Discovering Human-Object Interaction Concepts via Self-Compositional Learning

Zhi Hou[1], Baosheng Yu[1], and Dacheng Tao[1,2]

[1] The University of Sydney, Australia
[2] JD Explore Academy, China
zhou9878@uni.sydney.edu.au,
baosheng.yu@sydney.edu.au,dacheng.tao@gmail.com

Abstract. A comprehensive understanding of human-object interaction (HOI) requires detecting not only a small portion of predefined HOI concepts (or categories) but also other reasonable HOI concepts, while current approaches usually fail to explore a huge portion of unknown HOI concepts (i.e., unknown but reasonable combinations of verbs and objects). In this paper, 1) we introduce a novel and challenging task for a comprehensive HOI understanding, which is termed as **HOI Concept Discovery**; and 2) we devise a self-compositional learning framework (or **SCL**) for HOI concept discovery. Specifically, we maintain an online updated concept confidence matrix during training: 1) we assign pseudo labels for all composite HOI instances according to the concept confidence matrix for self-training; and 2) we update the concept confidence matrix using the predictions of all composite HOI instances. Therefore, the proposed method enables the learning on both known and unknown HOI concepts. We perform extensive experiments on several popular HOI datasets to demonstrate the effectiveness of the proposed method for HOI concept discovery, object affordance recognition and HOI detection. For example, the proposed self-compositional learning framework significantly improves the performance of 1) HOI concept discovery by over **10%** on HICO-DET and over **3%** on V-COCO, respectively; 2) object affordance recognition by over **9%** mAP on MS-COCO and HICO-DET; and 3) rare-first and non-rare-first unknown HOI detection relatively over **30%** and **20%**, respectively. Code is publicly available at https://github.com/zhihou7/HOI-CL.

Keywords: Human-Object Interaction, HOI Concept Discovery, Object Affordance Recognition

1 Introduction

Human-object interaction (HOI) plays a key role in analyzing the relationships between humans and their surrounding objects [21], which is of great importance for deep understanding on human activities/behaviors. Human-object interaction understanding has attracted extensive interests from the community, including image-based [7, 5, 17, 38, 52], video-based visual relationship analysis [11, 40],

video generation [42], and scene reconstruction [63]. However, the distribution of HOI samples is naturally long-tailed: most interactions are rare and some interactions do not even occur in most scenarios, since we can not obtain an interaction between human and object until someone conducts such action in real-world scenarios. Therefore, recent HOI approaches mainly focus on the analysis of very limited predefined HOI concepts/categories, leaving the learning on a huge number of unknown HOI concepts [9, 3] poorly investigated, including HOI detection and object affordance recognition [50, 25, 26]. For example, there are only 600 HOI categories known in HICO-DET [6], while we can find 9,360 possible verb-object combinations from 117 verbs and 80 objects.

Fig. 1. An illustration of unknown HOI detection via concept discovery. Given some known HOI concepts (*e.g.*, "drink_with cup", "drink_with bottle", and "hold bowl"), the task of concept discovery aims to identify novel HOI concepts (i.e., reasonable combinations between verbs and objects). For example, here we have some novel HOI concepts, "drink_with wine_glass", "fill bowl", and "fill bottle". Specifically, the proposed self-compositional learning framework jointly optimizes HOI concept discovery and HOI detection on unknown concepts in an end-to-end manner.

Object affordance is closely related to HOI understanding from an object-centric perspective. Specifically, two objects with similar attributes usually share the same affordance, *i.e.* , humans usually interact with similar objects in a similar way [18]. For example, cup, bowl, and bottle share the same attributes (*e.g.*, hollow), and all of these objects can be used to "drink with". Therefore, object affordance [18, 26] indicates whether each action can be applied into an object, *i.e.* , if a verb-object combination is reasonable, we then find a novel HOI concept/category. An illustration of unknown HOI detection via concept discovery is shown in Fig. 1. Recently, it has turned out that an HOI model

is not only capable of detecting interactions, but also able to recognize object affordances [26], especially novel object affordances using the composite HOI features. Particularly, novel object affordance recognition also indicates discovering novel reasonable verb-object combinations or HOI concepts. Inspired by this, we can introduce a simple baseline for HOI concept discovery by averaging the affordance predictions of training dataset into each object category [26].

Nevertheless, there are two main limitations when directly utilizing object affordance prediction [26] for concept discovery. First, the affordance prediction approach in [26] is time-consuming and unsuitable to be utilized during training phrase, since it requires to predict all possible combinations of verbs and objects using the whole training set. By contrast, we introduce an online HOI concept discovery method, which is able to collect concept confidence in a running mean manner with verb scores of all composite features in mini-batches during training. Second, also more importantly, the compositional learning approach [26] merely optimizes the composite samples with known concepts (*e.g.*, 600 categories on HICO-DET), ignoring a large number of composite samples with unknown concepts (unlabeled composite samples). As a result, the model is inevitably biased to known object affordances (or HOI concepts), and leads to the similar inferior performance to the one in Positive-Unlabeled learning [12, 14, 46]. That is, without negative samples for training, the network will tend to predict high confidence on those impossible verb-object combinations or overfit verb patterns (please refer to Appendix A for more analysis). Considering that the online concept discovery branch is able to predict concept confidence during optimization, we can then construct pseudo labels [35] for all composite HOIs belonging to either known or unknown categories. Inspired by this, we introduce a self-compositional learning strategy (or SCL) to jointly optimize all composite representations and improve concept predictions in an iterative manner. Specifically, SCL combines the object representations with different verb representations to compose new samples for optimization, and thus implicitly pays attention to the object representations and improves the discrimination of composite representations. By doing this, we can improve the object affordance learning, and then facilitate the HOI concept discovery.

