Affordance Transfer Learning for Human-Object Interaction Detection

Zhi Hou[1], Baosheng Yu[1], Yu Qiao[2,3], Xiaojiang Peng[4], Dacheng Tao[1]

[1] School of Computer Science, Faculty of Engineering, The University of Sydney, Australia
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] Shanghai AI Laboratory
[4] Shenzhen Technology University

`zhou9878@uni.sydney.edu.au, baosheng.yu@sydney.edu.au, yu.qiao@siat.ac.cn,`
`pengxiaojiang@sztu.edu.cn, dacheng.tao@sydney.edu.au`

Abstract

Reasoning the human-object interactions (HOI) is essential for deeper scene understanding, while object affordances (or functionalities) are of great importance for human to discover unseen HOIs with novel objects. Inspired by this, we introduce an affordance transfer learning approach to jointly detect HOIs with novel object and recognize affordances. Specifically, HOI representations can be decoupled into a combination of affordance and object representations, making it possible to compose novel interactions by combining affordance representations and novel object representations from additional images, i.e. transferring the affordance to novel objects. With the proposed affordance transfer learning, the model is also capable of inferring the affordances of novel objects from known affordance representations. The proposed method can thus be used to 1) improve the performance of HOI detection, especially for the HOIs with unseen objects; and 2) infer the affordances of novel objects. Experimental results on two datasets, HICO-DET and HOI-COCO (from V-COCO), demonstrate significant improvements over recent state-of-the-art methods for HOI detection and object affordance detection. Code is available at `https://github.com/zhihou7/HOI-CL`*.*
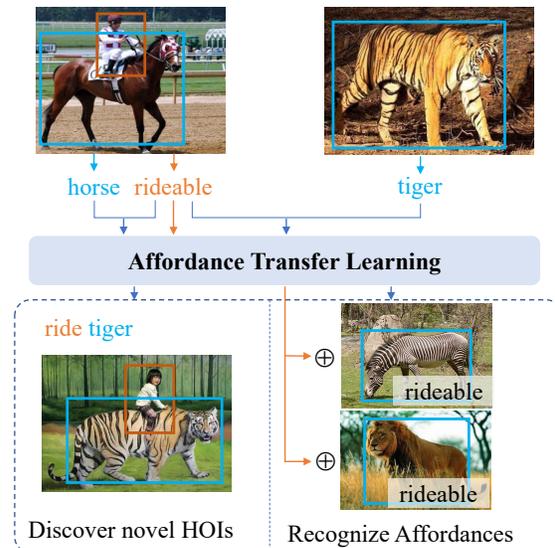
Figure 1. An intuitive example to demonstrate affordance transfer learning for jointly exploring human interactions with novel objects (*e.g.*, "tiger"), and recognizing the affordance of novel objects. The proposed method is able to learn from the unseen interaction samples (*e.g.*, "ride tiger") that are composed from affordance representations and novel object representations, which meanwhile transfers the affordance to novel objects and enables the object affordance recognition.

1. Introduction

Human-object interaction (HOI) detection aims to localize the human and objects in a given image, and recognize the interactions between the human and objects [4]. Considering the combinatorial nature of HOIs, there are always a variety of rare or unseen interactions with novel objects (*e.g.*, "ride tiger"), which remains a great challenge for the HOI detection model to detect unseen interactions with those novel objects.

The interactions between the human and object, $\langle human, verb, object \rangle$, can be captured by either a human-centric (actions) or object-centric (affordance) manner. Specifically, each HOI can be disentangled into a verb and an object, in which the verb also indicates one of the possible affordances (or functionalities) of the object [12, 16], *i.e.* what actions can be applied to a particular object [11]. Therefore, we are able to jointly learn the affordances of object from the HOI samples, making it possible to com-

pose new HOIs by combining the affordance representations in existing HOIs with novel object representations. Meanwhile, the composition of object representations and the corresponding affordance representations transfers the affordance representation (verb) to novel objects, which we term as affordance transfer learning or ATL. The affordance transfer learning empowers the shared affordance representation learning among different objects, and further facilitates the detection on HOIs with novel objects. For example, with the shared affordance representation (*e.g.*, "rideable") between "tiger" and "horse" as illustrated in Figure 1, we are able to compose new HOIs (*i.e.*, "ride tiger"), and thus enable the detection of unseen HOIs. The proposed affordance transfer learning framework further generalizes the compositional learning for HOI detection [18] with the ability to detect HOI with additional unseen objects, rather than only the objects from existing HOI samples.

The proposed affordance transfer learning also empowers the HOI detection model to learn object affordance in a weakly supervised manner. Recent HOI detection approaches [9, 21, 2, 22, 18, 8] usually fail to explore the possibility of object affordance recognition with the HOI detection model, and previous affordance learning methods [16] largely ignore transferring shared affordance representations from existing HOIs to novel objects by HOI detection model. By composing new HOI samples from the affordance representations and novel object representations, affordance transfer learning enables the HOI model to distinguish whether a novel object representation can be combined or not with an affordance representation (*i.e.*, verb). We thus recognize object affordances with HOI model as follows: 1) we maintain a feature bank of decoupled affordance representations from the HOI detection dataset; 2) we extract object representations from additional object detection datasets using the same HOI backbone network; and 3) we combine the object representations with all affordance representations in the feature bank as the input of the HOI classifier. Finally, we are able to obtain a set of HOI predictions, which are further used to infer the object affordances. Overall, the main contribution of this paper can be summarized as follows,

- We introduce an affordance transfer learning framework to exploit a broader source of data for HOI detection, especially for human interactions with novel objects.

