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

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

Abstract. Detecting Human-Object Interaction (HOI) in images is an important step towards high-level visual comprehension. Existing work often shed light on improving either human and object detection, or interaction recognition. However, due to the limitation of datasets, these methods tend to fit well on frequent interactions conditioned on the detected objects, yet largely ignoring the rare ones, which is referred to as the **object bias problem** in this paper. In this work, we for the first time, uncover the problem from two aspects: unbalanced interaction distribution and biased model learning. To overcome the object bias problem, we propose a novel plug-and-play Object-wise Debiasing Memory (ODM) method for re-balancing the distribution of interactions under detected objects. Equipped with carefully designed read and write strategies, the proposed ODM allows rare interaction instances to be more frequently sampled for training, thereby alleviating the object bias induced by the unbalanced interaction distribution. We apply this method to three advanced baselines and conduct experiments on the HICO-DET and HOI-COCO datasets. To quantitatively study the object bias problem, we advocate a new protocol for evaluating model performance. As demonstrated in the experimental results, our method brings consistent and significant improvements over baselines, especially on rare interactions under each object. In addition, when evaluating under the conventional standard setting, our method achieves new state-of-the-art on the two benchmarks.

1 Introduction

Benefiting from the advancement of visual detection systems, Human-Object Interaction (HOI) detection has drawn increasing research interests in recent years. It requires detecting both humans and objects in a given image, based on which the interactions (often expressed as verb phrases) should also be correctly recognized. HOI detection is of vital importance to human-centric visual understanding and also benefits other high-level vision tasks, such as image captioning [29] and visual question answering [1,15].

⋆ corresponding author

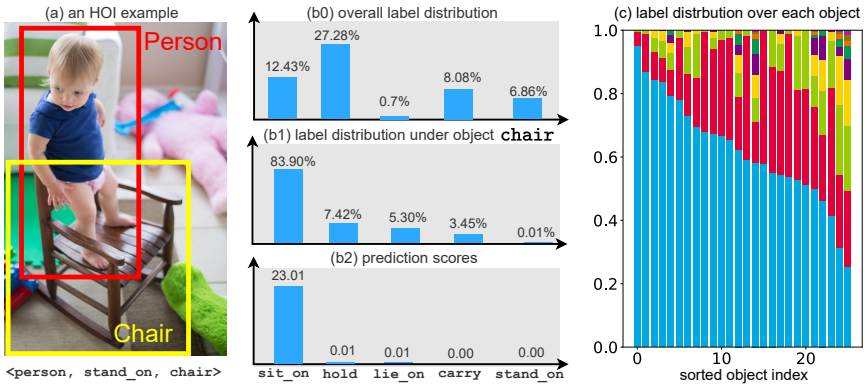

Fig. 1: An illustration of the object bias problem. Given the detected human-object pair in (a), the model [62] prediction (b2) is highly biased towards the object-conditional label distribution (b1), instead of the overall long-tail distribution in the training set (b0). As a result, the model predicts a more frequent verb `sit_on` for the object `chair`, leaving the true label `stand_on` ignored. (c) Label distribution from 25 randomly selected objects. It can be seen that most objects are dominated by one interaction (colored in blue).

Existing HOI detection efforts can be mainly categorized into two groups: two-stage and one-stage methods. Specifically, methods in the first group often leverage an off-the-shelf detector (*e.g.*, Faster R-CNN [42]) to initially detect the regions of humans and objects. The succeeding stage of interaction recognition can be enhanced with human part/pose understanding [18,49,30,10], graph-based message passing between humans and objects [41,67,12,52,47,62] or finer label space construction [27,64]. Some studies also exploit cross-dataset knowledge such as human-object interactiveness [32,33,56], cross-dataset objects [21] and word embeddings [57] to improve interaction recognition. Nonetheless, these approaches are often limited by deficiencies like inferior proposal generation or heavy inference overhead. To address these problems, one-stage methods often resort to performing detection and interaction classification within a single stage. Early studies treat HOI detection as a $\langle$`human`, `object`, `interaction`$\rangle$ point detection and matching [35,53,65] task. Recent approaches employ the Transformer-based detector [3] to aggregate contextual information and detect interaction in an anchor-free manner [44,68,5,25,60]. Nevertheless, increased training time is often encountered by this group of approaches.

Although existing methods have made progress over benchmarks, we observe one pervasive shortcoming that prevents them from further advancement. That is, the interaction prediction is strongly related to the detected object. Fig. 1 shows that given the detected object `chair` in (a), the model predicts (b2) the wrong verb `sit_on` with a very high confidence, rather than yields the true action - `stand_on`. Previous studies [60,20,21] mostly perceive this phenomenon as the outcome of learning from the long-tail label distribution from the overall training

