Discovering Human-Object Interaction Concepts via Self-Compositional Learning (Appendix)

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. In this appendix, we provide detailed analysis in Section A, and illustrate the annotation in Section B. Quantitative illustration of discovered concepts are provided in Section C. More experimental results (including ablation studies) in Section D and Section E. We also demonstrate the performance of HOI detection (*e.g.*, Qpic, Table 11 and Table 12) in Section F, the visualization results on V-COCO in Section G, and additional concept discovery approaches (e.g., language embedding, Re-Training) in Section H. Extensive experiments demonstrate 1) different hyper-parameters has an obvious impact on concept discovery; 2) self-training concept discovery achieves a large relative improvement by **over 190%** on novel object unknown concept detection; 3) self-training effectively improves two lines of HOI detection method, particularly for rare category.

A Detailed Analysis for the Motivation

Actually, after we generate the composite HOI features, we have features for both known and unknown concepts. We merely know the HOI features of the known concepts are existing, while we do not know whether the HOI features of unknown concepts are reasonable or not. This actually fall into a typical semi-supervised learning, in which part of samples are labeled (known). Therefore, inspired by the popular semi-supervised learning method, we propose to design a self-training strategy with pseudo labels.

SCL largely improves concept discovery. At first, during training, SCL involves both HOI instances from known or unknown concepts (via pseudo-labeling). Another important thing is that SCL uses both positive and negative unknown concepts, which prevents the model from only fitting the verb patterns. For example, the classifier may predict a reasonable concept for the verb "eat" regardless of the object representation, if there are no negative unknown concepts, *e.g.*, "eat TV". Lastly, as shown in Figure 1, SCL also reduces the risk of overfitting known concepts compared with ATL. *e.g.*, we observe high confidence for the novel concept "squeeze banana"(sort in 2027) in SCL, while the confidence of "squeeze banana" is merely 0.0017 (sort in 7554) in ATL.

Table 1. The performance of the proposed method for HOI concept discovery under different annotations. Better Annotation indicates we remove some wrongly labeled concepts in annotation. We report all performance using the average precision (AP) (%). UC means unknown concepts and KC means known concepts. SCL means self-compositional learning. SCL− means online concept discovery without self-training.

Method	Better Annotation	UC	KC
SCL−		22.36	83.04
SCL		33.26	93.06
SCL−	✓	22.25	83.04
SCL	✓	**33.58**	**92.65**

B Annotation

In order to evaluate the proposed method, we manually annotate the novel concepts for both HICO and V-COCO dataset. Specifically, we annotate the concepts that people can infer from existing concepts. The final set of concepts are provided in the supplemental material.

Statistically, there are about 1.3% and 1.9% mislabeled pairs on HICO-DET and V-COCO, respectively. Meanwhile, there are about 1.7% and 1.1% unlabeled pairs (including ambiguous verbs) on the remaining categories of HICO-DET and V-COCO.

To evaluate the effect of annotation quality of concept annotation on HOI concept discovery, we illustrate the result of different models with different annotations. We compare two versions of annotations, both of which are provided in supplemental materials. Specifically, the file "label_hoi_concept.csv" is the worse version, while "label_hoi_concept_new.csv" is the refined version. Table 1 shows SCL even achieves better performance when evaluate SCL with better annotation, while the performance of baseline is not improved. This experiments together with Table 1 in the main paper show the quality of current annotation is enough for the evaluation of the proposed method.

C Qualitative illustration

We also illustrate the discover concepts in this Section. Here, we choose the concepts after removing the known concepts from the prediction list because the confidence of known concepts in the prediction of SCL is usually very higher. We choose 5 concepts with high confidence and 5 concepts with low confidence to illustrate. Table 2 shows the discovered concepts in SCL are usually more reasonable. We provide the full prediction list with confidence in "result_conf_SCL.txt" in supplementary materials.

Table 2. The illustration of discovered concepts.