- We incorporate HOI detection network with decomposed affordances to infer the affordance of novel objects.

- The proposed method not only improves recent state-of-the-art HOI detection methods but also facilitates the recognition of object affordance at the same time.

2. Related Works

2.1. HOI Understanding

Human-object interaction (HOI) detection are receiving increasing attention from the community [4]. HOI detection aims to not only detect object and human in an image, but also reason the relationships between human and objects. Since Gputa *et al*. [13] presented a Human-Object Interaction approach, massive traditional methods [39, 40] were introduced for HOI recognition using spatial relation [13], pose [39], human part [40] in the early. Recently, Chao *et al*. [5] introduced a large HOI recognition data HICO [5] and a challenging HOI detection dataset HICO-DET [4] for HOI understanding. Meanwhile, Gupta *et al*. [14] introduced the V-COCO dataset, which mainly focuses on the grounding of verbs and their semantic roles.

Currently, there are a large number of HOI detection approaches [9, 21, 38, 35, 31, 2, 27, 34, 32, 42, 19] improving the HOI benchmark [4, 14, 18]. According to the target, current methods can be categoried into two-stage HOI detection, which contains common HOI detection [9, 21, 33, 35, 24] and few-and zero-shot HOI detection [31, 2, 37, 18, 27, 19, 34, 25], and one-stage HOI detection [22, 36, 3]. Recently, Hou *et al*. [18] propose a visual compositional learning framework to compose novel HOI samples between pair-wise HOI images for low-and zero-shot HOI detection. However, VCL [18] can not compose human-novel-object interactions and ignores the possibility of affordance recognition with HOI model. We introduce a novel framework, Affordance Transfer Learning, to transfer the affordance representation to novel objects via composing affordance and novel object representations, and thus enable the detection of HOIs with novel objects.

2.2. Object Affordance

James J. Gibson defined the word affordance in [11]. Object affordances are those action possibilities that are perceiveable by an actor [26, 11, 16], that is also the possibilities of Human-Object Interactions. In the early, Kjellstr *et al*. [20] investigated learning the affordances of objects from human demonstration. Yao & Li *et al*. [41] presented a weak supervised approach to discover object funtionalities from HOI data in the environment where the person is interacting with musical instruments. Fouhey *et al*. [7] introduced an approach to estimate functional surfaces by observing human actions. Recently, Fang *et al*. [6] introduces Demo2Vec to learn interaction region and action label from online video. Differently, we demonstrate to learn a shared affordance representation with HOI model, and recognize object affordance via classifying the composite HOIs of shared affordance representations from existing HOIs and the target object representation.
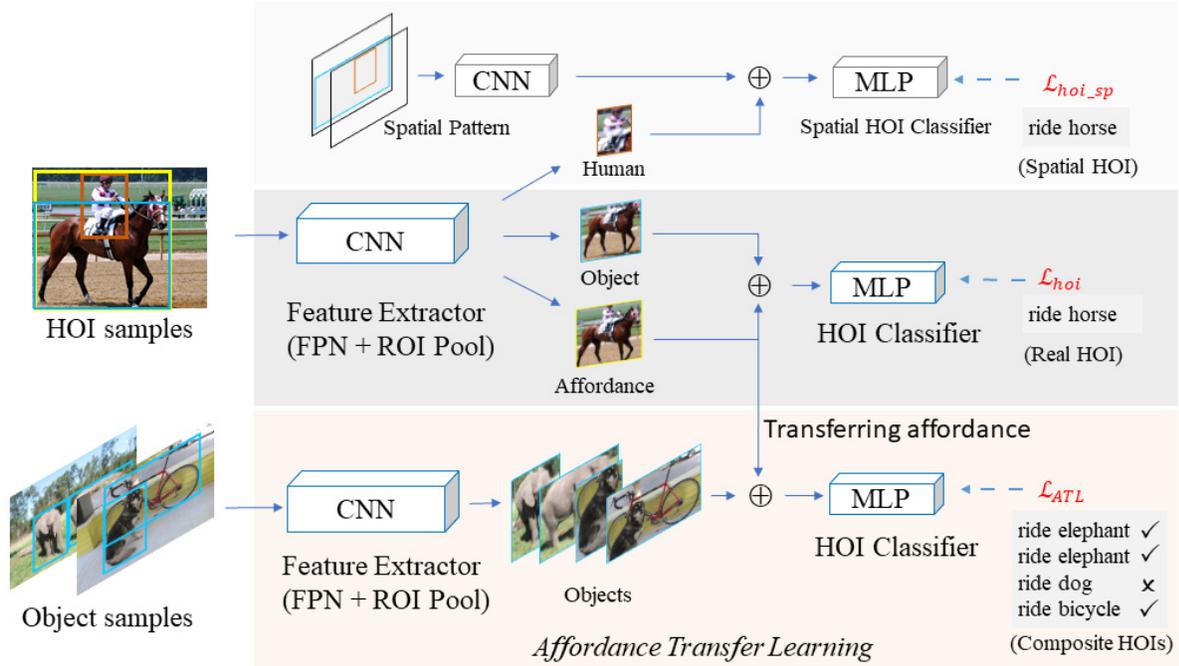
Figure 2. An overview of affordance transfer learning or ATL for HOI detection. We first extract the human, object, and affordance features via the ROI-Pooling from the feature pyramids [28], respectively. Meanwhile, we also extract new object features from an additional object datasets using the same backbone network. After that, we concatenate the affordance and the object features (from HOI datasets) as the real HOIs. We also compose new HOIs using the affordance features and the object features extracted from additional object datasets, which transfer the affordance to novel objects. Both the composite HOIs and real HOIs share the same HOI classifier. In addition, human features and spatial pattern features are combined to construct the spatial HOI branch.