set. Nevertheless, as we step further into this problem, we find it deviates a lot from the intuition of those methods. In particular, as shown in Fig. 1 (b0), the label `hold` dominates the training set and is twice frequent than `sit_on`. Out of expectation, the prediction score (Fig. 1 (b2)) for `hold` is only 0.01, which is 2,000 times smaller than that of `sit_on`. This observation brings our concern - is the wrong prediction really because of the long-tail label distribution in the overall training set? With this concern, we shift our focus to the interaction distribution under the detected object (Fig.1 (b1)), and discover a strong bias between the object and its conditional interaction distribution. Specifically, the model prediction conforms more with such object-induced bias, rather than the bias caused by the overall long-tail label distribution. In view of this, we can infer that during training, the object-induced bias drives the model to fit well on frequent interactions under each object, while overlooking the rare ones. However, rare classes are often more informative than non-rare ones [45,55]. Simply ignoring them undermines the model's representation ability, resulting in poor generalization and limited real-world applicability. Nonetheless, to the best of our knowledge, this bias problem has not been explored in the existing literature. As most objects struggle with the biased interaction distribution (Fig. 1 (c)), we therefore humbly suggest this problem to the community, and name it as the *object bias problem* in this work.

As a matter of fact, dealing with this problem is non-trivial due to the inherent distribution imbalance in existing benchmarks. However, building a balanced dataset is time and labor intensive. One alternative solution is to feed the model with balanced samples during training, which has been extensively proved effective in previous studies [43,55,28]. Yet, directly applying these methods to HOI detection is sub-optimal, as the object bias problem is actually induced by the class imbalance *under each object*, rather than that of the overall training set. To this end, we propose a novel Object-wise Debiasing Memory (ODM) module to achieve object-conditional class balancing. The proposed ODM is implemented with an object-indexed memory, upon which read and write strategies are designed to support the retrieval and storage of HOI features and labels. For memory reading, we take the label of each interactive instance as query to retrieve instances from the memory. Our read strategy assures that rare class instances are more frequently sampled, leading to a more balanced label distribution within the batch for training. On the other hand, the writing strategy is devised to store rare class instances with higher probability. In this way, the unbalanced interaction distribution under each object is mitigated, thus reducing the influence of the *object bias problem*.

We conduct extensive experiments over two benchmark datasets, namely HICO-DET [4] and HOI-COCO [21]. In addition, we also advocate a new object bias evaluation protocol to quantitatively evaluate the model performance under the object-biased condition. When equipped with our method, several advanced baselines are evidently shown to overcome the object bias problem, thereby achieving improved performance.

To summarize, our contributions are three-fold:

- We systematically study the object bias problem in the HOI detection task. To the best of our knowledge, we are the first to recognize and address this problem in the HOI literature.
- To alleviate the object bias problem, we propose a novel ODM module to facilitate the learning of a balanced classifier. The proposed ODM is model-agnostic and applicable to both one-stage and two-stage methods.
- We conduct extensive experiments on benchmark datasets, namely HICO-DET [4] and HOI-COCO [21]. When applying our method to several baselines, significant performance improvements, especially on rare interactions under each object, can be observed. As a side product, we achieve new state-of-the-art performance on the two datasets[3].

2 Related Work

2.1 Human-Object Interaction Detection

HOI detection [4] is challenging since it requires both precise detection and complex interaction reasoning capabilities. Existing methods have achieved some progress and mainly fall into two groups: two-stage and one-stage methods.

Two-stage methods adopt an off-the-shelf detector to perform detection, followed by an interaction prediction model over each human-object pair [4,13,47,33]. Previous approaches mostly endeavor to improve visual feature quality for interaction classification. For example, Qi *et.al.* [41] builds a holistic graph to assist information flow for all humans and objects, and Zhang *et.al.* [62] devises a bipartite graph utilizing relative spatial relation to promote interaction understanding. Besides, compositional models factorize the verb and object classification branches to improve generalization [31,21,22]. Beyond the visual appearance, more complementary cues are explored for the second stage, such as human pose and parts [30,38,18], language embeddings [26,2,57] and external knowledge [32,19].

One-stage methods perform both detection and interaction classification in an end-to-end manner. Besides detecting human and object regions, earlier one stage methods exploit either human-object interaction points [35,53] or their union regions [24] as interaction clues. With the success of Transformer[48] for object detection [3], some methods [5,68,25,44,60] present to formulate HOI detection as a set-prediction problem, where the anchor-free detection and attention-based global context aggregation are jointly operated.

Recently, some studies focus on the long-tail distribution problem in HOI detection benchmarks. For example, ATL [21] constructs new HOI instances from external object datasets in an affordance transfer fashion, while FCL [22] generates object features to fabricate more training samples. Besides, CDN [60] presents a dynamic re-weighting mechanism to tackle the long-tail problem. However, they mainly focus on the general long-tail distribution from the whole training set, leaving the *object bias problem* untouched in the literature.

[3] Code available: `https://github.com/daoyuan98/ODM`.

