do not require annotations of new object categories except their semantic word embeddings learned from text corpus [29, 26]. Hence, our method can scale HOIs more easily without the need of much human labor.

Zero-Shot Learning (ZSL) 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.

[1] https://github.com/scwangdyd/zero_shot_hoi

3. Approach

3.1. Problem Statement

The goal of human-object interaction (HOI) detection is to find one or multiple tuples <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 $b_h, b_o \in \mathbb{R}^4$ 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 $\mathcal{Y}_\mathcal{S} = \{1, \ldots, c_1\}$ and $\mathcal{Y}_\mathcal{U} = \{c_1 + 1, \ldots, c_1 + c_2\}$ 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.*, $y \in \mathcal{Y} = \mathcal{Y}_\mathcal{S} \cup \mathcal{Y}_\mathcal{U}$. To make this a feasible task, we assume that each object category $y \in \mathcal{Y}$ has a semantic embedding vector $\mathbf{q}_y \in \mathbb{R}^p$ 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 $p(b_h, y = \text{“person”}|x)$ 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

$$p(b_h, v, b_o, y|x) = p(v|b_h, b_o, y, x)p(b_o, y|x)p(b_h|x) \quad (1)$$

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.*, $p(b_h, b_o, y|x) = p(b_o, y|x)p(b_h|x)$. 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.

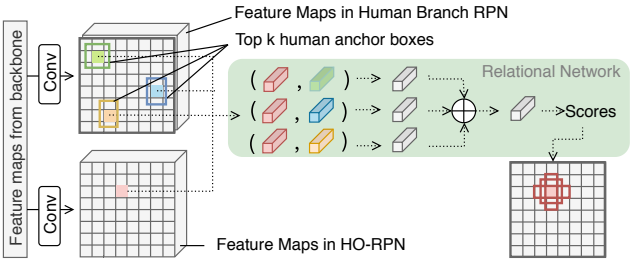

Figure 3: **Overview of HO-RPN**. Each cell in the feature map represents a sliding window position. The features of top k human anchor boxes will be used to score the object anchor boxes using a relational network.

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

$$p(v|b_h, b_o, y, x)p(b_o, y|b_h, x)p(b_h|x) \quad (2)$$

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 objectiveness. 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 a 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 interactiveness score of n anchor boxes with different shape, $s_r = [s_r^{(1)}, \dots, s_r^{(n)}] \in \mathbb{R}^n$, based on its visual features and relationships with detected humans. Let $\mathbf{x}_o^{(j)} \in \mathbb{R}^d$ denote the feature of j-th sliding window position in HO-RPN, and $\mathbf{x}_h^{(k)} \in \mathbb{R}^d$ denote the feature of k-th human anchor box from human RPN. Specifically, the relational network computes the score by

$$s_r = \sigma\Big(f\big(\sum_{k=1}^{K} g(\mathbf{x}_o^{(j)}, \mathbf{x}_h^{(k)})\big)\Big) \tag{3}$$

where $\sigma(\cdot)$ is a sigmoid function. Here $g(\cdot) : \mathbb{R}^{2d} \mapsto \mathbb{R}^d$ is a simple multi-layer perceptron (MLP), which processes the concatenated feature of $\mathbf{x}_o^{(j)}$ and $\mathbf{x}_h^{(k)}$. The role of $g(\cdot)$ is to infer if the object and human are interacting. Here $f(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^n$ 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 $y \in \mathcal{Y}_\mathcal{S}$ predicted by a classifier. Given $\{f_y(b_o)\}_{y \in \mathcal{Y}_\mathcal{S}}$, we denote the most likely seen category as

$$\hat{y}_1 := \arg\max_{y \in \mathcal{Y}_\mathcal{S}} f_y(b_o) \tag{4}$$

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

$$\mathbf{e} = \frac{1}{Z} \sum_{j=1}^{K} f_{\hat{y}_j}(b_o) \cdot \mathbf{q}_{\hat{y}_j} \tag{5}$$