Our main contributions can be summarized as follows: 1) we introduce a new task for a better and comprehensive understanding on human-object interactions; 2) we devise a self-compositional learning framework for HOI concept discovery and object affordance recognition simultaneously; and 3) we evaluate the proposed approach on two extended benchmarks, and it significantly improves the performance of HOI concept discovery, facilitates object affordance recognition with HOI model, and also enables HOI detection with novel concepts.

2 Related Work

2.1 Human-Object Interaction

HOI understanding [21] is of great importance for visual relationship reasoning [58] and action understanding [4, 64]. Different approaches have been inves-

tigated for HOI understanding from various aspects, including HOI detection [6, 37, 38, 65, 32, 8, 67, 52, 62], HOI recognition [7, 31, 28], video HOI [11, 30], compositional action recognition [40], 3D scene reconstruction [63, 10], video generation [42], and object affordance reasoning [15, 26]. Recently, compositional approaches (*e.g.*, VCL [25]) have been intensively proposed for HOI understanding using the structural characteristic [31, 25, 42, 36, 26]. Meanwhile, DETR-based methods (*e.g.*, Qpic [52]) achieve superior performance on HOI detection. However, these approaches mainly consider the perception of known HOI concepts, and pay no attention to HOI concept discovery. To fulfill the gap between learning on known and unknown concepts, a novel task, *i.e.* , HOI concept discovery, is explored in this paper. Currently, zero-shot HOI detection also attracts massive interests from the community [50, 2, 45, 25, 27]. However, those approaches merely consider known concepts and are unable to discover HOI concepts. Some HOI approaches [45, 2, 56, 55] expand the known concepts via leveraging language priors. However, that is limited to existing knowledge and can not discover concepts that never appear in the language prior knowledge. HOI concept discovery is able to address the problem, and enable unknown HOI concept detection.

2.2 Object Affordance Learning

The notation of affordance is formally introduced in [18], where object affordances are usually those action possibilities that are perceivable by an actor [43, 18, 19]. Noticeably, the action possibilities of an object also indicate the HOI concepts related to the object. Therefore, object affordance can also represent the existence of HOI concepts. Recent object affordance approaches mainly focus on the pixel-level affordance learning from human interaction demonstration [34, 16, 15, 23, 41, 13, 61]. Yao *et al.* [60] present a weakly supervised approach to discover object functionalities from HOI data in the musical instrument environment. Zhu *et al.* [66] introduce to reason affordances in knowledge-based representation. Recent approaches propose to generalize HOI detection to unseen HOIs via functionality generalization [2] or analogies [45]. However those approaches focus on HOI detection, ignoring object affordance recognition. Specifically, Hou *et al.* [26] introduce an affordance transfer learning (ATL) framework to enable HOI model to not only detect interactions but also recognize object affordances. Inspired by this, we further develop a self-compositional learning framework to facilitate the object affordance recognition with HOI model to discover novel HOI concepts for downstream HOI tasks.

2.3 Semi-Supervised Learning

Semi-supervised learning is a learning paradigm for constructing models that use both labeled and unlabeled data [59]. There are a wide variety of Deep Semi-Supervised Learning methods, such as Generative Networks [33, 51], Graph-Based methods [54, 20], Pseudo-Labeling methods [35, 57, 24]. HOI concept discovery shares a similar characteristic to semi-supervised learning approaches. *H*OI concept discovery has instances of labeled HOI concepts, but no instances

of unknown concepts. We thus compose HOI representations for unknown concepts according to [47]. With composite HOIs, concept discovery and object affordance recognition can be treated as PU learning [12]. Moreover, HOI concept discovery requires to discriminate whether the combinations (possible HOI concepts) are reasonable and existing. Considering each value of the concept confidences also represents the possibility of the composite HOI, we construct pseudo labels [35, 47] for composite features from the concept confidence matrix, and optimize the composite HOIs in an end-to-end way.

3 Approach

In this section, we first formulate the problem of HOI concept discovery and introduce the compositional learning framework. We then describe a baseline for HOI concept discovery via affordance prediction. Lastly, we introduce the proposed self-compositional learning framework for online HOI concept discovery and object affordance recognition.

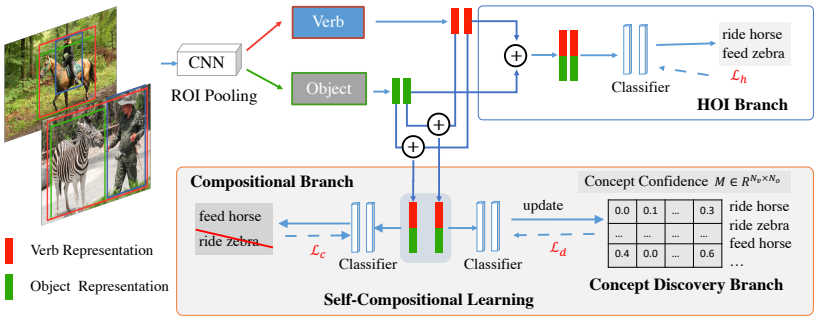

Fig. 2. Illustration of Self-Compositional Learning for HOI Concept Discovery. Specifically, following [25], verb and object features are extracted via RoI-Pooling from union box and object box respectively, which are then used to construct HOI features in HOI branch according to HOI annotation. Following [25], for SCL, verb and object features are further mutually combined to generate composite HOI features. Then, the feasible composite HOI features belonging to the known concepts are directly used to train the network in Compositional Branch. Here the classifier predicts verb classes directly. Meanwhile, we update the concept confidence $\mathbf{M} \in R^{N_v \times N_o}$, where N_v and N_o are the number of verb classes and object classes respectively, with the predictions of all composite HOI features. The concept discovery branch is optimized via a self-training approach to learn from composite HOI features with the concept confidence $\mathbf{M}$.