3. Method

In this section, we first give an overview of the proposed method, and then introduce the affordance transfer learning for HOI detection. Lastly, we describe the recognition of object affordances with the proposed HOI detection model.

3.1. Overview

The motivation of affordance transfer learning is to transfer the affordance to novel objects for exploring unseen HOIs, i.e., combining the affordance representations and novel object representations. The affordance transfer learning meanwhile enables HOI network to recognize object affordance. Similar to [9, 38, 15, 32], we utilize the popular two-stage HOI detection framework: 1) we first detect the objects in a given image using a common object detector (e.g., Faster-R-CNN [28]); and 2) we then construct HOIs from object and affordance representations to perform HOI classification. The main framework of our affordance transfer learning is illustrated in Figure 2, which consists of three branches, spatial HOI, real HOI, and composite HOI. Inspired by [9, 21, 18], we utilize the spatial HOI branch to further improve the HOI detection performance. Specifically, the spatial pattern representation consists of two 64×64 binary feature maps to indicate hu-

man and object relative positions, i.e., the pixels within the human (or object) bounding box are assigned the value 1. Both real HOIs and composite HOIs are constructed from object/affordance features and share the same HOI classifier [29], while the difference between them is that the objects in the composite HOIs are extracted from additional object image datasets. Furthermore, by transferring the shared affordance representations extracted from HOI samples to novel objects, we are also able to use the HOI detection model to recognize the affordance of novel objects.

3.2. Affordance Transfer Learning

The proposed affordance transfer learning first composes novel HOI samples between object representations from additional object images (e.g., images from COCO dataset [23]) and decoupled affordance representation as illustrated in Figure 2, and then generalize the affordance representation to novel objects via jointly optimizing the network with the composite HOIs. With the additional objects, the affordance transfer learning effectively decouples the affordance representation from the scenes, and then enables the composition of affordance and novel objects to recognize the affordance of novel objects. In this subsection, we introduce how to efficiently compose new HOIs and remove

invalid HOIs for affordance transfer learning.

Efficient HOI Composition. The label assignment for the composite HOIs can be integrated from the verb label (or the affordance label) and the object label. The object label $\widetilde{l}_o$ is provided by the object datasets, and we obtain the verb label l_v by decoupling the HOI label. Similar to [18], we decouple the HOI label space into a verb-HOI co-occurrence matrix $\mathbf{A}_v \in R^{N_v \times C}$ a the object-HOI co-occurrence matrix $\mathbf{A}_o \in R^{N_o \times C}$, where N_v, N_o, and C indicate the numbers of verbs, objects and HOI categories, respectively. Here, both $\mathbf{A}_v$ and $\mathbf{A}_o$ are binary matrix. Given the one-hot HOI label $y \in R^C$, we then obtain the verb label l_v from $y\mathbf{A}_v^T$. To compose a new HOI by the object $\widetilde{l}_o$ and verb l_v, we assign the label to the composite HOI as follows,

$$\bar{y} = (\widetilde{l}_o \mathbf{A}_o) \& (l_v \mathbf{A}_v), \tag{1}$$

where $\&$ indicates the element-wise "and" logical operation. Both $\widetilde{l}_o \mathbf{A}_o$ and $l_v \mathbf{A}_v$ are binary vectors (*i.e.* HOI labels), which represent all possible HOIs corresponding to $\widetilde{l}_o$ and l_v, respectively. Therefore, the intersection between $\widetilde{l}_o \mathbf{A}_o$ and $l_v \mathbf{A}_v$ indicates the label of the verb-object pair $\left\langle l_v, \widetilde{l}_o \right\rangle$, *i.e.*, the label of the composite HOI.

Invalid HOI Elimination. With additional object datasets, we can compose a large number of new types of HOI samples by combining object and affordance features. Considering a variety of object categories, there is nevertheless some invalid HOIs (*e.g.*, "ride orange"), *i.e.*, the invalid HOIs are out the space of ground truth HOI labels. Furthermore, the same verb might have different meanings in different scenes [10, 42], while the verbs in current HOI dataset (*e.g.*, HICO-DET) mainly represents action (affordance) and are usually not ambiguous [10]. Meanwhile, a variety of recent HOI detection methods do not distinguish the same affordance among different HOIs [31, 2, 15, 19, 37, 27]. To this end, we also equally treat the same affordance from different HOIs for the evaluation of the transfer affordance learning. Following [18], we simply remove those HOIs, which is out of the HOI label space (*e.g.*, "ride dog" in HICO-DET) as illustrated in the right part of Figure 2. In addition, the one-hot labels of those invalid HOIs are all zeros according to Equation 1. That is, we can easily remove those composite HOIs according to the one-hot labels.

