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

1. Overview

We provide more examples between HOI and affordance in Section B. ATL for One-Stage HOI detection is illustrated in Section C. Section D contains more details. Section E shows more comparison of object affordance recognition and the ablation study of the number of verbs on affordance recognition. We illustrate the Non-COCO classes that we select from Object365 on section F. Section G provides additional affordance results (mAP) and additional illustration of recent HOI approaches. Lastly, we compare prior approaches (*i.e*. VCL [6], FCL [5]) and ATL in detail.

2. More Examples of HOI and Object Affordance

Images labeled with HOI annotations simultaneously show the affordance of the objects. Therefore, we can not only learn to detect HOIs, but also learn to recognize the affordance of the objects. By combining the affordance representation with various kinds of its corresponding objects, we enable the HOI model to recognize the affordance of novel objects.

3. Affordance Transfer Learning for One-Stage HOI detection

Current HOI approaches mainly include one-stage methods [9, 16, 15] and two-stage methods [4, 6]. In main paper, we simultaneously evaluate ATL on both one-stage method and two-stage method. We implement ATL based on the code of [15], which implement HOI detection based on Faster-RCNN [12]. In details, we use 2 object images and 4 HOI images for each batch with 2 GPUs. Here, we regard the concatenation of features extracted from union and human boxes with RoI Align separately as verb feature. We regard the feature extracted from object boxes as object feature. We compose novel HOIs from object features and verb features between HOI images and object images. Different from two-stage method, we also compose object features and verb features between HOI images (*i.e*. VCL [6]). Meanwhile, during optimization, we keep the object detection learning for object images. For baseline, we remove the compositional approach (*i.e*. without compositional learning loss). Code is aviable at https://github.com/zhihou7/HOI-CL-OneStage.

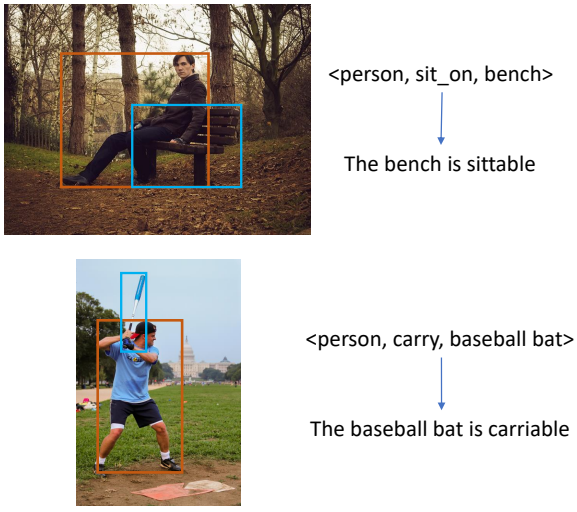

Figure 1. Examples about HOI and Affordance.

4. Supplementary Description

. In the Table 4 (object affordance recognition) in paper, we illustrate the affordance recognition of novel classes in zero-shot HOI detection on HICO-DET. All objects in HICO-DET, Val2017, Object365 with COCO classes are from the 12 novel classes (unseen objects). In the Novel Classes category, those objects are still Non-COCO classes. Besides, we can use both COCO images and HOI images

as object images in our experiments. But we do not find any improvement on HICO-DET. Thus in Table **??**, we do not include the result when we use two datasets as object images. It might be because there are nearly 900,000 object instances in COCO while HICO has only aroud 100, 000 object instances. For novel object zero-shot, there are too much many composite HOIs for seen HOIs, we thus remove some COCO object images for balancing the data.

5. Additional Ablation Study

The effect of different number of object images on affordance recognition. Table 1 illustrates the comparison of object affordance recognition among different number of object images in the minibatch on HICO-DET dataset. We find ATL with two images in each batch apparently improves the performance of object affordance recognition compared to one image and three images among COCO categories. Moreover, we find with more object images in each batch, ATL further improves the affordance recognition performance on Non-COCO classes. This means with multiple object images, ATL has better generalization of affordance recognition to novel classes.

The effect of the number of verbs on affordance recognition In affordance recognition, we randomly choose M instances for those affordances with more than M instances in dataset and all instances for other affordances. We ablate M in Table 2 under the ATL model with COCO objects and our baseline. The baseline is the model without compositional learning. Besides, when we use different M, we also update S_i. If we keep S_i same as the number when $M = 100$, all results will be very small when $M < 100$.

Table 2 shows the number goes stable after 20. This means we do not need to store a large number of templates of affordance representation.