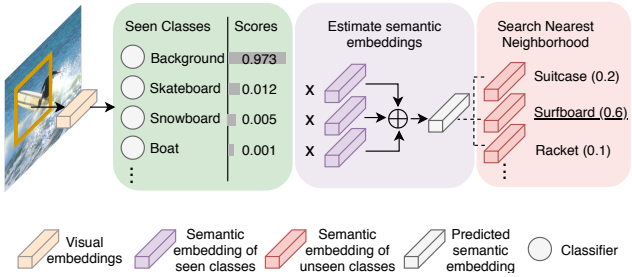

Figure 4: **The pipeline of zero-shot classification.** The softmax classifier for seen categories (including the background) first classifies the region proposal. If the background category has a high response, we then identify if it belongs to unseen categories. We estimate a semantic embedding using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

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

Besides the categories in $\mathcal{Y}_\mathcal{S}$, the background is also a category in the softmax classifier. If the softmax classifier is very confident in its prediction for $y =$ "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 $\mathbf{e}$ to the semantic embedding vector of unseen categories as

$$s_y = cos(\mathbf{e}, \mathbf{q}_y),\ y \in \mathcal{Y}_\mathcal{U} \tag{6}$$

where $cos(\cdot, \cdot)$ is the cosine similarity. If $s_y < \tau$ for any $y \in \mathcal{Y}_\mathcal{U}$, 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.*, $s_y = f_y(b_o), y \in \mathcal{Y}_\mathcal{S}$. 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

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

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

$$p(v|b_h, b_o, y, x) := \sigma\big(h_v(\mathbf{x}_o, \mathbf{x}_h, \mathbf{x}_{h,o})\big) \tag{8}$$

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

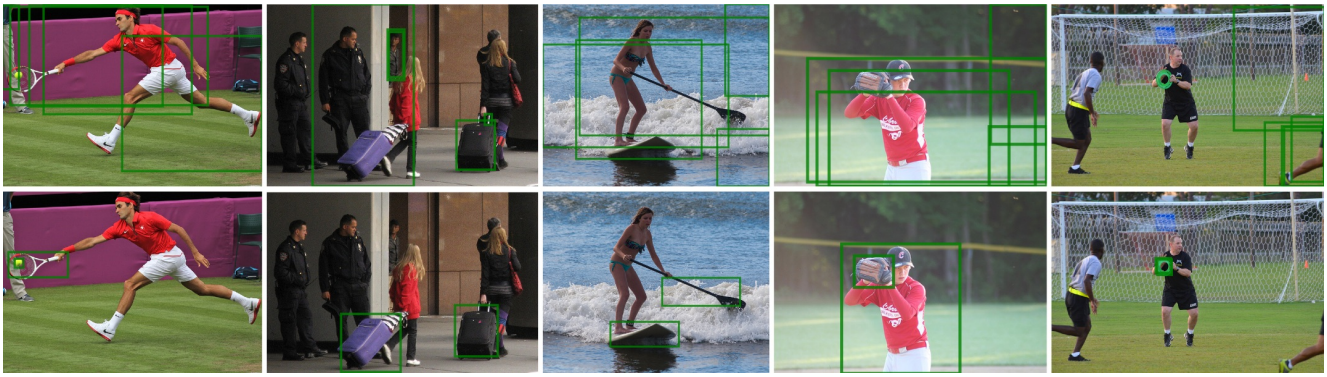

Figure 5: **Qualitative comparison of HO-RPN (bottom row) with RPN (upper row) on unseen novel objects.** In this figure, tennis racket, suitcase, surfboard, baseball glove, and frisbee are unseen by the model based on our `seen/unseen` split. For RPN (upper row), we visualize the top 5 object region proposals (except proposals have an IoU > 0.5 with humans). For our proposed HO-RPN, the top 2 object region proposals are enough to capture novel objects in the example images.