3.3. Object Affordance Recognition

In this subsection, we introduce how to infer the object affordance during the testing phase. Considering that we jointly optimize the decoupled components (*i.e.*, object features and affordance features from object and HOI images) in HOI samples and novel object samples with affordance transfer learning, the proposed method thus is able to distinguish whether a novel object is combinable or not with a
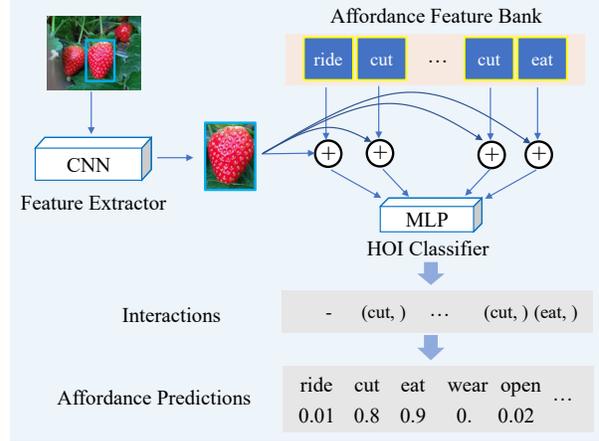

Figure 3. An illustration of object affordance recognition with HOI network. Here, we use verb to represent affordance. We first construct an affordance feature bank from the decoupled affordance representations. For any object (*e.g.* strawberry), we extract the object feature by the Feature Extractor according to bounding box. Then, the object feature is combined with all affordances in the bank to input into HOI classifier for obtaining predicted interactions. The interactions are further converted into affordances (*e.g.* eatable).

specific affordance (*i.e.*, valid HOIs). Therefore, we design a simple yet effective object affordance recognition method using the HOI detection model. Specifically, we first build an affordance feature bank as follows.

Affordance Feature Bank. We construct the affordance feature bank from HOI datasets (*e.g.* HICO-DET and HOI-COCO). In order to reduce storage space and computation, we randomly choose a maximun of M instances for each affordance in HICO-DET. In our experiment, M is 100. Then, we extract the features of those affordances to construct an off-the-shelf affordance feature bank.

Given an object feature extracted from the object image, we combine it with all affordances in the feature bank to obtain a set of HOIs. As illustrated in Figure 3, we obtain all HOI predictions from the HOI classifier. After that, we are able to convert all HOI predictions to affordance predictions according to the HOI-verb co-occurrence matrix $\mathbf{A}_v$. Specifically, we remove the predicted affordances whose label is not the same as the corresponding affordance labels in the feature bank. As a result, we obtain a list of affordances with many repeated elements. Let F_i denotes the frequency (count) of the affordance i and S_i indicates the number of affordance (or verb) i in the feature bank, we evaluate the probability of the affordance i as $\frac{F_i}{S_i}$.

3.4. Optimization and Inference

During the training stage, we train the proposed method with an unique loss L_{ATL} for transferring the affordance to novel objects. Meanwhile, similar to [9], we also incor-

porate spatial pattern loss L_{sp} to optimize the learning of the HOI spatial pattern representation. In addition, an HOI loss L_{hoi} is used for HOI classification. Lastly, the overall training loss function is defined as follows,

$$\mathcal{L} = \mathcal{L}_{hoi\_sp} + \lambda_1 \mathcal{L}_{hoi} + \lambda_2 \mathcal{L}_{ATL}, \qquad (2)$$

where λ_1 and λ_2 are two hyper-parameters to balance different losses. Both the feature extractors and the HOI detection modules are jointly trained in an end-to-end manner. $\mathcal{L}_{hoi\_sp}, \mathcal{L}_{hoi}, \mathcal{L}_{ATL}$ are binary cross entropy losses.

During the testing stage, affordance transfer learning module is not necessary. Similar to [9, 21, 18], we predict the final HOIs with spatial HOI predictions and verb-object HOI predictions. Formally, given a human-object bounding box pair (b_h, b_o), we predict the score $S_{h,o}^c$ as $s_h \cdot s_o \cdot s_{hoi}^c \cdot s_{sp}^c$ for each HOI category $c \in 1, ..., C$, where C denotes the total number of possible HOI types, s_h and s_o are the human and object detection scores respectively, s_{sp}^c represents the spatial HOI prediction score and s_{hoi}^c is the verb-object HOI prediction score.

4. Experiments

We conduct a number of experiments to evaluate the our method for HOI detection using two HOI datasets: HICO-DET [4] and HOI-COCO (built from V-COCO [14]). Furthermore, we also evaluate the HOI detection model with affordance transfer learning for object affordance recognition on COCO dataset [23], HICO-DET dataset [4], and COCO classes and non-COCO classes in Object365 [30].

4.1. Datasets and Evaluation Metrics

HICO-DET [4] dataset consists of 38,118 images in the training set and 9,658 test images over 600 types of interactions (80 object categories in COCO dataset and 117 unique verbs) with over 90,000 HOI instances.

