Supplementary Material
Visual Compositional Learning for Human-Object Interaction Detection

Zhi Hou[1,2], Xiaojiang Peng[2], Yu Qiao[2] *, and Dacheng Tao[1]

[1] UBTECH Sydney AI Centre, School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
zhou9878@uni.sydney.edu.au,dachengtao@sydney.edu.au
[2] Shenzhen Key Lab of Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{xj.peng, yu.qiao}@siat.ac.cn

1 Details of our baseline

We conduct experiments based on the code of [4] who released the code in their final version. We find there are two simple but very useful strategies for improvement: reweighting and postprocess for detection. Reweighting is that they allocate different weights for the cross entropy loss according to the number of classes. See 1. W is the weights that [4] provides.

$$L = W \cdot L_{cross\_entropy} \tag{1}$$

Another method is postprocess that they decrease the detection threshold for those images the the detector can't detect any objects and humans. We use the same test code to [4]. This strategy could improve recall largely. In Table 1, we can find the two strategies in line 72 in TIN_HICO.py and line 77 in test_HICO_pose_pattern_all_wise_pair.py from the released code of [4].

Table 1. Comparison of strategies from the released code of [4]

Strategy	Full (mAP %)	Rare (mAP %)	NonRare (mAP %)
w/o reweighting	16.87	10.07	18.90
w/o postprocess	17.14	12.92	18.40
our baseline	18.03	13.62	19.35

Besides, We also find different Hyper-Parameters also affect the performance. See next Section.

* corresponding author

2 Hyper-Parameters

In our proposed framework, there are two hyper-parameters λ_1 and λ_2. We evaluate the performance when we set different values for the two hyper-parameters.

From Table 2, when we increase the value of λ_1, We can witness a considerable increase in the Full category. If we choose the value more than 2.0 for λ_1, the performace slightly decreases. From Table 3, if we set 0.5 or 0.1, the performance is similar. But, when λ_2 is more than 1.0 or less than 0.1, the performance drops quickly.

Like [2, 4], we first detect the objects in the image and then use the object detection results to infer the HOI categories during test. We use the same score threshold (0.8 for human and 0.3 for object) same as [4] in resnet50 coco detector. We use 0.3 for human and 0.1 for object in resnet101 detector that is finetuned on HICO-DET dataset since the finetuned object detection result is largely better.

Table 2. The results of setting different values for λ_1 when λ_2 is 0.5 in HICO-DET.

λ_1	1.0	1.5	2.0	2.5	3
Full	18.96	18.95	19.43	19.29	19.34

Table 3. The results of setting different values for λ_2 when λ_1 is 2.0 in HICO-DET.

λ_2	0.05	0.1	0.5	1.0	1.5
Full	19.18	19.30	19.43	19.10	18.90

3 The effect of the number of interactions in minibatch

In order to compose enough interactions for Visual Compositional Learning, we increase the number of interactions in each minibatch while reducing the number of augmentations for each interaction and the number of negative interactions. Therefore, the batch size is nearly unchanged in our experiment and we can still optimize the network in a single GPU. We evaluate the effect in this section. We set the maximum number of interactions 5 in our experiment. Noticeably, most training images in HICO-DET only contain one interaction.

From Table 4, we can find the baseline model of different iteractions has similar results with 18.43 mAP and 18.47 mAP respectively. However, we witness a better improvement (1.0 mAP vs 0.44 mAP) if we increase the interaction classes in the minibatch. It shows that increasing the number of interactions is considerably beneficial for Visual Compositional Learning.

Table 4. The results of the number of interactions in minibatch in HICO-DET.

the number of interactions	VCL	Full	Rare	NonRare
1	-	18.41	14.17	19.68
1	✓	18.85	14.98	20.01
5	-	18.43	14.14	19.71
5	✓	**19.43**	**16.55**	**20.29**

4 The two branches in zero-shot HOI detection

Table 5. Two branches ablation study of the proposed Visual Compositional Learning framework in zero-shot HOI detection on HICO-DET test set during inference.