3.1 Problem Definition

HOI concept discovery aims to discover novel HOI concepts/categories using HOI instances from existing known HOI categories. Given a set of verb categories $\mathcal{V}$

and a set of object categories $\mathcal{O}$, let $\mathcal{S} = \mathcal{V} \times \mathcal{O}$ indicate the set of all possible verb-object combinations. Let $\mathcal{S}^k$, $\mathcal{S}^u$, and $\mathcal{S}^o$ denote three disjoint sets, known HOI concepts, unknown HOI concepts, and invalid concepts (or impossible verb-object combinations), respectively. That is, we have $\mathcal{S}^k \cap \mathcal{S}^u = \varnothing$ and $\mathcal{S}^k \cup \mathcal{S}^u = \mathcal{S}$ if $\mathcal{S}^o = \varnothing$. Let $\mathcal{T} = \{(h_i, c_i)\}_{i=1}^{L}$ indicate the training dataset, where h_i is a HOI instance (*i.e.* , verb-object visual representation pair), $c_i \in \mathcal{S}^k$ indicates the label of the i-th HOI instance and L is the total number of HOI instance.

We would also like to clarify the difference between the notations of "unknown HOI categories" and "unseen HOI categories" in current HOI approaches as follows. Let $\mathcal{S}^z$ indicate the set of "unseen HOI categories" and we then have $\mathcal{S}^z \subseteq \mathcal{S}^k$. Specifically, "unseen HOI category" indicates that the HOI concept is known but no corresponding HOI instances can be observed in the training data. Current HOI methods usually assume that unseen HOI categories $\mathcal{S}^z$ are known HOI categories via the prior knowledge [50, 31, 45, 2, 25]. Therefore, existing HOI methods can not directly detect/recognize HOIs with unknown HOI concepts. HOI concept discovery aims to find $\mathcal{S}^u$ from the existing HOI instances in $\mathcal{T}$ with only known HOI concepts in $\mathcal{S}^k$.

3.2 HOI Compositional Learning

Inspired by the compositional nature of HOI, *i.e.* , each HOI consists of a verb and an object, visual compositional learning has been intensively explored for HOI detection by combining visual verb and object representations [31, 25, 27, 26]. Let $\mathbf{h}_i = \langle \mathbf{x}_{v_i}, \mathbf{x}_{o_i} \rangle$ indicate a HOI instance, where $\mathbf{x}_{v_i}$ and $\mathbf{x}_{o_i}$ denote the verb and object representations, respectively. The HOI compositional learning then aims to achieve the following objective,

$$g_h(\langle \widetilde{\mathbf{x}}_{v_i}, \widetilde{\mathbf{x}}_{o_i} \rangle) \approx g_h(\langle \mathbf{x}_{v_i}, \mathbf{x}_{o_i} \rangle), \tag{1}$$

where g_h indicates the HOI classifier, $\mathbf{x}_{v_i}$ and $\mathbf{x}_{o_i}$ indicate the real verb-object representation pair (*i.e.* , annotated HOI pair in dataset), $\langle \widetilde{\mathbf{x}}_{v_i}, \widetilde{\mathbf{x}}_{o_i} \rangle$ indicates the composite verb-object pair. Specifically, $\widetilde{\mathbf{x}}_{o_i}$ can be obtained from either real HOIs [25], fabricated objects or language embedding [27, 2, 45], or external object datasets [26], while $\widetilde{\mathbf{x}}_{v_i}$ can be from real HOIs (annotated verb-object pair) and language embeddings [31, 45]. As a result, when composite HOIs are similar to real HOIs, we are then able to augment HOI training samples in a compositional manner. However, current compositional approaches for HOI detection [25, 26] simply remove the composite HOI instances out of the label space, which may also remove a large number of feasible HOIs (*e.g.*, "ride zebra" as shown Figure 2). Furthermore, the compositional approach can not only augment the training data for HOI recognition, but also provide a method to determinate whether $\widetilde{\mathbf{x}}_{v_i}$ and $\widetilde{\mathbf{x}}_{o_i}$ are combinable to form a new HOI or not [26], *i.e.* , discovering the HOI concepts.

3.3 Self-Compositional Learning

In this subsection, we introduce the proposed self-compositional learning framework for HOI concept discovery as follows. As shown in Figure 2, the main

HOI concept discovery framework falls into the popular two-stage HOI detection framework [25]. Specifically, we compose novel HOI samples from pair-wise images to optimize the typical HOI branch (annotated HOIs), compositional branch (the composite HOIs out of the label space are removed [25, 26]) and the new concept discovery branch (all composite HOIs are used). The main challenge of HOI concept discovery is the lack of instances for unknown HOI concepts, but we can infer to discover new concepts according to the shared verbs and objects. Specifically, we find that the affordance transfer learning [26] can be used for not only the object affordance recognition but also the HOI concept discovery, and we thus first introduce the affordance-based method as a baseline as follows.

Affordance Prediction The affordance transfer learning [26] or ATL is introduced for affordance recognition using the HOI detection model. However, it has been ignored that the affordance prediction can also enable HOI concept discovery, *i.e.* , predicting a new affordance for an object although the affordance is not labeled during training. We describe a vanilla approach for HOI concept discovery using affordance prediction [26]. Specifically, we predict the affordances for all objects in the training set according to [26]. Then, we average the affordance predictions according to each object category to obtain the HOI concept confidence matrix $\mathbf{M} \in R^{N_v \times N_o}$, where each value represents the concept confidence of the corresponding combination between a verb and an object. N_v and N_o are the numbers of verb and object categories, respectively. For simplicity, we may use both vector and matrix forms of the confidence matrix $\mathbf{M} \in R^{N_v N_o}$ and $\mathbf{M} \in R^{N_v \times N_o}$ in this paper. Though affordance prediction can be used for HOI concept discovery, it is time-consuming since it requires to predict affordances of all objects in training set. Specifically, we need an extra offline affordance prediction process to infer concepts with the computational complexity $O(N^2)$ in [26], where N is the number of total training HOIs, *e.g.*, it takes 8 hours with one GPU to infer the concept matrix $\mathbf{M}$ on HICO-DET. However, we can treat the verb representation as affordance representation [26], and obtain the affordance predictions for all objects in each mini-batch during training stage. Inspired by the running mean manner in [29], we devise an online HOI concept discovery framework via averaging the predictions in each mini-batch.