Method	Concepts with high confidence	Concepts with low confidence
SCL−	type_on sink,inspect refrigerator,feed suitcase, inspect chair,carry stop_sign	zip zebra, sign dog, chase broccoli, set parking_meter, tag teddy_bear
SCL	ride bear, board truck, carry bowl, wash fire_hydrant, hop_on motorcycle	zip zebra, flush parking_meter, stop_at hair_drier, stop_at microwave

Table 3. Ablation studies of different modules on HICO-DET. UC means unknown concepts and KC means known concepts. Verb aux loss means Verb auxiliary loss (*i.e.* , binary cross entropy loss). Results are reported by average precision (%).

Spatial branch	Verb aux loss	Union Verb	UC	KC
✓	✓	✓	32.56	94.39
-	✓	✓	**33.26**	93.06
✓	-	✓	29.56	93.36
✓	✓	-	28.30	94.27

D Ablation Studies

D.1 Modules

We conduct ablation studies on three modules: verb auxiliary loss [7], union verb [5], and spatial branch [4]. Union verb indicates that we extract verb representation from the union box of human and object. When we remove the union verb representation, we directly extract verb representation from the human bounding box; In our experiment, we remove the spatial branch. Here, we demonstrate we achieve better performance without the spatial branch.

Spatial branch. We remove the spatial branch in [4], which is very effective for HOI detection. We find that the spatial branch degrades the performance of HOI concept discovery: the performance of HOI concept discovery increases from 32.56% to 33.26% without spatial branch, as shown in Table 3. We thus remove spatial branch.

Verb auxiliary loss. We follow [7] to utilize a verb auxiliary loss to regularize verb representations. As shown in Table 3, the model without using a verb auxiliary loss drops by nearly 3% on unseen concepts, which demonstrates the importance of verb auxiliary loss for HOI concept discovery.

Union verb. Table 3 demonstrates that extracting verb representation from union box is of great importance for HOI concept discovery. When we extract verb representation from human bounding box, the result of HOI concept discovery apparently drops from 32.56% to 28.30%.

Though verb auxiliary loss and union verb representation are very helpful for concept discovery, the performance without the two strategies still outperform our baseline, *i.e.* , online concept discovery without self-training.

D.2 Convergence Analysis

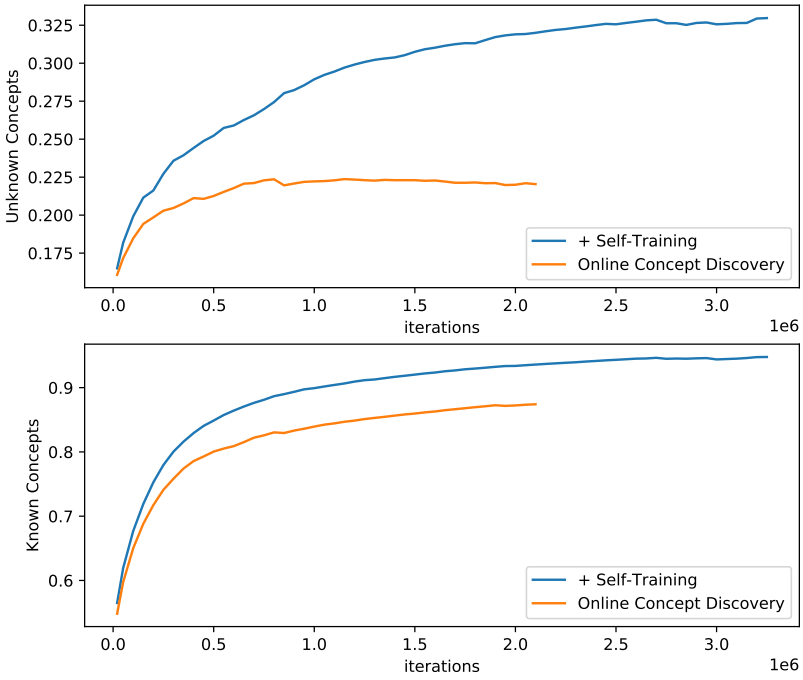

Fig. 1. Illustration of the convergence with self-training strategy.