6. Non-COCO classes

For evaluating ATL on affordance recognition of unseen classes, we manually select 12 non-coco classes from object365: glove, microphone, american football, strawberry, flashlight, tape, baozi, durian, boots, ship, flower, basketball. The actions that we can act on those objects (*i.e*. affordance) on HOI-COCO and HICO-DET are list on Table 3 and Table 4 respectively.

We further provide some visual examples of the Non-COCO classes in Figure 2. ATL can recognize the affordance of those objects without being interacted by combining the affordance representation and those object features.

7. Additional Results and Comparision

We find the metrics (Recall, Precision, F1) the paper (first version) uses is not much robust. F1 is sensitive to the confidence. Thus, we further evaluate the affordance recognition in Table 5 by Mean average Precision (mAP) (%). Table 5 shows the compositional learning approach consistently improves the baseline among all categories.

Due to the limitation of space in main paper. Other recent HOI detection methods are provided in Table 8.

8. Discussion between Prior Approaches

ATL extends VCL [6] by composing verbs and objects from object detection datasets which do not have HOI annotations. ATL presents a way to explore a broader source of data for HOI detection. Meanwhile, ATL finds that the HOI network trained with compositional learning can be simultaneously applied to affordance recognition. Meanwhile, ATL shows with more data, ATL can improve the generalization of affordance recognition on new dataset.

Prior to ATL, Fabricated Compositional Learning [5] was presented to fabricated objects to ease the open long-tailed issue for HOI detection. FCL [5] inspires our to compose novel HOIs from verb features from HOI images and object features from external object datasets. Compared to VCL [6] and ATL [6], FCL [6] is more flexible to generate balanced objects for each verb, and thus achieves better performance on some zero-shot settings. However, FCL also has some limitations. Although FCL achieves similar even better performance to ATL in HOI detection, Table 5 shows the model of FCL in fact is unable to recognize affordance. In addition, Table 6 further shows although FCL [6] achieves also good results on Novel Object HOI detection with a generic object detector, the results of FCL [6] on Unseen category goes to zero without a generic object detector.

We further illustrates the complementary between FCL and VCL in Table 7. Here, we fuse the prediction results of the two model to evaluate the complementary. We find this can largely improves the result.

References

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

[2] 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. 4

[3] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 4

[4] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *BMVC*, 2018. 1

[5] Zhi Hou, Yu Baosheng, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 1, 2, 4

Table 1. Additional Ablation study of object affordance recognition with HOI network among different number of object images on HICO-DET. Val2017 is the validation 2017 of COCO [10]. Subset of Object365 is the validation of Object365 [13] with only COCO labels. Novel classes are selected from Object365 with non-COCO labels. Object means what object dataset we use. The content in parentheses indicates the number of images in each batch.

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
ATL (1)	HICO	COCO	16.63	62.91	24.73	12.47	39.92	17.45	12.47	52.45	18.66	7.30	18.44	9.73
ATL (2)	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 (3)	HICO	COCO	27.36	78.21	38.12	21.84	53.57	28.25	13.85	57.42	20.94	**12.15**	**26.07**	**15.56**

Table 2. The effect of different number of verbs in affordance feature bank. Mean average Precision (mAP) (%) is reported. Dataset means the evaluation object dataset. HICO-DET means the test set of HICO-DET. Val2017 means the validation set of COCO2017.

$\#M$	Dataset	1	5	10	20	40	80	100
Baseline	Val2017	13.39	15.90	17.69	18.74	19.25	19.67	**19.71**
ATL (COCO)	Val2017	52.98	53.74	55.40	55.19	54.88	55.77	**56.05**
Baseline	HICO-DET	14.77	18.30	20.22	21.70	22.21	23.00	**23.18**
ATL (COCO)	HICO-DET	56.04	58.03	**59.14**	57.84	56.61	57.23	57.41

Table 3. Affordances of Non-COCO classes from Object365 on HOI-COCO.

name	verbs/affordances
glove	carry, throw, hold
microphone	talk_on_phone, carry, throw, look, hold
american football	kick, carry, throw, look, hit, hold
strawberry	cut, eat, carry, throw, hold
flashlight	carry, throw, hold
tape	carry, throw, hold
baozi	eat, carry, look, hold
durian	eat, carry, hold
boots	carry, hold
ship	ride, sit, lay, look
flower	look, hold
basketball	throw, hold

Table 4. Affordances of Non-COCO classes from Object365 on HICO-DET.

name	verbs/affordances
glove	buy, carry, hold, lift, pick_up, wear
microphone	carry, hold, lift, pick_up
american football	block, carry, catch, hold, kick, lift, pick_up, throw
strawberry	buy, eat, hold, lift, move
flashlight	buy, hold, lift, pick_up
tape	buy, hold, lift, pick_up
baozi	buy, eat, hold, lift, pick_up
durian	buy, hold, lift, pick_up
boots	buy, hold, lift, pick_up, wear
ship	adjust, board
flower	buy, hold, hose, lift, pick_up
basketball	block, hold, kick, lift, pick_up, throw

Figure 2. Examples of Non-COCO classes and its affordances.