Online Concept Discovery As shown in Figure 2, we keep a HOI concept confidence vector during training, $\mathbf{M} \in R^{N_v N_o}$, where each value represents the concept confidence of the corresponding combination between a verb and an object. To achieve this, we first extract all verb and object representations among pair-wise images in each batch as $\mathbf{x}_v$ and $\mathbf{x}_o$. We then combine each verb representation and all object representations to generate the composite HOI representations $\mathbf{x}_h$. After that, we use the composite HOI representations as the input to the verb classifier and obtain the corresponding verb predictions $\hat{\mathbf{Y}}_v \in R^{NN \times N_v}$, where N indicates the number of real HOI instances (*i.e.* , verb-object pair) in each mini-batch and NN is then the number of all composite verb-object pairs (including unknown HOI concepts). Let $\mathbf{Y}_v \in R^{N \times N_v}$ and $\mathbf{Y}_o \in$

$R^{N\times N_o}$ denote the label of verb representations $\mathbf{x}_v$ and object representations $\mathbf{x}_o$, respectively. We then have all composite HOI labels $\mathbf{Y}_h = \mathbf{Y}_v \otimes \mathbf{Y}_o$, where $\mathbf{Y}_h \in R^{NN\times N_v N_o}$, and the superscripts h, v, and o indicate HOI, verb, and object, respectively. Similar to affordance prediction, we repeat $\hat{\mathbf{Y}}_v$ by N_o times to obtain concept predictions $\hat{\mathbf{Y}}_h \in R^{NN\times N_v N_o}$. Finally, we update $\mathbf{M}$ in a running mean manner [29] as follows,

$$\mathbf{M} \leftarrow \frac{\mathbf{M}\odot\mathbf{C} + \sum_i^{NN} \hat{\mathbf{Y}}_h(i,:)\odot\mathbf{Y}_h(i,:)}{\mathbf{C} + \sum_i^{NN}\mathbf{Y}_h(i,:)}, \tag{2}$$

$$\mathbf{C} \leftarrow \mathbf{C} + \sum_i^{NN}\mathbf{Y}_h(i,:), \tag{3}$$

where $\odot$ indicates the element-wise multiplication, $\hat{\mathbf{Y}}_h(i,:)\odot\mathbf{Y}_h(i,:)$ aims to filter out predictions whose labels are not $\mathbf{Y}_h(i,:)$, each value of $\mathbf{C} \in R^{N_v N_o}$ indicates the total number of composite HOI instances in each verb-object pair (including unknown HOI categories). Actually, $\hat{\mathbf{Y}}_h(i,:) \odot \mathbf{Y}_h(i,:)$ follows the affordance prediction process [26]. The normalization with $\mathbf{C}$ is to avoid the model bias to frequent categories. Specifically, both $\mathbf{M}$ and $\mathbf{C}$ are zero-initialized. With the optimization of HOI detection, we can obtain the vector $\mathbf{M}$ to indicate the HOI concept confidence of each combination between verbs and objects.

Self-Training Existing HOI compositional learning approaches [25, 27, 26] usually only consider the known HOI concepts and simply discard the composite HOIs out of label space during optimization. Therefore, there are only positive data for object affordance learning, leaving a large number of unlabeled composite HOIs ignored. Considering that the concept confidence on HOI concept discovery also demonstrates the confidence of affordances (verbs) that can be applied to an object category, we thus try to explore the potential of all composite HOIs, i.e., both labeled and unlabeled composite HOIs, in a semi-supervised way. Inspired by the way used in PU learning [12] and pseudo-label learning [35], we devise a self-training strategy by assigning the pseudo labels to each verb-object combination instance using the concept confidence matrix $\mathbf{M}$, and optimize the network with the pseudo labels in an end-to-end way. With the self-training, the online concept discovery can gradually improve the concept confidence $\mathbf{M}$, and in turn optimize the HOI model for object affordance learning with the concept confidence. Specifically, we construct the pseudo labels $\tilde{\mathbf{Y}}_v \in R^{NN\times N_v}$ from the concept confidence matrix $\mathbf{M} \in R^{N_v\times N_o}$ for composite HOIs $\mathbf{x}_h$ as follows,

$$\tilde{\mathbf{Y}}_v(i,:) = \sum_j^{N_o} \frac{\mathbf{M}(:,j)}{\max(\mathbf{M})} \odot \mathbf{Y}_h(i,:,j), \tag{4}$$

where $0 \le j < N_o$ indicates the index of object category, $0 \le i < NN$ is the index of HOI representations. Here, N is the number of HOIs in each mini-batch, and is usually very small on HICO-DET and V-COCO. Thus the time

complexity of Equation 4 is small. The labels of composite HOIs are reshaped as $\mathbf{Y}_h \in R^{NN \times N_v \times N_o}$. Noticeably, in each label $\mathbf{Y}_h(i,:,:)$, there is only one vector $\mathbf{Y}_h(i,:,j)$ larger than 0 because each HOI has only one object. As a result, we obtain pseudo verb label $\tilde{\mathbf{Y}}_v(i,:)$ for HOI $\mathbf{x}_{h_i}$. Finally, we use composite HOIs with pseudo labels to train the models, and the loss function is defined as follows,

$$\mathcal{L}_d = \frac{1}{NN}\sum_{i}^{NN}(\frac{1}{N_v}\sum_{k}^{N_v}\mathcal{L}_{\text{BCE}}(\frac{\mathbf{Z}(i,k)}{T}, \tilde{\mathbf{Y}}_v(i,k))), \tag{5}$$

where $\mathbf{Z}(i,:)$ is the prediction of the i-th composite HOI, $0 \leq k < N_v$ means the index of predictions, T is the temperature hyper-parameter to smooth the predictions (the default value is 1 in experiment), $\mathcal{L}_{\text{BCE}}$ indicates the binary cross entropy loss. Finally, we optimize the network using $\mathcal{L}_d$, $\mathcal{L}_h$ and $\mathcal{L}_c$ in an end-to-end way, where $\mathcal{L}_h$ indicate the typical classification loss for known HOIs and $\mathcal{L}_c$ is the compositional learning loss [25].

