Chairs Can be Stood on: Overcoming Object Bias in Human-Object Interaction Detection Supplementary Material

Guangzhi Wang[1], Yangyang Guo[2⋆], Yongkang Wong[2], and Mohan Kankanhalli[2]

[1] Institute of Data Science, National University of Singapore
[2] School of Computing, National University of Singapore
guangzhi.wang@u.nus.edu guoyang.eric@gmail.com
yongkang.wong@nus.edu.sg mohan@comp.nus.edu.sg

1 Additional Experiments

1.1 Known Object Setting

Following previous work [2, 11, 21], we also report results on the HICO-DET dataset [2] with Known Object (KO) setting in Tab. 1 and Tab. 2, respectively. It can be observed that our method surpasses the baselines under this setting.

Table 1: Performance comparison on HICO-DET under the Known Object (KO) setting with *pre-trained* detector.

Method	Full	Rare	Non-rare
iCAN [6]	16.26	11.33	17.73
TIN [12]	19.17	15.51	20.26
DRG [5]	23.40	21.75	23.89
VCL [7]	22.00	19.09	22.87
DJ-RN [10]	23.69	20.64	24.60
SCG* [21]	24.53	20.00	25.88
SCG + Ours	**25.54**	**21.93**	**26.61**

Table 2: Performance comparison on HICO-DET under the Known Object (KO) setting with *fine-tuned* detector.

Method	Full	Rare	Non-rare
PPDM [13]	24.58	16.65	26.84
HOI-Trans [22]	26.15	19.24	28.22
ATL [8]	27.38	22.09	28.96
AS-Net [3]	31.74	27.07	33.14
FCL [9]	31.31	25.62	33.02
SCG* [21]	33.74	26.41	35.95
SCG + Ours	**34.52**	**27.34**	**36.67**

1.2 Hyper-parameter Analysis

The detailed analysis of the coordination of these two classifiers with respect to λ is shown in Tab. 3. It can be seen both classifiers are essential for the performance improvement.

1.3 Efficiency and Memory Comparison

We compare the memory and computational cost with SCG [21] in Tab. 4. Note that the adopted detector (*i.e.*, Faster R-CNN [15]) is not counted as the detection results can be obtained via one-pass inference for all images before training. It can be observed the overhead brought by our method is negligible for both training and test.

⋆ corresponding author

Table 3: Performance comparison with different λ.

λ	**Full**	**Rare**	**Non-rare**	**OR**	**ONR**	**AVE**
0.2	21.32	17.39	22.50	19.33	24.66	22.00
0.4	**21.50**	**17.59**	**22.67**	**19.47**	**25.01**	**22.14**
0.6	21.41	17.44	22.60	19.33	24.98	22.16
0.8	21.17	17.10	22.38	19.02	25.01	22.02

Table 4: Efficiency and memory comparison.

	train/img	**test/img**	**#param (train)**	**#param (test)**
SCG [21]	428.16ms	248.50ms	16.04M	16.04M
+Ours	440.72ms	251.32ms	17.12M	16.50M

2 Implementations

2.1 Overall Implementations

We conducted all experiments on 4 Nvidia 2080Ti GPUs. Due to resource limitation, we reduced the batch size of SCG and QPIC to 8 and linearly scaled their learning rate.[3] For QPIC, we started from a trained model and finetuned it with the proposed method for a total of 15 epochs. The learning rate is decayed by 0.1 at the 10-th epoch. For the other two baselines, we followed the default scheduling and started the training of f_m from the $3rd$ epoch for stability. λ is empirically set to 0.4 in all experiments.

For the proposed method, we parameterize f_m as a three layer Multi-layer Perceptron with ReLU activation function. For each memory cell, we set the size n to 16 for each object and k to 4. About the write operation, τ^o is set to the third smallest w^o (for objects with more than 5 associated verbs) or 0 (other objects). Regarding to other aspects with respect to base models (feature extractor, sampling strategy, and loss function), we adopted their default settings. More details are as follows.

2.2 Implementation of Baselines

The implementation details of the baselines are listed in Tab. 5. In this table, the batch size is represented as *number of images per GPU* × *number of GPUs*. BCE stands for binary cross-entropy loss. Kindly find the codes with the corresponding model names in the zip file.

2.3 More Hyper-parameter Settings

Due to resource limitation, we used a smaller batch size and scaled learning rate for both the baseline (SCG) and our method in all previous experiments.

[3] This may slightly influence the performance and result in inconsistency between reported and reproduced ones.

Table 5: Implementation details of baseline methods

Method	SCG [21]	HOID [18]	QPIC [16]
Venue	ICCV'21	CVPR'20	CVPR'21
Batch Size (default)	4 × 8	4 × 4	2 × 8
Batch Size (ours)	2 × 4	4 × 4	2 × 4
Feature Extractor	ResNet50-FPN	ResNet50-FPN	ResNet-50
Proposal Generation	Faster-RCNN [15]	HO-RPN [18]	QPIC [16]/DETR [1]
Interaction Loss	BCE+focal [14]	BCE	BCE
Interactiveness Score [12]	Yes	No	No
Low-grade Suppression [12]	Yes	No	No
Sample Strategy	#human&#object	pos/neg ratio	None

Table 6: Model performance of different hyper-parameter settings.

