Spatially Conditioned Graphs for Detecting Human–Object Interactions Supplementary Material

Frederic Z. Zhang[1,3] Dylan Campbell[2,3] Stephen Gould[1,3]
[1]The Australian National University [2]University of Oxford
[3]Australian Centre for Robotic Vision
{firstname.lastname}@anu.edu.au dylan@robots.ox.ac.uk

A. Known Object Setting for HICO-DET

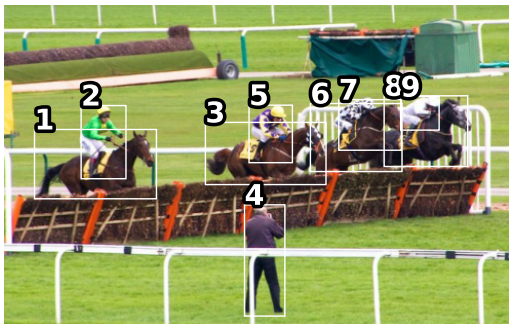
While the default setting for HICO-DET [1] has been the more popular evaluation protocol, there is an additional less frequently reported known object setting, where the object types of ground truth interactions in images are considered known, thus automatically removing predicted interactive pairs with other object types. For interested readers, we provide the performance of our model in comparison with other methods under the known object setting in Table 1.

Table 1. HOI detection performance (mAP×100) on the HICO-DET [1] test set under the known object setting. The most competitive method in each category is in bold, while the second best is underlined.

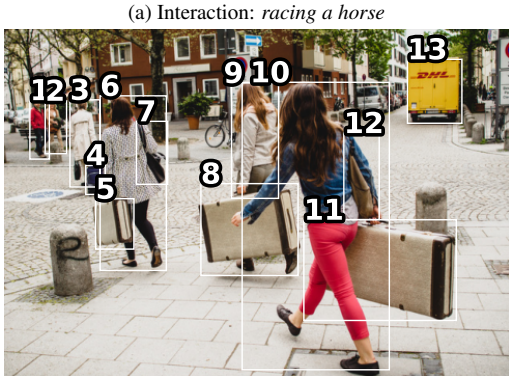
Method	Backbone	Full	Rare	Non-rare
DETECTOR PRE-TRAINED ON MS COCO				
HO-RCNN [1]	CaffeNet	10.41	8.94	10.85
iCAN [3]	ResNet-50	16.26	11.33	17.73
TIN [6]	ResNet-50	19.17	15.51	20.26
DRG [2]	ResNet-50-FPN	23.40	21.75	23.89
VCL [4]	ResNet50	22.00	19.09	22.87
IDN [5]	ResNet50	**26.43**	**25.01**	**26.85**
Ours	ResNet-50-FPN	<u>25.53</u>	<u>21.79</u>	<u>26.64</u>
DETECTOR FINE-TUNED ON HICO-DET				
PPDM [7]	Hourglass-104	24.58	16.65	26.84
VCL [4]	ResNet50	25.98	19.12	28.03
DRG [2]	ResNet-50-FPN	27.98	23.11	<u>29.43</u>
IDN [5]	ResNet50	<u>28.24</u>	<u>24.47</u>	29.37
Ours	ResNet-50-FPN	**34.37**	**27.18**	**36.52**
ORACLE DETECTOR				
Ours	ResNet-50-FPN	**51.75**	**41.40**	**54.84**

B. Additional Qualitative Results

We show more qualitative results to demonstrate the strength of our model in Figure 1. We intentionally select images that have many human instances and multiple human–object pairs of the same interaction. In Figure 1a, there are 20 combinatorial human–horse pairs, with 4 of



(a) Interaction: *racing a horse*



(b) Interaction: *carrying a suitcase*

Figure 1. Qualitative results. The scores corresponding to (a) are shown in Table 2 and the scores corresponding to (b) are shown in Table 3.

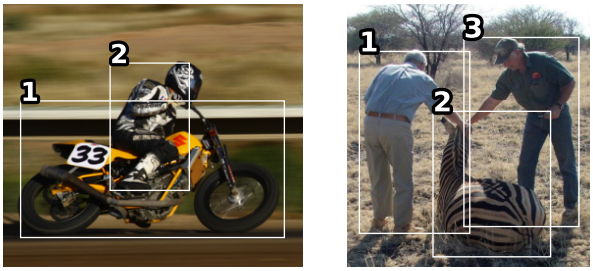
Table 2. Scores for the interaction *racing a horse* in Figure 1a. Each column corresponds to pairs with the same human instance. Each row corresponds to pairs with the same horse instance.