To some extent, the self-training approach makes use of all composite HOIs, and thus significantly enriches the training data. As a result, the self-training strategy usually requires more iterations to converge to a better result. Figure 1 illustrates the comparison of convergence between online concept discovery and self-training. For online concept discovery, we observe that the model begins to overfit the known concepts after 2,000,000 iterations, and we thus have an early stop during the optimization. We notice that the result on unknown concepts of self-training increases to 32.%, while the baseline (*i.e.* , online concept discovery) begins to overfit after 800,000 iterations. This might be because the self-training utilizes all composite HOIs including many impossible combinations (*i.e.* , negative samples for HOI concept discovery).

Table 4. Ablation studies of hyper-parameters on V-COCO. UC means unknown concepts and KC means known concepts. Results are reported by average precision (%).

λ_3	0.5	0.5	0.5	0.25	1.	2.	4.
T	1	2	0.5	1.	1.	1.	1.
UC (%)	29.52	28.60	29.69	28.06	29.94	**31.33**	29.78
KC (%)	97.57	96.76	97.57	95.32	97.87	97.81	97.94

Table 5. Ablation studies of hyper-parameter T on HICO-DET. Here, we run all experiments with only 1,000,000 iterations and remove the spatial branch to evaluate T. UC means unknown concepts and KC means known concepts. Results are reported by average precision (%).

T	2	1	0.5	0.25	0.125
UC (%)	27.15	30.36	33.54	**33.66**	33.25
KC (%)	85.53	88.72	91.71	93.62	**94.32**

D.3 Hyper-parameters

In the main paper, we have several hyper-parameters (*i.e.* λ_1, λ_2, λ_3, T, where $\lambda_1 = 2.$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$ and $T = 1.$). For λ_1 and λ_2, we follow the settings in [5]. For λ_3 and T, we perform ablation studies on V-COCO as shown in Table 4. We notice that both T and λ_3 have an important effect on the HOI concept discovery. As shown in Table 4, the performance increases from 29.52% to **31.33%** on unseen concepts when we set $\lambda_3 = 2.$, which is much better than the results reported in the main paper. This also illustrates that $\mathcal{L}_d$ is more important than $\mathcal{L}_{CL}$ for HOI concept discovery.

In our experiment, we apply the temperature T to predictions. As shown in Table 4, we find that when T decreases to 0.5, the performance also slightly increases from 29.52% to 29.69%. Thus, we further conduct ablation experiments on T in Table 5. Specifically, to quickly evaluate the effect of T, we remove spatial branch and run all experiments with 1,000,000 iterations. Noticeably, when we set $T = 0.25$, the performance on concept discovery further increases from 30.36% to **33.66%**, which indicates a smaller temperature helps HOI concept discovery. In our experiments, we also find this result further increases to over 35.% when $T = 0.5$ after convergence, which is much better than the result (33.26%) of $T = 1$. This might be because smaller temperature is less sensitive to noise data, since composite HOIs can be regard as noise data.

D.4 Normalization for Pseudo-labels

In our experiment, we normalize the confidence matrix for pseudo-labels. Table 6 illustrates the normalization approach has a slight effect on the concept discovery performance.

Table 6. Illustration of normalized pseudo labels on HICO-DET and V-COCO. Experiments results are reported by average precision (%). Here, the SCL model uses spatial branch.

Method	HICO-DET		V-COCO	
	UC (%)	KC (%)	UC (%)	KC (%)
SCL	**32.56**	**94.39**	**29.52**	**97.57**
w/o normalization	32.30	94.2	29.32	97.93

Table 7. 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 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 [11](ResNet-50) with the released code in zero-shot setting and use the discovered concepts of SCL to evaluate HOI detection with unknown concepts.