4 Experiments

In this section, we first introduce the datasets and evaluation metrics. We then compare the baseline and the proposed method for HOI concept discovery and object affordance recognition. We also demonstrate the effectiveness of the proposed method for HOI detection with unknown concepts and zero-shot HOI detection. Lastly, we provide some visualizations results of self-compositional learning. Moreover, ablation studies and the full results of HOI detection with self-compositional learning are provided in Appendix D, F, respectively.

4.1 Datasets and Evaluation Metrics

Datasets. We extend two popular HOI detection datasets, HICO-DET [6] and V-COCO [22], to evaluate the performance of different methods for HOI concept discovery. Specifically, we first manually annotate all the possible verb-object combinations on HICO-DET (117 verbs and 80 objects) and V-COCO (24 verbs and 80 objects). As a result, we obtain 1,681 concepts on HICO-DET and 401 concepts on V-COCO, *i.e.* , 1,681 of 9,360 verb-object combinations on HICO-DET and 401 of 1,920 verb-object combinations on V-COCO are reasonable. Besides, 600 of 1,681 HOI concepts on HICO-DET and 222 of 401 HOI concepts on V-COCO are known according to existing annotations. Thus, the HOI concept discovery task requires to discover the other 1,081 concepts on HICO-DET and 179 concepts on V-COCO. See more details about the annotation process, the statistics of annotations, and the novel HOI concepts in Appendix B.

Evaluation Metrics. HOI concept discovery aims to discover all reasonable combinations between verbs and objects according to existing HOI training samples. We report the performance by using the average precision (AP) for concept discovery and mean AP (or mAP) for object affordance recognition. For HOI detection, we also report the performance using mAP. We follow [26]

to evaluate object affordance recognition with HOI model on COCO validation 2017 [39], Object 365 validation [49], HICO-DET test set [6] and Novel Objects from Object 365 [49].

4.2 Implementation Details

We implement the proposed method with TensorFlow [1]. During training, we have two HOI images (randomly selected) in each mini-batch and we follow [17] to augment ground truth boxes via random crop and random shift. We use a modified HOI compositional learning framework, *i.e.* , we directly predict the verb classes and optimize the composite HOIs using SCL. Following [25, 27], the overall loss function is defined as $\mathcal{L} = \lambda_1\mathcal{L}_h + \lambda_2\mathcal{L}_c + \lambda_3\mathcal{L}_d$, where $\lambda_1 = 2$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$ on HICO-DET, and $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$ on V-COCO, respectively. Following [27], we also include a sigmoid loss for verb representation and the loss weight is 0.3 on HICO-DET. For self-training, we remove the composite HOIs when its corresponding concept confidence is 0, *i.e.* , the concept confidence has not been updated. If not stated, the backbone is ResNet-101. The Classifier is a two-layer MLP. We train the model for 3.0M iterations on HICO-DET and 300K iterations on HOI-COCO with an initial learning rate of 0.01. For zero-shot HOI detection, we keep human and objects with the score larger than 0.3 and 0.1 on HICO-DET, respectively. See more ablation studies (*e.g.*, hyper-parameters, modules) in Appendix. Experiments are conducted using a single Tesla V100 GPU (16GB), except for experiments on Qpic [52], which uses four V100 GPUs with PyTorch [44].

4.3 HOI Concept Discovery

Baseline and Methods. We perform experiments to evaluate the effectiveness of our proposed method for HOI concept discovery. For a fair comparison, we build several baselines and methods as follows,

- **Random**: we randomly generate the concept confidence to evaluate the performance.
- **Affordance**: discover concepts via affordance prediction [26] as described in Sec 3.3.
- **GAT** [53]: build a graph attention network to mine the relationship among verbs during HOI detection, and discover concepts via affordance prediction.
- **Qpic*** [52]: convert verb and object predictions of [52] to concept confidence similar as online discovery.
- **Qpic* [52] +SCL**: utilize concept confidence to update verb labels, and optimize the network (Self-Training). Here, we have no composite HOIs.

Please refer to the Appendix for more details, comparisons (*e.g.*, re-training, language embedding), and qualitative discovered concepts with analysis.

Results Comparison. Table 1 shows affordance prediction is capable of HOI concept discovery since affordance transfer learning [26] also transfers affordances to novel objects. Affordance prediction achieves 24.38% mAP on HICO-DET and 21.36% mAP on V-COCO, respectively, significantly better than the

Table 1. The performance of the proposed method for HOI concept discovery. We report all performance using the average precision (AP) (%). SCL means self-compositional learning. SCL− means online concept discovery without self-training.