HOI-COCO is built from the V-COCO dataset [14], which contains 10,346 images with 16,199 person instances. Each annotated person in V-COCO has binary labels for 26 different actions. V-COCO mainly focuses on verb recognition, and has limited object categories (only two). Thus we construct a new benchmark HOI-COCO for the evaluation of verb-object pairs as follows. We use 21 actions from all 26 actions in V-COCO (*i.e.*, five non-interaction actions, "walk", "run, "smile", "stand" and "point ", are removed). As a result, we build HOI-COCO benchmark with 222 HOI categories over 21 verbs and 80 objects. Meanwhile, we use the same train/val split in V-COCO for HOI-COCO. Similar to HICO-DET [4], we evaluate the performance on HOI-COCO under three different settings: Full (222 types), Rare (97 types), and NonRare (115 types). The HOI type in Rare category contains less than 10 training instances, and the distribution of HOI categories is long-tailed.

COCO [23] dataset is a widely-used benchmark for common object detection with 80 different object classes. Considering that both HICO-DET [4] and HOI-COCO consist of the same object label sets to COCO, we thus directly incorporate the COCO dataset as the additional object dataset in our experiments.

Object365 [30] is a recently proposed large-scale common object detection dataset with 365 object categories. The domain of Object365 is different from COCO [23]. In detail, we select objects that are labeled as COCO classes from Object365 validation dataset to evaluate the affordance recognition of objects on new domain. Meanwhile, we choose 12 new types of objects and label manually the affordance of those objects according to the HICO-DET and HOI-COCO, respectively. Those objects are used to evaluate affordance recognition on new types of objects. See more details in supplementary materials.

Evaluation Metrics. We follow the standard evaluation metric [9, 38] and report mean average precision for HICO-DET dataset [4] and HOI-COCO. A prediction is a true positive only when the detected human and object bounding boxes have IoUs larger than 0.5 with reference to ground truth, and the HOI category is accurately predicted. Object affordance recognition is a multiple label classification problem (*i.e.* an object usually has multiple affordances). Thus, we compare Precision, Recall and F1-Score for evaluating object affordance recognition.

4.2. Implementation Details

For HICO-DET, similar to recent methods [2, 22, 36, 18], we use the object detector fine-tuned on HICO-DET, *i.e.* the detector provided in [18]. For HOI-COCO, we directly use the object detector pre-trained on COCO. Besides, all HOI classifiers consist of two fully-connected layers with 1024 hidden units. To compare with recent methods on HICO-DET, we use two object images in each mini-batch. On HOI-COCO, we only use one object image for evaluation. Besides, we also use an auxiliary verb loss [17] to improve our baseline. During training, following [9, 21, 18], we augment the ground truth boxes via random crop and random shift. During inference, we keep human and objects with the score larger than 0.3 and 0.1 on HICO-DET respectively. Following [18], we set $\lambda_1 = 2$, $\lambda_2 = 0.5$ on HICO-DET, and $\lambda_1 = 0.5$, $\lambda_2 = 0.5$ on HOI-COCO, respectively. To prevent composite interactions from dominating the training of the model, we keep the number of composite interactions not more than the number of objects in each mini-batch by randomly sampling composite HOIs. We train the model for 1.2M iterations on HICO-DET dataset and 300K iterations on HOI-COCO with an initial learning rare of 0.01. For object affordance recognition, we use the actions of each HOI dataset as affordances and remove the "no interaction" categories on HICO-DET

Table 1. Comparison to recent state-of-the-art methods with fine-tuned detector on HICO-DET dataset [4]. The content in brackets indicates the source of the object images. The last two rows are one-stage HOI detection results.

Method	Default			Known Object		
	Full	Rare	NonRare	Full	Rare	NonRare
FG [2]	21.96	16.43	23.62	-	-	-
IP-Net [36]	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [22]	21.73	13.78	24.10	24.58	16.65	26.84
VCL [18]	23.63	17.21	25.55	25.98	19.12	28.03
DRG [8]	24.53	19.47	26.04	27.98	23.11	29.43
ATL (HICO-DET)	23.67	17.64	25.47	26.01	19.60	27.93
ATL (COCO)	**24.50**	**18.53**	**26.28**	**27.23**	**21.27**	**29.00**
ATL (HICO-DET) ^{DRG}	27.68	20.31	29.89	30.05	22.40	32.34
ATL (COCO) ^{DRG}	**28.53**	**21.64**	**30.59**	**31.18**	**24.15**	**33.29**
Baseline (One-Stage)	22.77	16.54	24.63	26.31	21.60	27.72
ATL (One-Stage)	**23.81**	**17.43**	**25.72**	**27.38**	**22.09**	**28.96**

Table 2. Comparison to recent state-of-the-art methods on HOI-COCO dataset.

Method	object data	Full	Rare	NonRare
Baseline	-	22.86	6.87	35.27
VCL [18]	HOI-COCO	23.53	8.29	35.36
ATL	HOI-COCO	23.40	8.01	35.34
ATL	COCO	24.84	9.79	36.51
ATL	COCO, HICO-DET	**25.29**	**9.85**	**37.27**

Table 3. Comparison of Zero Shot Detection results of our proposed method. UC means unseen composition HOI detection. NO means novel object HOI detection. * means we only use the boxes of the detection results. Here, the baseline means we do not use affordance transfer learning (*i.e.* without L_{ATL}).