Index	**Setting**	**Full↑**	**Rare↑**	**Non-rare↑**
1	Baseline (4 GPUs * 2 image, unscaled lr)	19.94	14.70	21.50
2	Baseline (4 GPUs * 1 image, scaled lr)	20.75	15.96	22.18
3	+ Ours	**21.16**	**17.41**	**22.28**
4	Baseline (4 GPUs * 2 image, scaled lr)	20.99	16.30	22.40
5	+ Ours	**21.50**	**17.59**	**22.67**

We also studied the performance of baseline and our method under other hyper-parameter settings in Tab. 6. It can be observed that a) smaller batch size results in worse performance (2&4, 3&5). b) linearly scaling learning rate with respect to batch size can prohibit performance degradation to some degree (1&4). c) Under different training settings, our method outperforms the baseline by a considerable margin (2&3, 4&5).

3 Discussion on Debiasing Baselines

Re-weighting Methods For re-weighting methods (*i.e.*, inverse frequency weighting and CB-Loss [4]), we followed their conventions and computed the number of HOI instances (*i.e.* interactive human-object pairs) in the training set to facilitate the weight calculation. However, this leads to severe performance degradation. We conjecture that there are mainly two reasons. Firstly, these loss functions are all designed for reducing the general bias, instead of the *object bias* studied in this paper. Secondly, these re-weighting strategies interfere a lot the original training process, which requires complex interaction recognition and reasoning. In contrast, our proposed method allows dynamic adjustment with respect to each HOI instance in the training process, thereby improving the performance.

General Debiasing Methods For Adversarial Training (AT) [19], we trained the model with another classifier, whose output dimension equals to the number of object classes (*i.e.*, 80 in HICO-DET). Given each human-object feature, a cross-entropy loss with a flat label, *i.e.*, $\mathbf{1}/N_o$ is introduced to the original training process, so that the representation is expected to be object-agnostic. For Domain Independent Training (DIT) [19], we trained the model with another classifier. The output dimension of this classifier equals to the number of total interactions (*i.e.*, 600 in HICO-DET). During inference, the interaction prediction is taken as the maximum probability over all interactions involving this verb. We observe significant performance degradation with these methods. The key reason to this is that these methods ignore the object factor in their representations, which is essential for interaction recognition.

SGG Debiasing Methods The original TDE [17] aims to alleviate the contextual bias in Scene Graph Generation (SGG). Besides the original forward pass, it conducts a second forward pass in the same model by masking (*e.g.*, set to zero) both the subjects and the objects. The final prediction is taken as the subtraction between the original logits and the logits in the second pass. In this way, the biasing effects caused by factors other than the subject and object are expected to be eliminated. In this work, to alleviate the object bias, we conduct a second forward pass by masking everything other than the object. Then, similarly, the final output logits is obtained by subtracting this logits from the original ones. By doing the subtraction, the output is expected to be less affected by the object bias problem, following the intuition of [17]. In PCPL [20], we take the representation of an HOI class as the average of all features that involve this interaction class. We argue that the failure of these methods may result from the ignorance of multi-label setting, which results in different logits in TDE (since single-label classification is conducted for SGG.) and imprecise class embedding estimation in PCPL (because an embedding for one instance may be counted into multiple classes, confusing the representations).

4 More Visualizations

4.1 More Memory Evolutions

We show the evolution of label distribution under another four randomly picked objects in Fig. 1. It can be observed that the model prefers to sample some frequent class instances at early iterations due to their dominance. When it comes to later training steps, rare class instances gain more attention with the help of the proposed ODM. By the end of the first epoch (*i.e.*, 4.5k iterations), the tail classes under each object is more frequently sampled.

4.2 More Qualitative Results

We provide additional qualitative results in Fig. 2. It can be seen that our method can effectively alleviate the *object bias problem* by reducing false negative errors.