Method	HICO-DET		V-COCO	
	Unknown (%)	Known (%)	Unknown (%)	Known (%)
Random	12.52	6.56	12.53	13.54
Affordance [26]	24.38	57.92	20.91	95.71
GAT [53]	26.35	76.05	18.35	98.09
Qpic* [52]	27.53	87.68	15.03	13.21
SCL−	22.25	83.04	24.89	96.70
Qpic* [52] + SCL	28.44	88.91	15.48	13.34
SCL	**33.58**	**92.65**	**28.77**	**98.95**

random baseline. With graph attention network, the performance is further improved a bit. Noticeably, [26] completely ignores the possibility of HOI concept discovery via affordance prediction. Due to the strong ability of verb and object prediction, Qpic achieves 27.42% on HICO-DET, better than affordance prediction. However, Qpic has poor performance on V-COCO. The inference process of affordance prediction for concept discovery is time-consuming (over 8 hours with one GPU). Thus we devise an efficient online concept discovery method which directly predicts all concept confidences. Specifically, the online concept discovery method (SCL−) achieves 22.25% mAP on HICO-DET, which is slightly worse than the result of affordance prediction. On V-COCO, the online concept discovery method improves the performance of concept discovery by **3.98%** compared to the affordance prediction. The main reason for the above observation might be due to that V-COCO is a small dataset and the HOI model can easily overfit known concepts on V-COCO. Particularly, SCL significantly improves the performance of HOI concept discovery from 22.36% to **33.58%** on HICO-DET and from 24.89% to **28.77%** on V-COCO, respectively. We find we can also utilize self-training to improve concept discovery on Qpic [52] (ResNet-50) though the improvement is limited, which might be because verbs and objects are entangled with Qpic. Lastly, we meanwhile find SCL largely improves concept discovery of known concepts on both HICO-DET and V-COCO.

4.4 Object Affordance Recognition

Following [26] that has discussed average precision (AP) is more robust for evaluating object affordance, we evaluate object affordance recognition with AP on HICO-DET. Table 2 illustrates SCL largely improves SCL− (without self-training) by **over 9%** on Val2017, Object365, HICO-DET under the same training iterations. SCL requires more iterations to converge, and SCL greatly improves previous methods on all datasets with 3M iterations (Please refer to Appendix for convergence analysis). Noticeably, SCL directly predicts verb rather than HOI categories, and removes the spatial branch. Thus, SCL without self-training (SCL−) is a bit worse than ATL. Previous approaches ignore the un-

Table 2. Comparison of object affordance recognition with HOI network (trained on HICO-DET) among different datasets. Val2017 is the validation 2017 of COCO [39]. Obj365 is the validation of Object365 [49] with only COCO labels. Novel classes are selected from Object365 with non-COCO labels. ATL∗ means ATL optimized with COCO data. Numbers are copied from the appendix in [26]. Unknown affordances indicate we evaluate with our annotated affordances. Previous approaches [25, 26] are usually trained by less 0.8M iterations (Please refer to the released checkpoint in [25, 26]). We thus also illustrate SCL under 0.8M iterations by default. SCL− means SCL without self-training. Results are reported by Mean Average Precision (%).

Method	Known Affordances				Unknown Affordances			
	Val2017	Obj365	HICO	Novel	Val2017	Obj365	HICO	Novel
FCL [27]	25.11	25.21	37.32	6.80	-	-	-	-
VCL [25]	36.74	35.73	43.15	12.05	28.71	27.58	32.76	12.05
ATL [26]	52.01	50.94	59.44	15.64	36.80	34.38	42.00	15.64
ATL∗ [26]	56.05	40.83	57.41	8.52	37.01	30.21	43.29	8.52
SCL−	50.51	43.52	57.29	14.46	44.21	41.37	48.68	14.46
SCL	**59.64**	**52.70**	**67.05**	14.90	**47.68**	**42.05**	**52.95**	14.90
SCL (3M iters)	**72.08**	**57.53**	**82.47**	**18.55**	**56.19**	**46.32**	**64.50**	**18.55**

known affordance recognition. We use the released models of [26] to evaluate the results on novel affordance recognition. Here, affordances of novel classes (annotated by hand [26]) are the same in the two settings. We find SCL improves the performance considerably by **over 10%** on Val2017 and HICO-DET.

4.5 HOI Detection with Unknown Concepts

HOI concept discovery enables zero-shot HOI detection with unknown concepts by first discovering unknown concepts and then performing HOI detection. The experimental results of HOI detection with unknown concepts are shown in Table 3. We follow [25] to evaluate HOI detection with 120 unknown concepts in two settings: rare first selection and non-rare first selection, *i.e.* , we select 120 unknown concepts from head and tail classes respectively. Different from [25, 27] where the existence of unseen categories is known and the HOI samples for unseen categories are composed during optimization, HOI detection with unknown concepts does not know the existence of unseen categories. Therefore, we select top-K concepts according to the confidence score during inference to evaluate the performance of HOI detection with unknown concepts (that is also zero-shot) in the default mode [6].

As shown in Table 3, with more selected unknown concepts according to concept confidence, the proposed approach further improves the performance on unseen categories on both rare first and non-rare first settings. Specifically, it demonstrates a large difference between rare first unknown concepts HOI detection and non-rare first unknown concepts HOI detection in Table 3. Considering that the factors (verbs and objects) of rare-first unknown concepts are rare in the training set [27], the recall is very low and thus degrades the performance on

Table 3. Illustration of HOI detection with unknown concepts and zero-shot HOI detection with SCL. K is the number of selected unknown concepts. HOI detection results are reported by mean average precision (mAP)(%). We also report the recall rate of the unseen categories in the top-K novel concepts. "K = all" indicates the results of selecting all concepts, *i.e.* , common zero-shot. $*$ means we train Qpic [52](ResNet-50) with the released code in zero-shot setting and use the discovered concepts of SCL to evaluate HOI detection with unknown concepts. Un indicates Unknown/Unseen, Kn indicates Known/Seen, while Rec indicates Recall.