Method	K	Rare First				Non-rare First			
		Unknown	Known	Full	Recall (%)	Unknown	Known	Full	Recall (%)
Baseline	0	1.68	22.10	18.52	0.00	5.86	16.30	14.21	0.00
Baseline	120	3.06	22.10	18.29	10.83	6.16	16.30	14.27	21.67
Baseline	240	3.28	22.10	18.34	13.33	6.90	16.30	14.42	25.00
Baseline	360	3.86	22.10	18.45	15.83	7.29	16.30	14.50	30.83
Baseline	all	9.62	22.10	19.61	100.00	12.82	16.30	15.60	100.00
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

E HOI Detection with Unknown Concepts

E.1 Additional Comparisons

Table 7 demonstrates SCL consistent improves the baseline (*i.e.* , SCL without Self-Training). Here, we use the same concepts for a fair comparison. Thus, the recall is the same. Meanwhile, Table 7 also shows Self-Training effectively improves the HOI detection. when we select all concepts to evaluate HOI detection, it is common zero-shot HOI detection, *i.e.* , all unseen classes are known. Particularly, for application, one can directly detect unknown concepts with concept discovery from the model itself, *e.g.*, Qpic [11]. Here, we mainly demonstrate different methods with the same concept confidence for a fair comparison.

E.2 Novel Objects

In the main paper, we illustrate the result on two compositional zero-shot settings. Here, we further illustrate the effectiveness of HOI concept discovery for novel object HOI detection. Novel object HOI detection requires to detect HOI

Table 8. Illustration of the effectiveness of HOI concept discovery for HOI detection with unknown concepts (novel objects). K is the number of selected unknown concepts. HOI detection results are reported by mean average precision (mAP)(%). Recall is evaluated for the unseen categories under the top-k novel concepts. The last row indicates the results of selecting all concepts.

K	Unseen	Seen	Full	Recall (%)
0	3.92	19.45	16.86	0.00
100	11.41	19.45	18.11	41.00
200	12.40	19.45	18.28	48.00
300	13.52	19.45	18.46	52.00
400	13.52	19.45	18.46	52.00
500	13.91	19.45	18.53	56.00
600	13.91	19.45	18.53	56.00
all	17.19	19.45	19.07	100.00

Table 9. Additional Comparison on HOI concept discovery. We report all performance using the average precision (AP) (%). UC means unknown concepts and KC means known concepts. SCL means self-compositional learning. SCL− means online concept discovery without self-training. SCL (COCO) means we train the network via composing between verbs from HICO and objects from COCO 2014 training set.

Method	HICO-DET		V-COCO	
	UC (%)	KC (%)	UC (%)	KC (%)
Random	12.52	6.56	12.53	13.54
language embedding	16.08	29.64	-	-
Re-Training	26.09	50.32	-	-
SCL− (COCO)	17.01	55.50	26.04	81.47
SCL (COCO)	31.92	86.43	27.90	90.04
SCL−	22.36	83.04	26.64	95.59
SCL	**33.26**	**93.06**	**29.52**	**97.57**

with novel objects, *i.e.* , the object of an unseen HOI is never seen in the HOI training set. We follow [6] to select 100 categories as unknown concepts. The remaining categories do not include the objects of unseen categories. Here we use a unique object detector to detect objects. To enable the novel object HOI detection and novel object HOI concept discovery, we follow [6] to incorporate external objects (*e.g.*COCO [9]) to compose novel object HOI samples. Specifically, we only choose the novel types of objects from COCO [9] as objects images in the framework [6] for novel object HOI detection with unknown concepts.

Table 8 demonstrates concept discovery largely improves the performance on unseen category from 3.92% to **11.41%** (relatively by 191%) with top 100 unknown concepts. We meanwhile find the recall increases to 41.00% with only the top 100 unknown concepts. Nevertheless, when we select all unknown concepts, the performance on unseen category is 17.19%. This shows we should improve the performance of concept discovery.