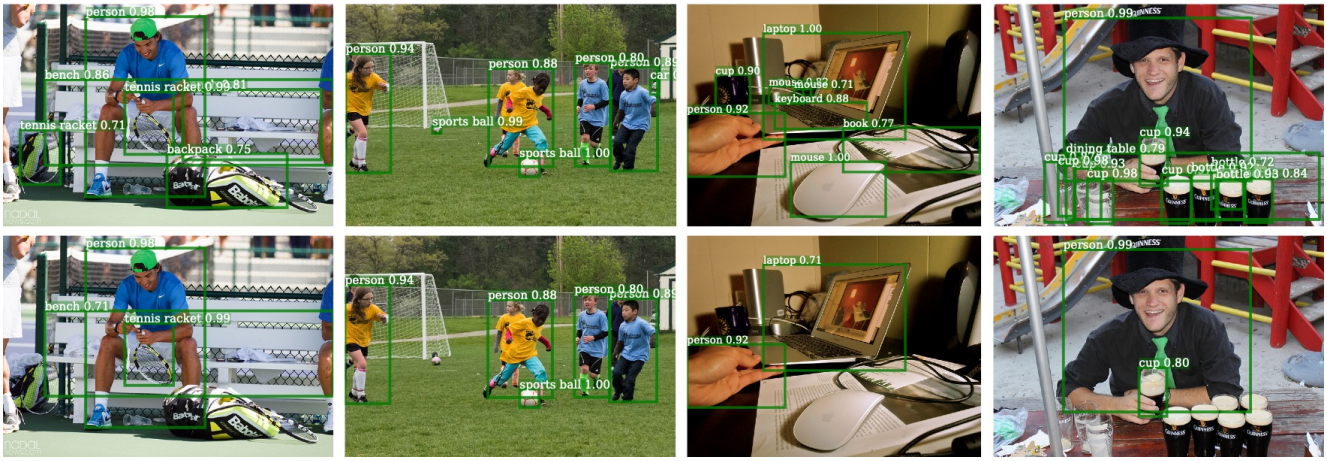

Figure 6: **Qualitative results of interacting object detection.** Compared with Faster RCNN (top row), our model (bottom row) only detects interacting objects. Boxes with a score > 0.7 are visualized.

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`

Recall@k Methods	Train set COCO	Train set VCOCO	$k=100$ all	$k=100$ seen	$k=100$ unseen	$k=500$ all	$k=500$ seen	$k=500$ unseen
RPN w/ FPN	✓		83.6	88.7	50.3	90.2	93.5	68.6
RPN w/ FPN	✓	✓	87.1	90.7	63.8	94.4	96.1	83.5
HO-RPN (ours)		✓	**89.8**	**92.2**	**74.6**	**95.5**	**96.6**	**88.5**

(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.

Methods	VCOCO	HICO-DET
Faster RCNN	28.2	33.2
InteractNet [11]	36.2	39.3
Interactiveness [21]	36.6	41.4
HOID w/o s_r	29.4	33.6
HOID (ours)	**42.7**	**45.7**

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

#boxes	VCOCO	HICO-DET	Time (s)
4	36.2	43.7	0.214
8	42.7	45.7	0.287
12	42.8	45.9	0.356

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

Methods	VCOCO	HICO-DET
ZSOD [2]	3.3	5.1
ZS-HOID w/o s_r	3.4	5.3
ZS-HOID (ours)	**11.5**	**11.3**

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

Embeddings	VCOCO	HICO-DET
GloVe [29]	10.4	9.1
FastText [16]	10.2	9.9
GoogleNews [26]	**11.5**	**11.3**

(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 $\{v, y\}_{v\in\mathcal{V}, y\in\mathcal{Y}}$ 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 finetune 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 noninteracting 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 noninteracting 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.

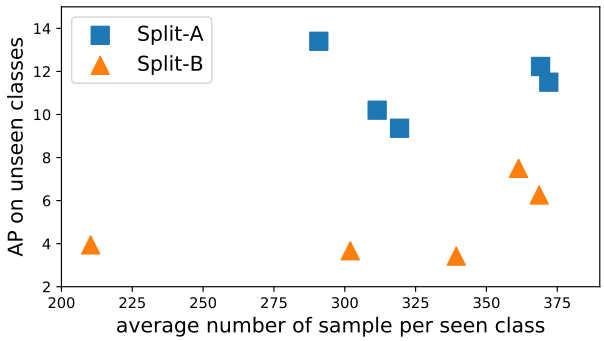

Figure 7: **Ablation on various seen/unseen splits**. Each point corresponds to an experiment. The x-axis represents the average number of training samples per class. The y-axis is the AP@IoU=0.5 (%) over unseen classes.