Method	Type	Unseen	Seen	Full
Shen *et al.* [31]	UC	5.62	-	6.26
FG [2]	UC	10.93	12.60	12.26
VCL [18] (rare first)	UC	10.06	24.28	21.43
ATL (rare first)	UC	9.18	**24.67**	**21.57**
VCL [18] (non-rare first)	UC	16.22	18.52	18.06
ATL (non-rare first)	UC	**18.25**	**18.78**	**18.67**
FG [2]	NO	11.22	14.36	13.84
Baseline	NO	12.84	20.63	19.33
ATL (HICO-DET)	NO	11.35	20.96	19.36
ATL (COCO)	NO	**15.11**	**21.54**	**20.47**
Baseline*	NO	0.00	14.13	11.77
ATL (HICO-DET)*	NO	0.00	13.67	11.39
ATL (COCO)*	NO	**5.05**	**14.69**	**13.08**

dataset. We keep the object affordance predictions if the affordance score is large than 0.5. All experiments are conducted on a single Tesla V100 GPU with TensorFlow [1].

4.3. HOI Detection

HICO-DET. We report the performance on three different settings: Full (600 categories), Rare (138 categories) and NonRare (462 categories) in "Default" and "Known" modes on HICO-DET. As shown in Table 1, the proposed method outperforms recent state-of-the-art methods among all categories. Furthermore, with better object detection results provided in [8], the performance of ATL dramatically increases to **28.53%**. Meanwhile, we find ATL is more effective on Rare category. Specifically, when using the objects from the training set of HICO-DET, the proposed method is similar to VCL [18] as shown in Table 1. ATL also improves the baseline effectively based on One-Stage method. Here, the baseline is the model without compositional learning. Details of One-Stage method is provided in supplementary materials.

HOI-COCO. We find the proposed method has similar performance to VCL when using HOI-COCO as the source of object images in Table 2. Here, we evaluate the performance of VCL on HOI-COCO dataset using the official code from [18]. When using the COCO object dataset, the proposed method signifcantly improves the performance, especially on Rare categories, *e.g.*, over **1.5%** than VCL and **2.9%** than the baseline, respectively. Meanwhile, the proposed method also gives a larger improvement than baseline in NonRare category comparing with VCL, suggesting that ATL also increases the diversity of HOIs via composing new samples. Furthermore, when using both HICO-DET and COCO to provide object images, we further improve the performance to **25.29%**.

4.4. Zero-Shot HOI detection

The proposed affordance transfer learning enables the detection of HOIs with novel objects due to the mecha-

nism of composing HOI samples of unseen classes. Therefore, we evaluate the proposed method for zero-shot HOI detection on HICO-DET [4]. We report the performance on two settings: 1) Unseen Composition and 2) Novel Object. Specifically, Unseen Composition means there are unseen HOIs in the test but the verbs and objects of the unseen HOIs exist in training data, while the objects of unseen HOIs in novel object HOI detection do not exist in training data. For compositional zero-shot learning, we follow [18] to evaluate on rare-first unseen HOIs (firstly select tail HOIs in HICO-DET as unseen data) and non-rare first unseen HOIs (firstly select head HOIs in HICO-DET as unseen data). We evaluate zero-shot HOI detection on three categories: Unseen (120 categories), Seen (480 categories) and Full (600 categories). For novel object HOI detection, similar to [2], we choose 100 unseen categories (includes 12 unseen objects) and 500 seen categories. We choose the object detector provided in [18] to compare fairly with [18].

Compositional Zero-Shot HOI Detection. In Table 3, we find our approach effectively improves the non-rare first zero-shot HOI detection. Meanwhile, our approach achieves better result on seen category in rare first zero-shot HOI detection. Particularly, the affordances in tail part of HOIs are usually rare, the composite samples of tail HOIs with additional objects are much less than that of head

Table 4. Comparison of object affordance recognition with HOI network among different datasets (based on Mean average Precision). Val2017 is the validation 2017 of COCO [23]. Subset of Object365 is the validation of Object365 [30] with only COCO labels. Novel classes are selected from Object365 with non-COCO labels. Object means what object dataset we use. ATL^{ZS} means novel object zero-shot HOI detection model in Table 3 on HICO-DET. For ATL^{ZS}, we show the results of the 12 classes of novel objects in Val2017, Subset of Object365 and HICO-DET.

Method	HOI Data	Object	Val2017 of COCO			Subset of Object365			HICO-DET			Novel classes		
			Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1	Rec	Prec	F1
Baseline	HOI	-	28.62	32.34	27.08	21.75	22.20	19.83	36.64	49.83	37.67	12.39	8.63	9.62
VCL [18]	HOI	HOI	76.93	71.79	72.15	68.60	67.52	65.82	87.98	82.59	83.84	54.75	35.85	40.43
ATL	HOI	HOI	80.71	72.79	74.44	71.76	67.34	67.13	90.29	83.21	85.30	**58.73**	37.75	42.75
ATL	HOI	COCO	90.94	87.33	87.65	82.95	82.13	80.80	93.35	90.77	91.02	53.65	40.94	43.57
Baseline	HICO	-	8.11	29.21	11.81	6.77	26.1	9.97	8.11	29.21	15.55	8.12	15.87	8.78
VCL [18]	HICO	HICO	9.63	43.62	14.89	10.66	38.69	15.77	10.76	53.54	16.82	7.81	22.63	11.02
ATL	HICO	HICO	14.01	46.45	20.24	17.71	50.92	24.61	15.54	52.25	22.54	**12.78**	**28.8**	**16.78**
ATL	HICO	COCO	33.69	79.54	44.32	28.25	63.56	35.24	30.27	73.53	40.31	12.41	14.56	12.86
ATL^{ZS}	HICO	HICO	4.28	22.96	6.98	3.54	19.35	5.8	6.02	32.22	9.93	5.02	11.63	6.79
ATL^{ZS}	HICO	COCO	**19.41**	**66.70**	**29.01**	**15.57**	**55.58**	**23.49**	**19.36**	**67.55**	**28.81**	**14.00**	**28.60**	**18.07**

HOIs. Therefore, our approach achieves even worse result on unseen category.