Table 10. Illustration of the effectiveness of self-training on HOI detection based on ground truth box. Results are reported by mean average precision (%).

Method	Full	Rare	NonRare
SCL	**42.92**	**36.60**	**44.81**
w/o Self-Training	42.66	35.81	44.70

Table 11. Illustration of the effectiveness of self-training for Qpic (ResNet-50). Results are reported by mean average precision (%). $*$ means we use the released code to reproduce the results for a fair comparison. S1 means Scenario 1, while S2 means Scenario 2.

Method	HICO-DET			V-COCO	
	Full	Rare	NonRare	S1	S2
GGNet [12]	23.47	16.48	25.60	-	54.7
ATL [6]	23.81	17.43	25.72	-	-
HOTR [8]	25.10	17.34	27.42	55.2	**64.4**
AS-Net[2]	28.87	24.25	30.25	-	53.9
Qpic [11]	29.07	21.85	31.23	58.8	61.0
Qpic* [11]	29.19	23.01	31.04	61.29	62.10
Qpic + SCL	**29.75**	**24.78**	**31.23**	**61.55**	**62.38**

F HOI Detection

One-Stage Method. We also evaluate SCL on Qpic [11], *i.e.* , the state-of-the-art HOI detection method based on Transformer, for HOI detection. Code is provided in https://github.com/zhihou7/SCL. We first obtain concept confidence similar as Section 3.3.2 in the main paper. Denote $\hat{\mathbf{Y}}_v \in R^{N \times N_v}$ as verb predictions, $\hat{\mathbf{Y}}_o \in R^{N \times N_o}$ as verb predictions, we obtain concept predictions $\hat{\mathbf{Y}}_h$ as follows,

$$\hat{\mathbf{Y}}_h = \hat{\mathbf{Y}}_v \otimes \hat{\mathbf{Y}}_o. \tag{1}$$

Then, we update M according to Equation 2 and Equation 3 in the main paper. After training, we evaluate HOI concept discovery with M.

For self-training on Qpic [11], we use M to update the verb label $\mathbf{Y}_v \in R^{N \times N_v}$ for annotated HOIs. Here, we do not have composite HOIs because Qpic has entangled verb and object predictions, and we update verb labels with $\mathbf{M}$. Specifically, given an HOI with a verb labeled as $y_v \in R_v^N$ and an object labeled as $y_o \in R_o^N$, where $0 \leq y_o < N_o$ denotes the index of object category, we update y_v as follows,

$$\widetilde{y}_v = max(y_v + \mathbf{M}(:, y_o), 1) \tag{2}$$

where max means we clip the value to 1 if the value is larger than 1. Then, we obtain pseudo verb label $\widetilde{y}_v$ to optimize the samples of the HOI similar as Equation 7 (here, we only have annotated HOI samples). We think the running concept confidence $\mathbf{M}$ have **implicitly counted the distribution of verb and object in the dataset**. Meanwhile, the denominator in Equation 2 can

also normalize the confidence according to the frequency, and thus ease the long-tailed issue. Thus, with the pseudo labels constructed from **M**, we can re-balance the distribution of the dataset, which is a bit similar to re-weighting strategy [1, 3]. However, SCL does not require to set the weights for each class manually.

Table 12 demonstrates SCL greatly improves Qpic on Unseen category on rare first zero-shot detection, while SCL significantly facilitates rare category on non-rare first zero-shot detection. In Full HOI detection on HICO-DET, Table 11 shows SCL largely facilitates HOI detection on rare category. Particularly, the seen category in rare first setting includes 120 rare classes, while the seen category in non-rare first setting only includes 18 classes (all rare classes are in unseen category in non-rare first setting). Thus, SCL actually improves HOI detection for rare category. We think the concept confidence matrix internally learns the distribution of verb and objects and in the dataset. *e.g.*, given an object, **M** illustrates the corresponding verb distribution.