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 HO-RPN, since the AP drops to 3.4 and 5.3 when we ablate it from our model.

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.

Methods	No pose estimator	HICO-DET (default) Full	Rare	Non-rare
VSRL [12]	✓	9.09	7.02	9.71
InteractNet [11]	✓	9.94	7.16	10.77
GPNN [32]	✓	13.11	9.34	14.23
iCAN [10]	✓	14.84	10.45	16.15
Knowledge [44]	✓	14.70	13.26	15.13
Contextual Attention [40]	✓	16.24	11.16	17.75
No-Frills (no pose) [13]	✓	16.96	11.95	18.46
HOID (ours)	✓	**17.85**	12.85	**19.34**
No-Frills (with pose) [13]		17.18	12.17	18.68
Interactiveness [21]		17.22	13.51	18.32
PMFNet [39]		17.46	**15.65**	18.00

Table 2: **HOI detection.** We compare against state-of-the-art HOI detection methods. Results are mAP (%) evaluated on HICO-DET `test` set. In this experiment, all 80 MS-COCO categories are used for training.

Methods	`Seen`	`Unseen`	`All`
ZSOD + InteractNet [11]	38.64	10.97	27.88
ZSOD + Interactiveness [21]	39.70	13.67	29.58
ZS-HOID w/o s_r (our baseline)	37.53	10.61	27.06
ZS-HOID (ours)	**43.13**	**19.88**	**34.09**

Table 3: **Experimental comparison of human-novel-object interaction detection.** Results are mAP (%) evaluated on V-COCO `test` set based on our `seen`/`unseen` split.