Instance index	2	4	5	7	9
1	**0.2031**	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	**0.5913**	0.0002	0.0000
6	0.0000	0.0000	0.0013	**0.0178**	0.0030
8	0.0000	0.0000	0.0001	0.0034	**0.1412**

Table 3. Scores for the interaction *carrying a suitcase* in Figure 1b. Each column corresponds to pairs with the same human instance. Each row corresponds to pairs with the same suitcase instance. Missing indices correspond to detections other than suitcases.

Instance index	**1**	**2**	**3**	**6**	**9**	**10**
4	0.0000	0.0000	**0.0391**	0.0021	0.0000	0.0000
5	0.0000	0.0000	0.0000	**0.1278**	0.0000	0.0004
8	0.0000	0.0000	0.0000	0.0178	**0.2791**	0.1098
11	0.0000	0.0000	0.0000	0.0003	0.0000	**0.3858**

them being interactive. As shown in Table 2, our model is able to assign highest scores to all four interactive pairs and suppress all non-interactive pairs. However, we do notice that small and clustered boxes can reduce the confidence of our model, e.g. person (7) and horse (6). This issue can also be seen in Figure 1b and Table 3. Our model is able to find the correct human–suitcase pairs $(10, 11)$, $(9, 8)$, $(6, 5)$ and predict high scores for them. Yet the positive pair $(3, 4)$ receives a very low score due to the size of the bounding boxes and less confident object detection scores. We also notice that person (10) and suitcase (8) receive a fairly high score for *carrying a suitcase*. This is due to the close relative location between the pair and a plausible gesture from the person. In such scenarios, access to the depth information could be helpful.



(a) Interaction: *racing a motocycle* (b) Interaction: *petting a zebra*

Figure 2. Qualitative results where images contain a small number of clean human and object instances.

We also show some qualitative results where our model does not improve upon previous methods in Figure 2. For examples such as in Figure 2a, where there is only one human–object pair, our graphical model is not particularly superior as there are only one human and object node each passing messages between each other. And in Figure 2b, when both human–zebra pairs are in fact interactive under the interaction *petting a zebra*, we found that the baseline model with appearance only is also able to correctly assign high scores to both pairs, as shown in Table 4.

To sum up, we found that our graphical model with spatial conditioning is more competitive on images with large number of human and object instances, particularly when there are multiple ground truth pairs of the same interaction, but does not improve upon previous methods on clean images with very few distractions.

Table 4. Scores for the interaction *petting a zebra* in Figure 2b

Human–zebra pairs	Scores (baseline)	Scores (ours)
(1, 2)	0.6782	0.7019
(1, 3)	0.6945	0.6799

C. Additional Ablations

Apart from the main contribution of the paper, we found a few other training techniques beneficial to our model. First, a larger batch size helps to stablise the focal loss. We normalise the focal loss by the number of positive logits, which in itself is a very unstable statistic. Increasing the batch size from 4 to 32 results in roughly 0.8 mAP improvement. Second, using AdamW [8] instead of SGD contributes about 1 mAP to our model's performance. We attribute this improvement to the similarity between graphical models and transformers [9], for which AdamW is the *de facto* choice of optimiser. Last, we observe a further 1 mAP improvement from fine-tuning the backbone.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018. 1

[2] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. DRG: Dual relation graph for human-object interaction detection. *Eur. Conf. Comput. Vis.*, 2020. 1

[3] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. *Brit. Mach. Vis. Conf.*, 2018. 1

[4] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Eur. Conf. Comput. Vis.*, 2020. 1

[5] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *Adv. Neural Inform. Process. Syst.*, 2020. 1

[6] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. *Int. Conf. Comput. Vis.*, 2019. 1

[7] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. PPDM: Parallel point detection and matching for real-time human-object interaction detection. *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 1

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *Int. Conf. Learn. Represent.*, 2018. 2

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30:5998–6008, 2017. 2