Table 12. Zero-Shot HOI detection based on Qpic. Results are reported by mean average precision (%). Here, we split the classes of HOI into four categories in zero-shot setting, *i.e.* , Seen are categorized into rare and non-rare.

Method	Unseen	Rare	NonRare	Full
Qpic [11] (non-rare first)	21.03	19.12	25.59	23.19
Qpic+SCL (non-rare first)	**21.73**	**22.43**	**26.03**	**24.34**
Qpic [11] (rare first)	15.24	16.72	30.98	27.40
Qpic+SCL (rare first)	**19.07**	16.19	30.89	**28.08**

Two-Stage method. Considering the HOI concept discovery is mainly based on two-stage HOI detection approaches [5], it is direct and simple to evaluate the performance of self-training on HOI detection. Table 10 demonstrates the HOI detection results on ground truth boxes. Noticeably, we directly predict the verb category, rather than HOI category. Thus, the baseline of HOI detection (*i.e.* visual compositional learning [5]) is a bit worse. We can find self-training also slightly improves the performance, especially on rare category.

G Visualization

In this section, we provide more visualized illustrations.

More Grad-CAM Visualizations Figure 2 demonstrates the visualization of Qpic and Qpic+SCL: the second row is Qpic and the third row is Qpic+SCL, where we observe a similar trend to the Gram-CAM illustration in main paper.

Concept Visualization. We illustrate the visualized comparisons of concept discovery in Figure 3. According to the ground truth and known concepts, we find some verb (affordance) classes can be applied to most of objects (the row is highlighted in the ground truth figure). This observation is reasonable because some kinds of actions can be applied to most of objects in visual world, *e.g.*, hold.

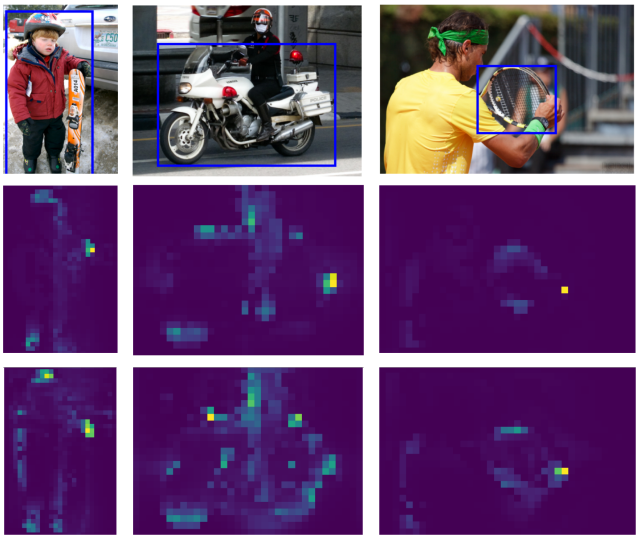

Fig. 2. Visualized Illustration of SCL+Qpic and Qpic [11].

As shown in Figure 3, there are many false positive predictions in the results of affordance prediction, and affordance prediction tends to overfit the known concepts, especially those with frequently appeared verbs. Methods of online HOI concept discovery on V-COCO have fewer false positive predictions compared to affordance prediction. However, the two methods tend to predict concepts composed of frequent verbs in known concepts due to the verb and object imbalance issues in HOI dataset [7]. Particularly, the false positive predictions are largely eased with self-training (*e.g.*, the top right region). In addition, the blank columns in Figure 3 are because there are only 69 objects in V-COCO training set, and we can ease it via training network with additional object images [6] as illustrated in the last figure of Figure 3. See more visualized results on HICO-DET and V-COCO in the supplemental material. Particularly, we further notice there are dependencies between verb classes (See verb dependency analysis).