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

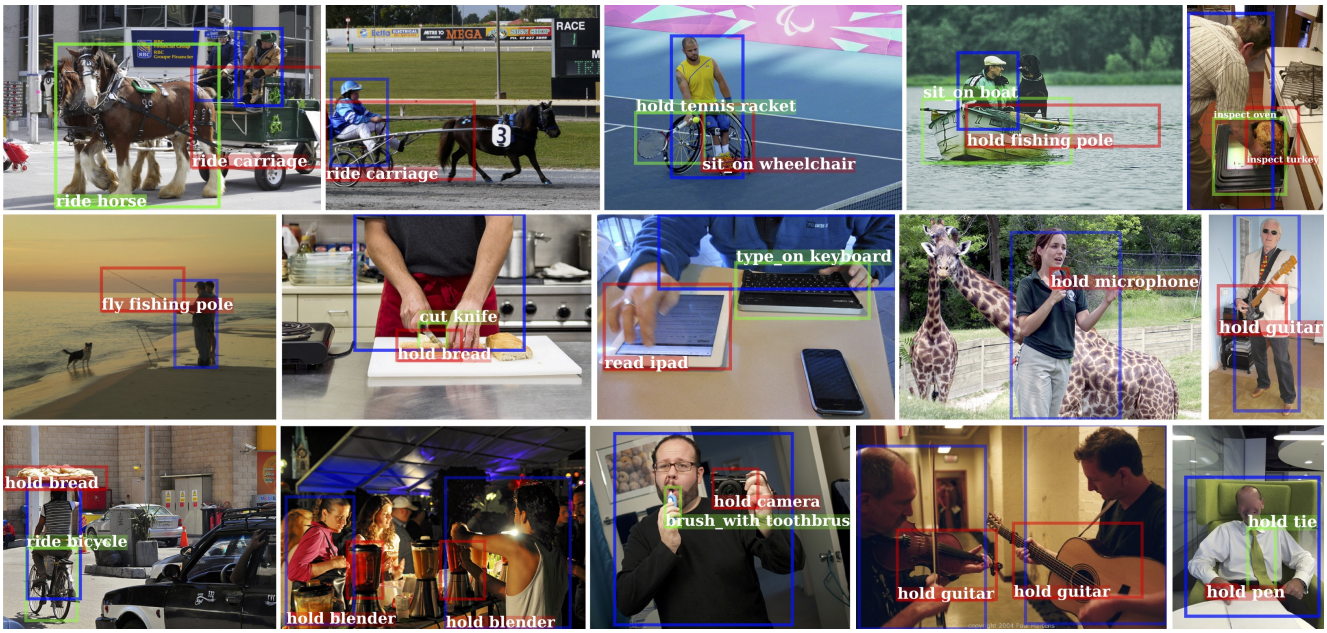

Figure 8: **Results of our model on human-novel-object interaction detection.** Our model is trained with annotations of all 80 MS-COCO object categories (green boxes), while it can detect interactions with unseen object categories (red boxes).

	Methods	accessory	animal	appliance	electronic	food	furniture	indoor	kitchen	outdoor	sports	vehicle	all
Seen	ZSOD+InteractNet [11]	9.55	16.09	4.82	8.88	5.54	6.12	9.67	6.30	11.02	12.93	14.64	12.31
	ZSOD+Interactiveness[21]	15.04	19.10	7.90	12.23	7.59	6.76	10.33	10.66	25.84	15.63	17.40	15.92
	ZS-HOID w/o s_r (our baseline)	9.05	14.40	4.60	7.94	5.17	7.20	9.16	5.88	11.08	11.57	12.98	11.11
	ZS-HOID (ours)	**19.51**	**25.76**	**11.56**	**16.23**	**11.69**	**10.30**	**18.71**	**12.11**	**30.70**	**21.51**	**20.88**	**21.19**
Unseen	ZSOD+InteractNet [11]	0.82	1.64	0.23	0.00	0.82	3.87	0.67	0.07	0.00	2.41	2.20	1.62
	ZSOD+Interactiveness[21]	2.13	3.10	0.07	0.61	0.24	0.82	0.64	0.69	1.97	2.60	1.67	1.52
	ZS-HOID w/o s_r (our baseline)	0.92	1.44	0.13	0.00	0.75	3.41	0.58	0.06	0.00	2.05	2.18	1.45
	ZS-HOID (ours)	**5.32**	**4.98**	**4.63**	**0.34**	**1.48**	**4.04**	**0.01**	**0.44**	**3.31**	**3.50**	**3.71**	**3.02**

Table 4: **Experimental comparison of human-novel-object interaction detection.** Results are mAP (%) evaluated on HICO-DET `test` set based on our `seen`/`unseen` split.

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 $1.86\times$ 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 Delta-NTU 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

[1] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. 2020. 1, 2

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 2, 3, 5, 6, 7

[3] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2, 5

[5] 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. 1

[6] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[7] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 3

[8] 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

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2

[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, 7

[11] Georgia Gkioxari, Ross Girshick, Piotr Dollr, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 1, 2, 3, 6, 7, 8

[12] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1, 2, 3, 5, 7

[13] 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, 3, 7

[14] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

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

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651, 2016. 6, 7

[17] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *The European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, 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

[19] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009. 2

[20] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *The European Conference on Computer Vision (ECCV)*, 2018. 3

[21] 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. 1, 2, 3, 6, 7, 8

[22] T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1, 2, 3, 5, 6

[23] 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, 5

[24] Jingen Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, 2011. 2

[25] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016. 2

[26] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013. 2, 6, 7

[27] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[28] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014. 2, 4

[29] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 2, 6, 7

[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. 1, 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] 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, 3, 7

[33] Shafin Rahman, Salman Khan, and Nick Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018. 2, 5

[34] Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. *International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 5

[35] 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, 6

[36] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30*. 2017. 4

[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

[39] 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. 1, 2, 7

[40] 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. 1, 2, 7

[41] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[42] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[43] 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. 1, 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

[45] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[46] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. On exploring undetermined relationships for visual relationship detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[47] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[48] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[49] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Generalized zero-shot recognition based on visually semantic embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2