Method	Unseen	Seen	Full
Verb-Object branch (rare first)	**7.85**	15.48	13.95
Spatial-Human branch (rare first)	4.33	15.92	13.60
Two branches (rare first)	7.55	**18.84**	**16.58**
Verb-Object branch (non-rare first)	**10.61**	10.95	10.88
Spatial-Human branch (non-rare first)	5.71	11.82	10.60
Two branches (non-rare first)	9.13	**13.67**	**12.76**

We evaluate the contribution of the two branches in zero-shot HOI detection. From Table 5, we can find the performance of verb-object branch in Seen category and Full category is similar to that of spatial-human branch, while verb-object branch is 3.52% and **4.90%** better than spatial branch in selecting rare first and selecting non-rare first respectively in the Unseen category. **Particularly, after we fuse the result of the two branches, the Unseen category witnesses a considerable decrease in the two selecting strategies.** This illustrates that the additional spatial-human branch contributes to the full performance while *the verb-object branch with VCL efficiently benefits the zero-shot recognition.*

5 Verb Polysemy Problem

There is a verb polysemy problem in HOI detection, that is the verb "play" has different meanings between "play guita" and "play football". But, HICO restricts itself to a single sense of a verb (with the exceptions of a couple of verbs) [1, 3], which means that the verb polysemy problem is not serious. Previous HOI approaches [5–7] usually regard the verb from different HOIs as same, and successfully achieve good performance. We also conduct a simple experiment to validate this problem. We use the language priors to choose the suitable composited HOIs according to the object similarity of word embedding in Table 6.

We can find the improvement of language priors is very limited. This experiment also demonstrates the verb polysemy problem is not serious in HICO-DET dataset.

Table 6. Illustrations of VCL with language priors.

Strategy	Full (mAP %)	Rare (mAP %)	NonRare (mAP %)
VCL	19.43	16.55	20.29
VCL + Language prior	19.56	16.27	20.55

6 Visual Illustration of zero-shot HOI detection

Similar to Figure 4 in the paper, we qualitatively show that the proposed Visual Compositional Learning framework can detect those unseen interactions efficiently in Figure 1 while the baseline model without Visual Compositional Learning misdetects on HICO-DET. It shows our proposed Visual Compositional Learning framework is significantly beneficial for Unseen categories.

Fig. 1. Some HOI detections detected by the proposed Compositional Learning and the model without Compositional Learning in zero-shot HOI detection (selecting nonrare first). The first row is the results of our baseline model without VCL. The second row is the results of the proposed composition learning. The unseen interactions are marked with purple. We illustrate top 5 score results for the human object pair.

7 Unseen labels on HICO-DET dataset

In zero-shot detection in HICO-DET, we select randomly unseen labels for zero-shot detection. In detail, we first sorted the labels according to the number of

instances of categories. Then we select the HOIs out for unseen data according to the sorted label list and meanwhile make sure that all types of objects and verbs exist in seen data. we provide the unseen label id in two zero-shot learning settings.

rare first ids: 509, 279, 280, 402, 504, 286, 499, 498, 289, 485, 303, 311, 325, 439, 351, 358, 66, 427, 379, 418, 70, 416, 389, 90, 395, 76, 397, 84, 135, 262, 401, 592, 560, 586, 548, 593, 526, 181, 257, 539, 535, 260, 596, 345, 189, 205, 206, 429, 179, 350, 405, 522, 449, 261, 255, 546, 547, 44, 22, 334, 599, 239, 315, 317, 229, 158, 195, 238, 364, 222, 281, 149, 399, 83, 127, 254, 398, 403, 555, 552, 520, 531, 440, 436, 482, 274, 8, 188, 216, 597, 77, 407, 556, 469, 474, 107, 390, 410, 27, 381, 463, 99, 184, 100, 292, 517, 80, 333, 62, 354, 104, 55, 50, 198, 168, 391, 192, 595, 136, 581

non-rare first ids: 38, 41, 20, 18, 245, 11, 19, 154, 459, 42, 155, 139, 60, 461, 577, 153, 582, 89, 141, 576, 75, 212, 472, 61, 457, 146, 208, 94, 471, 131, 248, 544, 515, 566, 370, 481, 226, 250, 470, 323, 169, 480, 479, 230, 385, 73, 159, 190, 377, 176, 249, 371, 284, 48, 583, 53, 162, 140, 185, 106, 294, 56, 320, 152, 374, 338, 29, 594, 346, 456, 589, 45, 23, 67, 478, 223, 493, 228, 240, 215, 91, 115, 337, 559, 7, 218, 518, 297, 191, 266, 304, 6, 572, 529, 312, 9, 308, 417, 197, 193, 163, 455, 25, 54, 575, 446, 387, 483, 534, 340, 508, 110, 329, 246, 173, 506, 383, 93, 516, 64

References

1. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1017–1025 (2015)
2. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. arXiv preprint arXiv:1808.10437 (2018)
3. Gella, S., Keller, F., Lapata, M.: Disambiguating visual verbs. IEEE transactions on pattern analysis and machine intelligence **41**(2), 311–322 (2017)
4. 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. arXiv preprint arXiv:1811.08264 (2018)
5. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
6. Shen, L., Yeung, S., Hoffman, J., Mori, G., Fei-Fei, L.: Scaling human-object interaction recognition through zero-shot learning. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1568–1576. IEEE (2018)
7. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)