Method	K	Rare First				Non-rare First			
		Un	Kn	Full	Rec (%)	Un	Kn	Full	Rec (%)
SCL	0	1.68	22.72	18.52	0.00	5.86	16.70	14.53	0.00
SCL	120	2.26	22.72	18.71	10.83	7.05	16.70	14.77	21.67
SCL	240	3.66	22.72	18.91	15.00	7.17	16.70	14.80	25.00
SCL	360	4.09	22.72	19.00	15.83	7.91	16.70	14.94	30.83
SCL	all	9.64	22.72	19.78	100.00	13.30	16.70	16.02	100.00
Qpic$*$ [52]	0	0.0	**30.47**	24.37	0.00	0.0	23.73	18.98	0.0
Qpic$*$ [52]	120	2.32	30.47	24.84	10.83	14.90	22.19	20.58	21.67
Qpic$*$ [52]	240	3.35	30.47	25.04	15.00	14.90	22.79	21.22	25.00
Qpic$*$ [52]	360	3.72	30.47	25.12	15.83	14.91	23.13	21.48	30.83
Qpic$*$ [52]	all	15.24	30.44	27.40	100.00	21.03	23.73	23.19	100.00
ATL [26]	all	9.18	24.67	21.57	100.00	18.25	18.78	18.67	100.00
FCL [27]	all	13.16	24.23	22.01	100.00	18.66	19.55	19.37	100.00
Qpic + SCL	all	**19.07**	30.39	**28.08**	100.00	**21.73**	**25.00**	**24.34**	100.00

unknown categories. However, with concept discovery, the results with top 120 concepts on unknown categories are improved by relatively **34.52%** (absolutely 0.58%) on rare first unknown concepts setting and by relatively **20.31%** (absolutely 1.19%) on non-rare first setting, respectively. with more concepts, the performance on unknown categories is also increasingly improved.

We also utilize the discovered concept confidences with SCL to evaluate HOI detection with unknown concepts on Qpic [52]. For a fair comparison, we use the same concept confidences to SCL. Without concept discovery, the performance of Qpic [52] degrades to 0 on Unseen categories though Qpic significantly improves zero-shot HOI detection. Lastly, we show zero-shot HOI detection (the unseen categories are known) in Table 3 (Those rows where K is all). We find that SCL significantly improves Qpic, and *forms a new state-of-the-art* on zero-shot setting though we merely use ResNet-50 as backbone in Qpic. We consider SCL improves the detection of rare classes (include unseen categories in rare first and seen categories in non-rare first) via stating the distribution of verb and object. See Appendix D for more analysis, *e.g.*, SCL improves Qpic particularly for rare categories on Full HICO-DET.

4.6 Visualization

Figure 3 illustrates the Grad-CAM under different methods. We find the proposed SCL focus on the details of objects and small objects, while the baseline

and VCL mainly highlight the region of human and the interaction region, *e.g.*, SCL highlights the details of the motorbike, particularly the front-wheel (last row). Besides, SCL also helps the model via emphasizing the learning of small objects (*e.g.*, frisbee and bottle in the last two columns), while previous works ignore the small objects. This demonstrates SCL facilitates affordance recognition and HOI concept discovery via exploring more details of objects. A similar trend can be observed in Appedix G (Qpic+SCL).

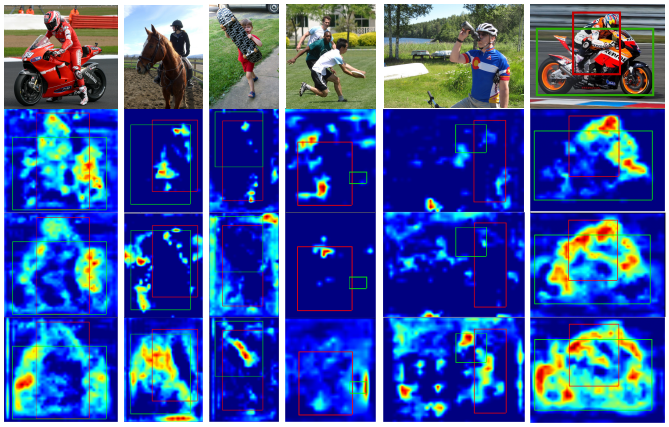

Fig. 3. A visual comparison of recent methods using the Grad-CAM [48] tool. The first row is input image, the second row is baseline without compositional approach, the third row is VCL [25] and the last row is the proposed SCL. We do not compare with ATL [26], since that ATL uses extra training datasets. Here, we compare all models using the same dataset.

5 Conclusion

We propose a novel task, Human-Object Interaction Concept Discovery, which aims to discover all reasonable combinations (*i.e.* , HOI concepts) between verbs and objects according to a few training samples of known HOI concepts/categories. Furthermore, we introduce a self-compositional learning or SCL framework for HOI concept discovery. SCL maintains an online updated concept confidence matrix, and assigns pseudo labels according to the matrix for all composite HOI features, and thus optimize both known and unknown composite HOI features via self-training. SCL facilitates affordance recognition of HOI model and HOI concept discovery via enabling the learning on both known and unknown HOI concepts. Extensive experiments demonstrate SCL improves HOI concept discovery on HICO-DET and V-COCO and object affordance recognition with HOI model, enables HOI detection with unknown concepts, and improves zero-shot HOI detection.