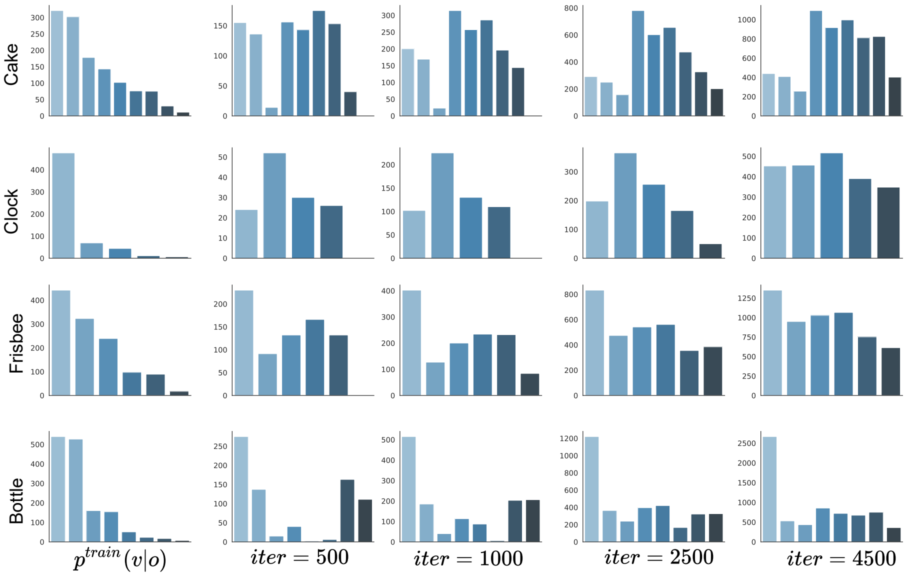

Fig. 1: Extra examples for the evolution of accumulated verb distribution after reading from the proposed ODM. The leftmost column shows $p^{train}(v|o)$ and the other 4 columns represent the sampled verb distributions at different iterations.

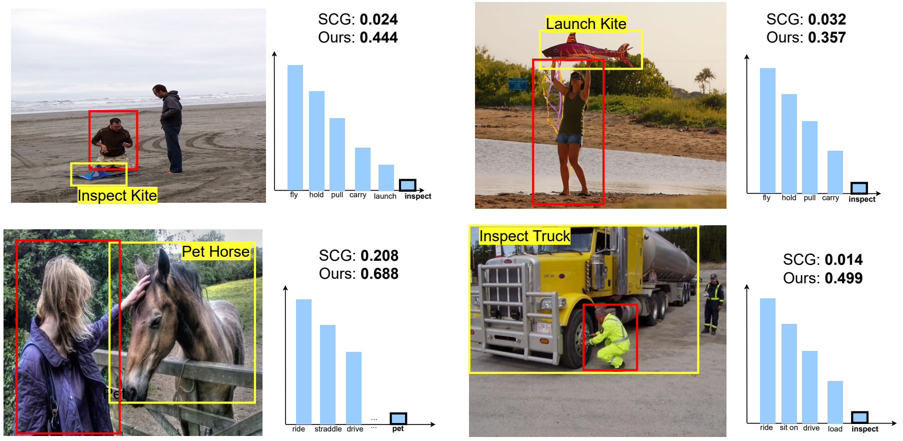

Fig. 2: Additional qualitative results. Human and object are bounded by red box and yellow box, where the tag indicates the ground truth interaction. For each example, the object-conditional verb distribution on training set $p^{train}(v|o)$ are shown, where the involved verb is bold.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020)
2. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018)
3. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating hoi detection as adaptive set prediction. In: CVPR (2021)
4. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR (2019)
5. Gao, C., Xu, J., Zou, Y., Huang, J.B.: DRG: Dual relation graph for human-object interaction detection. In: ECCV (2020)
6. Gao, C., Zou, Y., Huang, J.B.: iCAN: Instance-centric attention network for human-object interaction detection. In: BMVC (2018)
7. Hou, Z., Peng, X., Qiao, Y., Tao, D.: Visual compositional learning for human-object interaction detection. In: ECCV (2020)
8. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for human-object interaction detection. In: CVPR (2021)
9. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Detecting human-object interaction via fabricated compositional learning. In: CVPR (2021)
10. Li, Y.L., Liu, X., Lu, H., Wang, S., Liu, J., Li, J., Lu, C.: Detailed 2d-3d joint representation for human-object interaction. In: CVPR (2020)
11. Li, Y.L., Liu, X., Wu, X., Li, Y., Lu, C.: HOI analysis: Integrating and decomposing human-object interaction. In: NeurIPS (2020)
12. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. In: CVPR (2019)
13. Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J.: PPDM: Parallel point detection and matching for real-time human-object interaction detection. In: CVPR (2020)
14. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection (2017)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015)
16. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR (2021)
17. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: CVPR (2020)
18. Wang, S., Yap, K.H., Yuan, J., Tan, Y.P.: Discovering human interactions with novel objects via zero-shot learning. In: CVPR (2020)
19. Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. In: CVPR (2020)
20. Yan, S., Shen, C., Jin, Z., Huang, J., Jiang, R., Chen, Y., Hua, X.S.: Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In: ACM MM (2020)
21. Zhang, F.Z., Campbell, D., Gould, S.: Spatially conditioned graphs for detecting human-object interactions. In: ICCV (2021)
22. Zou, C., Wang, B., Hu, Y., Liu, J., Wu, Q., Zhao, Y., Li, B., Zhang, C., Zhang, C., Wei, Y., et al.: End-to-end human object interaction detection with hoi transformer. In: CVPR (2021)