Novel Object HOI Detection. Table 3 demonstrates that transferring affordance representation to novel objects effectively facilitates the detection of unseen HOIs with novel objects. Here we use the network without affordance transfer learning as our baseline. We find using HICO-DET (remove HOIs with unseen objects) as object images even degrades the performance on unseen categories compared to the baseline because we compose massive seen HOI samples but not unseen HOI samples with HICO-DET. Besides, similar to [2], we use an generic object detector to enable HOI detection with novel object, which provides a strong baseline. While we only use the boxes of the detector (not use the object label predicted by detector), the performances of baseline and ATL (HICO-DET) on unseen category decrease to 0. However, ATL (COCO) still achieves 5.05% on unseen category.

4.5. Object Affordance Recognition

Table 4 shows ATL significantly improves the baseline by **over 40 %** in F1 score among all datasets with COCO categories, and by **over 30%** on novel object classes on HOI-COCO. On HICO-DET, ATL improves the baseline by **nearly 10%** among all categories in datasets with COCO categories, and by **around 5%** on novel object classes. Those experiments indicate that the affordance transfer learning via composing novel HOIs effectively disentangles the affordance and object representations from the scenes and endows the HOI network with the ability of affordance recognition. Noticeably, ATL^{ZS} (COCO), that composing interactions from affordances and novel objects, largely improves the baseline model ATL^{ZS} (HICO) on the 12 novel classes among all evaluation datasets.

With the object images from the HOI dataset, our method

Table 5. Illustration of the number of object images in each batch on HICO-DET dataset.

#Images	Full	Rare	NonRare
1	24.07	18.17	25.83
2	**24.50**	**18.53**	**26.28**
3	24.19	17.33	26.24

Table 6. Illustration of the effect of different object detectors on HOI detection in HICO-DET. Fine-tuned detector is provided in [18]. GT means ground truth boxes. The last column is the detection mAP on HICO-DET test dataset.

Model	Detector	Full	Rare	NonRare	mAP
Baseline	COCO	**21.07**	**16.79**	**22.35**	20.82
ATL	COCO	20.08	15.57	21.43	20.82
Baseline	Fine-tuned	23.44	16.80	25.43	30.79
ATL	Fine-tuned	**24.50**	**18.53**	**26.28**	30.79
Baseline	GT	43.32	33.84	46.15	100
ATL	GT	**44.27**	**35.52**	**46.89**	100

is similar to VCL [18] on HOI-COCO dataset, because both two methods compose HOI samples between two images. On HICO-DET, the proposed method has a better performance than VCL [18], while the two methods have similar performance on HOI detection, as shown in Table 1. We find Recall, Precision and F1-Score are sensitive to the threshold. Thus, we further illustrates the Mean average Precision results of all models in Supplementary Materials.

4.6. Ablation Studies

The number of object images in each batch. Table 5 shows ATL achieves best performance with 2 object images. We think more object images increase the diversity of object features and balance the object distribution. However, too much object images also hampers the performance.

Object detector. Due to the domain shift between

Table 7. Illustration of effect of domain shift on ATL between object images and HOI images on HOI-COCO dataset. Sub-COCO is a subset of COCO images that we randomly choose the same number of object instances to the objects of HICO-DET from COCO dataset.

Method	Object images	Full	Rare	NonRare
ATL	HICO-DET	24.21	9.52	35.61
ATL	Sub-COCO	**24.74**	**9.60**	**36.50**

Figure 4. Comparison of object affordance recognition (F1) between ATL and the conversion from object detection results on HICO-DET. Confidence is the object detection confidence for choosing object boxes. Red is our method and Blue is the conversion from object detection results.

HICO-DET and COCO, COCO detector usually achieves worse result. we thus use the same fine-tuned object detector as [18]. Table 6 illustrates better detected object boxes improves the performance largely. Meanwhile, we find ATL is apparently sensitive to worse boxes. Under worse object detector (*i.e.* COCO detector), ATL does not improve the result. It might be because composing affordance features and object features from additional images results in poor generalization to worse boxes. When we transfer affordance representation to objects from a large number of additional images via composing novel HOI samples, we improve the scene generalization (*i.e.* the model generalizes to novel scenes) of the affordance representation learning, while degrading the generalization to worse object boxes on HICO-DET test set. The object affordance recognition in Table 4 illustrates the scene generalization of affordance and object representations. Noticeably, worse object detector largely hampers HOI detection in two-stage method. Thus, it is necessary to utilize better object detector for evaluating HOI detection, and ATL further improves HOI detection effectively with better object detector.