Acknowledgments Mr. Zhi Hou and Dr. Baosheng Yu are supported by ARC FL-170100117, DP-180103424, IC-190100031, and LE-200100049.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th symposium on operating systems design and implementation (OSDI). pp. 265–283 (2016)
2. Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. In: AAAI (2020)
3. Best, J.B.: Cognitive psychology. West Publishing Co (1986)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
5. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 ieee winter conference on applications of computer vision (wacv). pp. 381–389. IEEE (2018)
6. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV. pp. 381–389. IEEE (2018)
7. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: ICCV. pp. 1017–1025 (2015)
8. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating hoi detection as adaptive set prediction. In: CVPR. pp. 9004–9013 (2021)
9. Coren, S.: Sensation and perception. Handbook of psychology pp. 85–108 (2003)
10. Dabral, R., Shimada, S., Jain, A., Theobalt, C., Golyanik, V.: Gravity-aware monocular 3d human-object reconstruction. In: ICCV. pp. 12365–12374 (2021)
11. Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV. pp. 720–736 (2018)
12. De Comité, F., Denis, F., Gilleron, R., Letouzey, F.: Positive and unlabeled examples help learning. In: International Conference on Algorithmic Learning Theory. pp. 219–230. Springer (1999)
13. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1778–1787 (2021)
14. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. pp. 213–220 (2008)
15. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: CVPR (2018)
16. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. IJCV **110**, 259–274 (2014)
17. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. BMVC (2018)
18. Gibson, J.J.: The ecological approach to visual perception (1979)
19. Gibson, J.J.: The ecological approach to visual perception: classic edition. Psychology Press (2014)
20. Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E.: Neural message passing for quantum chemistry. In: ICML. pp. 1263–1272. PMLR (2017)
21. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE PAMI **31**(10), 1775–1789 (2009)

22. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
23. Hassan, M., Dharmaratne, A.: Attribute based affordance detection from human-object interaction images. In: Image and Video Technology. pp. 220–232. Springer (2015)
24. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
25. Hou, Z., Peng, X., Qiao, Y., Tao, D.: Visual compositional learning for human-object interaction detection. In: ECCV (2020)
26. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for human-object interaction detection. In: CVPR (2021)
27. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Detecting human-object interaction via fabricated compositional learning. In: CVPR (2021)
28. Huynh, D., Elhamifar, E.: Interaction compass: Multi-label zero-shot learning of human-object interactions via spatial relations. In: ICCV. pp. 8472–8483 (2021)
29. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456. PMLR (2015)
30. Ji, J., Desai, R., Niebles, J.C.: Detecting human-object relationships in videos. In: ICCV. pp. 8106–8116 (2021)
31. Kato, K., Li, Y., Gupta, A.: Compositional learning for human object interaction. In: ECCV. pp. 234–251 (2018)
32. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: Hotr: End-to-end human-object interaction detection with transformers. In: CVPR. pp. 74–83 (2021)
33. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: NIPS (2014)
34. Kjellström, H., Romero, J., Kragić, D.: Visual object-action recognition: Inferring object affordances from human demonstration. Computer Vision and Image Understanding **115**(1), 81–90 (2011)
35. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML (2013)
36. Li, Y.L., Liu, X., Wu, X., Li, Y., Lu, C.: Hoi analysis: Integrating and decomposing human-object interaction. NeuIPS **33** (2020)
37. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y.F., Lu, C.: Transferable interactiveness prior for human-object interaction detection. In: CVPR (2019)
38. Liao, Y., Liu, S., Wang, F., Chen, Y., Feng, J.: Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: CVPR (2020)
39. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
40. Materzynska, J., Xiao, T., Herzig, R., Xu, H., Wang, X., Darrell, T.: Something-else: Compositional action recognition with spatial-temporal interaction networks. In: CVPR. pp. 1049–1059 (2020)
41. Nagarajan, T., Grauman, K.: Learning affordance landscapes for interaction exploration in 3d environments. Advances in Neural Information Processing Systems **33**, 2005–2015 (2020)
42. Nawhal, M., Zhai, M., Lehrmann, A., Sigal, L., Mori, G.: Generating videos of zero-shot compositions of actions and objects. In: ECCV (2020)
43. Norman, D.A.: The Design of Everyday Things. Basic Books, Inc., USA (2002)

44. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) NeurIPS, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
45. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: ICCV (October 2019)
46. Scott, C., Blanchard, G.: Novelty detection: Unlabeled data definitely help. In: Artificial intelligence and statistics. pp. 464–471. PMLR (2009)
47. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory **11**(3), 363–371 (1965)
48. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
49. Shao, S., Li, Z., Zhang, T., Peng, C., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV (2019)
50. Shen, L., Yeung, S., Hoffman, J., Mori, G., Fei-Fei, L.: Scaling human-object interaction recognition through zero-shot learning. In: WACV. pp. 1568–1576. IEEE (2018)
51. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 (2015)
52. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR (2021)
53. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: ICLR (2017)
54. Wang, D., Cui, P., Zhu, W.: Structural deep network embedding. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge Discovery and Data Mining. pp. 1225–1234 (2016)
55. Wang, S., Yap, K.H., Ding, H., Wu, J., Yuan, J., Tan, Y.P.: Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13475–13484 (2021)
56. Wang, S., Yap, K.H., Yuan, J., Tan, Y.P.: Discovering human interactions with novel objects via zero-shot learning. In: CVPR. pp. 11652–11661 (2020)
57. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: CVPR. pp. 10687–10698 (2020)
58. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5419 (2017)
59. Yang, X., Song, Z., King, I., Xu, Z.: A survey on deep semi-supervised learning. arXiv preprint arXiv:2103.00550 (2021)
60. Yao, B., Ma, J., Li, F.F.: Discovering object functionality. In: ICCV (2013)
61. Zhai, W., Luo, H., Zhang, J., Cao, Y., Tao, D.: One-shot object affordance detection in the wild. arXiv preprint arXiv:2108.03658 (2021)
62. Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage hoi detection. In: Advances in Neural Information Processing Systems. vol. 34 (2021)

63. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: ECCV (2020)
64. Zheng, S., Chen, S., Jin, Q.: Skeleton-based interactive graph network for human object interaction detection. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020)
65. Zhong, X., Ding, C., Qu, X., Tao, D.: Polysemy deciphering network for human-object interaction detection. In: European Conference on Computer Vision. pp. 69–85. Springer (2020)
66. Zhu, Y., Fathi, A., Fei-Fei, L.: Reasoning about object affordances in a knowledge base representation. In: ECCV. pp. 408–424. Springer (2014)
67. Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al.: End-to-end human object interaction detection with hoi transformer. In: CVPR. pp. 11825–11834 (2021)