[6] 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. 1, 2, 4

[7] 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. 4

[8] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, pages 10166–10175, 2020. 4

[9] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, and Jiashi Feng.

Table 5. Comparison of object affordance recognition with HOI network among different datasets. Val2017 is the validation 2017 of COCO [10]. Object365 is the validation of Object365 [13] 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. All results are reported by Mean average Precision (mAP)(%).

Method	HOI Data	Object	Val2017	Object365	HICO-DET	Novel classes
Baseline	HOI	-	31.91	26.16	44.00	14.27
FCL [5]	HOI	-	41.89	32.20	55.95	18.84
VCL [6]	HOI	HOI	76.43	69.04	86.89	32.36
ATL	HOI	HOI	76.52	69.27	87.20	34.20
ATL	HOI	COCO	**90.84**	**85.83**	**92.79**	**36.28**
Baseline	HICO	-	19.71	17.86	23.18	6.80
FCL [6]	HICO	-	25.11	25.21	37.32	6.80
VCL [6]	HICO	HICO	36.74	35.73	43.15	12.05
ATL	HICO	HICO	52.01	**50.94**	**59.44**	**15.64**
ATL	HICO	COCO	**56.05**	40.83	57.41	8.52
ATL^{ZS}	HICO	HICO	24.21	20.88	28.56	12.26
ATL^{ZS}	HICO	COCO	**35.55**	**31.77**	**39.45**	**13.25**

Table 6. Comparison of Zero Shot Detection results of between FCL [5] and ATL. NO means novel object HOI detection. * means we only use the boxes of the detection results.

Method	Type	Unseen	Seen	Full
FCL [5]	NO	15.38	21.30	20.32
ATL (COCO)	NO	15.11	**21.54**	**20.47**
FCL [5]	NO	0.00	13.71	11.43
ATL (COCO)*	NO	**5.05**	**14.69**	**13.08**

Table 7. Evaluation of the complementary between ATL and FCL. We use the released model of FCL [5].

Method	Full	Rare	Non-Rare
FCL [5]	24.68	20.03	26.07
ATL (COCO)	24.50	18.53	26.28
FCL + ATL	25.63	21.18	26.95

Table 8. Additional Illustration of recent HOI detection approaches.

Method	Default			Known Object		
	Full	Rare	NonRare	Full	Rare	NonRare
FG [1]	21.96	16.43	23.62	-	-	-
VSGNet [14]	19.80	16.05	20.91	-	-	-
DJ-RN [8]	21.34	18.53	22.18	23.69	20.64	24.60
IP-Net [16]	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [9]	21.73	13.78	24.10	24.58	16.65	26.84
Kim *et al.* [2]	17.58	11.72	19.33	19.76	14.68	21.27
ACP [7]	20.59	15.92	21.98	-	-	-
PD-Net [17]	20.81	15.90	22.28	24.78	18.88	26.54
FCMNet [11]	20.41	17.34	21.56	22.04	18.97	23.12
VCL [6]	23.63	17.21	25.55	25.98	19.12	28.03
DRG [3]	24.53	19.47	26.04	27.98	23.11	29.43
ATL (COCO) ^{VCL}	24.50	18.53	26.28	27.23	21.27	29.00
ATL (COCO) ^{DRG}	28.53	21.64	30.59	31.18	24.15	33.29

Ppdm: Parallel point detection and matching for real-time human-object interaction detection. *CVPR*, 2020. 1, 4

[10] 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, 4

[11] Y Liu, Q Chen, and A Zisserman. Amplifying key cues for human-object-interaction detection. *Lecture Notes in Computer Science*. 4

[12] 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. 1

[13] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, 2019. 3, 4

[14] Oytun Ulutan, ASM Iftekhar, and BS Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. *CVPR*, 2020. 4

[15] 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. 1

[16] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. *CVPR*, 2020. 1, 4

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