H Additional Concept Discovery Approaches

We provide More comparisons in this Section. For a fair comparison with ATL [6] (*i.e.* , affordance prediction), we use the same number of verbs (21 verbs) on V-COCO. The code includes how to convert V-COCO to 21 verbs, *i.e.* merge "_instr" and "_obj" and remove actions without object (*e.g.*, stand, smile, run).

Language embedding baseline. In the main paper, we illustrate a random baseline. Here we further illustrate the results with language embedding [10]. Different from extracting verb/object features from real HOI images, we use the corresponding language embedding representations of verb/object as input, *i.e.*

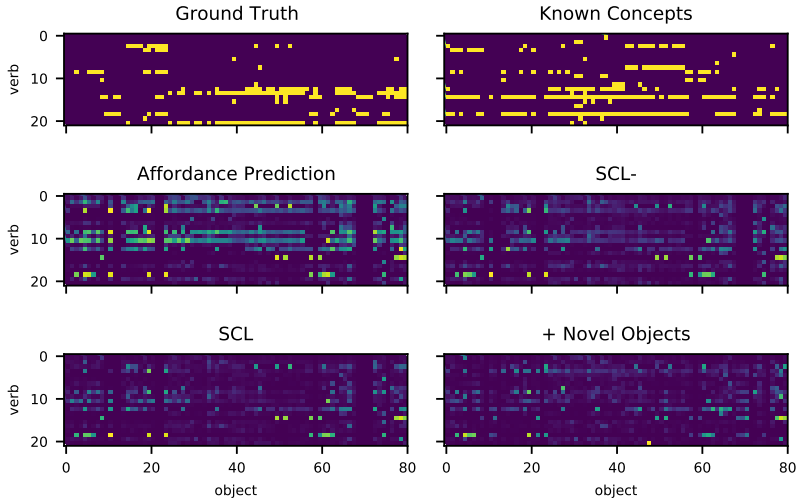

Fig. 3. Visualized Comparison of different methods on V-COCO dataset. The column is the object classes and the row represents the verb classes. Known Concepts are the concepts that we have known. SCL− means online concept discovery without self-training. For better illustration, we filter out known concepts in proposed methods. "+ Novel Objects" means self-training with novel object images.

discovering concepts from language embedding. Table 9 shows the performance is just a bit better than random result, and is much worse than online concept discovery. Similar to the main paper, when we evaluate the unknown concepts, we mask out the known concepts to avoid the disturbance from known concepts.

Re-Training. We first train the HOI model via visual compositional learning [5], and then predict the concept confidence. Next, we use the predicted concept confidence to provide pseudo labels for the composite HOIs. Table 9 shows the performance of Re-Training is worse than SCL.

With COCO dataset. Table 9 also demonstrates the baseline (SCL−) with COCO datasets has poor performance on concept discovery. We think it is because the domain shift between COCO dataset and HICO-DET dataset. However, SCL still achieves significant improvement on concept discovery.

Qpic+SCL. The details are provided in Section D.

I Object Affordance Recognition

SCL requires more iterations to converge, and achieves better performance on object affordance recognition. Table 13 shows the performance of the model without self-training does not improve with more training iterations.

Table 13. Comparison of object affordance recognition with HOI network (trained on HICO-DET) among different datasets. Val2017 is the validation 2017 of COCO [9]. Here we illustrate the result of SCL− under different training iterations.

Method	Type	Val2017	Obj365	HICO	Novel
SCL− (0.8M iters)	U	43.61	41.14	47.56	14.46
SCL− (1.5M iters)	U	44.07	39.05	50.27	10.19

References

1. Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *ICML*, pages 872–881. PMLR, 2019.
2. Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, pages 9004–9013, 2021.
3. Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, pages 9268–9277, 2019.
4. Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *BMVC*, 2018.
5. Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020.
6. Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021.
7. Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021.
8. Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, pages 74–83, 2021.
9. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.
10. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.
11. Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021.
12. Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021.