Domain difference. From the large performance gap between different object detectors in Table 6, we find the HICO-DET dataset has a different domain to COCO. Table 7 shows with the same number of object instances, COCO dataset improves the performance larger than HICO-DET dataset due to the domain difference on HOI-COCO. There is a similar trend in Table 1 and Table 2. With the same COCO dataset, our method facilitates HOI detection on HOI-COCO dataset better than that on HICO-DET.
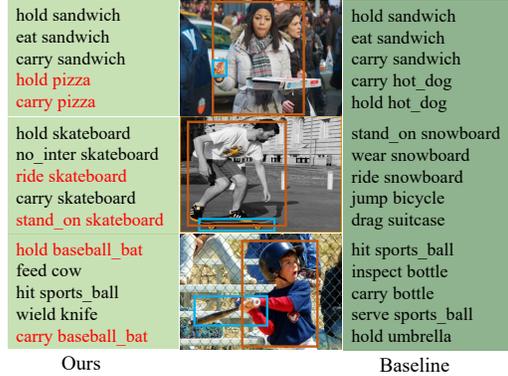

Figure 5. Illustration of unseen object zero-shot detection result (top 5) between the proposed method and Baseline. The correct results are highlighted in red.

Affordance comparison with object detection results. Our method can also be applied to detected boxes of an object detector. For a robust comparison, we directly compare ATL with the object affordance result converted from object detection results according to the object affordance annotation (*i.e.* the ground truth affordances of an object category) on HICO-DET test set. Here we use the detected box of a COCO pretrained Faster-RCNN. We train our model on HOI-COCO dataset and COCO (2014) dataset, which has a same training set to COCO pretrained Faster-RCNN. Figure 4 illustrates ATL achieves better affordance recognition results among different confidences. Meanwhile, ATL has better performance than object affordance detection when the confidence of detected box is lower.

4.7. Qualitative Results

We demonstrate the result of exploring unseen HOIs with novel objects in Figure 5. We find the baseline can not recognize the object at all, while the proposed method effectively detects the HOI with unseen objects.

5. Conclusion

In this paper, we introduce a novel approach, affordance transfer learning or ATL, to transfer the affordance to novel objects via composing objects (from object images) and affordances (from HOI images) for HOI detection. ATL effectively facilitates HOI detection in long-tailed settings, especially for HOIs with novel objects. In addition, we devise a simple yet effective method to incorporate HOI detection model for object affordance recognition and ATL significantly improves the performance of the HOI detection model for object affordance recognition.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th symposium on operating systems design and implementation (OSDI)*, pages 265–283, 2016. 6

[2] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. *AAAI*, 2020. 2, 4, 5, 6, 7

[3] Jaewoo Kang Hyunwoo J. Kim Bumsoo Kim, Taeho Choi. Uniondet: union-level detector towards real-time human-object interaction detection. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 2

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 1, 2, 5, 6

[5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015. 2

[6] Kuan Fang, Te Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018. 2

[7] David F. Fouhey, Vincent Delaitre, A. Gupta, Alexei A. Efros, I. Laptev, and Josef Sivic. People watching: Human actions as a cue for single view geometry. *International Journal of Computer Vision*, 110:259–274, 2014. 2

[8] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 2, 6

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *BMVC*, 2018. 2, 3, 4, 5

[10] Spandana Gella, Frank Keller, and Mirella Lapata. Disambiguating visual verbs. *IEEE PAMI*, pages 1–1, 2017. 4

[11] James J. Gibson. The ecological approach to visual perception. 1979. 1, 2

[12] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, pages 8359–8367, 2018. 1

[13] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE PAMI*, 31(10):1775–1789, 2009. 2

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

[15] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9677–9685, 2019. 3, 4

[16] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey, 2018. 1, 2

[17] Zhi Hou, Yu Baosheng, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 5

[18] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 4, 5, 6, 7, 8

[19] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 4

[20] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011. 2

[21] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yan-Feng Wang, and Cewu Lu. Transferable interactiveness prior for human-object interaction detection. *CVPR*, 2018. 2, 3, 5

[22] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. *CVPR*, 2020. 2, 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 5, 7

[24] Y Liu, Q Chen, and A Zisserman. Amplifying key cues for human-object-interaction detection. *Lecture Notes in Computer Science*. 2

[25] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACMMM*, 2020. 2

[26] Donald A. Norman. *The Design of Everyday Things*. Basic Books, Inc., USA, 2002. 2

[27] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 4

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, pages 91–99, 2015. 3

[29] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR*, pages 1745–1752. IEEE, 2011. 3

[30] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 5, 7

[31] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1568–1576. IEEE, 2018. 2, 4, 6

[32] Oytun Ulutan, ASM Iftekhar, and BS Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. *CVPR*, 2020. 2, 3

[33] 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)*, pages 9469–9478, 2019. 2

[34] Suchen Wang, Kim-Hui Yap, Junsong Yuan, and Yap-Peng Tan. Discovering human interactions with novel objects via zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11652–11661, 2020. 2

[35] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 5694–5702, 2019. 2

[36] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. *CVPR*, 2020. 2, 5, 6

[37] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 22(6):1423–1432, 2019. 2, 4

[38] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 2, 3, 5

[39] Bangpeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE PAMI*, 34(9):1691–1703, 2012. 2

[40] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*, pages 1331–1338. IEEE, 2011. 2

[41] Bangpeng Yao, Jiayuan Ma, and Fei Fei Li. Discovering object functionality. In *IEEE International Conference on Computer Vision*, 2013. 2

[42] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human-object interaction detection. In *ECCV*, 2020. 2, 4