2.2 Bias Identification and Mitigation

Previous practices on the bias problem mainly follow an identification then mitigation paradigm. Pertaining to the bias identification, Zhao *et.al.* [63] finds that the gender bias contained in datasets can be further amplified by the model trained on them. Manjunatha *et.al.* [39] explicitly discovers the bias in Visual Question Answering [1] via association rule mining, while Guo *et.al.* [16] alleviates the bias through loss re-scaling. Lately, Li and Xu [34] unearths unknown biased attributes of a classifier with generative models. To mitigate the bias problem, adversarial training [11] is employed to learn bias irrelevant representations [61]. Recently, Wang *et.al.* [54] benchmarks previous mitigation methods and presents a combination of domain-conditional models for de-biasing, while Choi *et.al.* [8] tackles the unbalanced distribution with the weak supervision from a small reference dataset.

3 Object Bias Identification

The *object bias problem* in HOI detection refers to predicting interactions based on the unbalanced label distribution under each object. In the following, we demonstrate that the *object bias* problem comes from two aspects: (1) the conditionally unbalanced label distribution induced by objects and (2) the biased model training on the datasets.

3.1 Unbalanced Verb Distribution

The objective of HOI detection is to detect and classify $\langle \texttt{human}, \texttt{verb}, \texttt{object} \rangle$ triplets, where the most challenging and crucial part is verb classification.[4] Denote the whole verb set as $\mathcal{V}$, $\mathcal{V}_o$ for object o represents a subset of all verbs, i.e., $\mathcal{V}_o \in \mathcal{V}$ and $|\mathcal{V}_o| < |\mathcal{V}|$. We use $p(v|o)$ to represent the verb distribution conditioned on object o, and $p_o(v)$ to signify the global verb distribution involving only verbs for object o in the training set. The latter is employed to re-normalize the number of verbs associated with o, leaving other irrelevant verbs unaffected. From Fig. 2, we can observe that these two distributions are both skewed and actually different from each other. Besides, a globally frequent class can be a rare one after object conditioning and vice versa. For example, `hold` is the most frequent verb from a global view, while the object `vase` sees verb `make` most (Fig. 2). It thus brings our question: among these two long-tailed distributions, which one dominates more for the final verb classification?

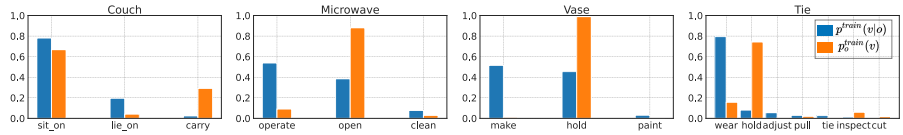

Fig. 2: Comparison between the object-conditional verb distribution $p^{train}(v|o)$ and the overall re-normalized distribution $p_o^{train}(v)$ for four objects in the training set of HICO-DET [4].

[4] With the detected object, verb classification is required to recognize the interaction.

3.2 Biased Model Learning

To delineate the second aspect, we exemplarily study the behavior of the state-of-the-art model SCG [62] on the HICO-DET dataset. Denote $\hat{y}(v|o)$ as the averaged verb score output by SCG conditioned on object o, and $p^{test}(v|o)$ as the counterpart of $p^{train}(v|o)$ on the test set. We compare them in Fig. 3 (a) and observe that $\hat{y}(v|o)$ pays less attention to conditionally rare classes in the training set (e.g., `paint` for `vase`, `clean` for `microwave`). In contrast, conditionally frequent classes (e.g., `operate` for `microwave`, `wear` for `tie` and `sit on` for `couch`) gain higher scores regardless of their prediction correctness. To quantify how much *bias* the model has learned, we compute the Jensen-Shannon Divergence [36] between $\hat{y}(v|o)$ and $p_o^{train}(v)$, $p^{train}(v|o)$, $p^{test}(v|o)$ and visualize them in Fig. 3 (b). We can see that $\hat{y}(v|o)$ is closer to $p^{train}(v|o)$ than $p_o^{train}(v)$, indicating the model leans towards the object-conditional statistics, rather than the overall label distribution in the training set. Besides, $\hat{y}(v|o)$ is even more similar to $p^{train}(v|o)$ than the ground-truth distribution $p^{test}(v|o)$. This implies, if we can counteract the learning of the *object bias*, there is a large potential of performance improvements with existing methods.

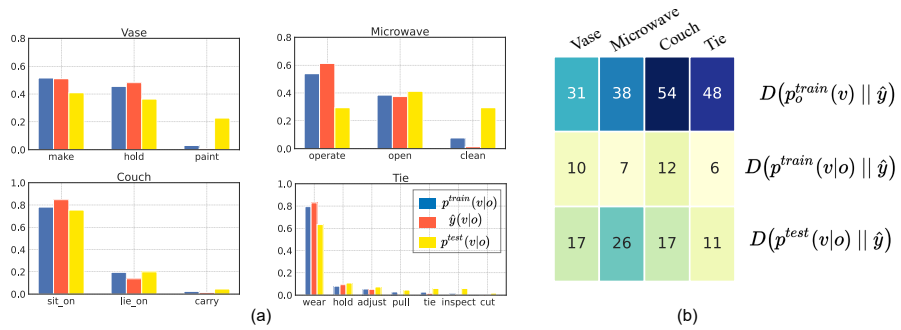

Fig. 3: (a) The model output analysis of SCG [62] on HICO-DET [4] test set with four objects. We show the difference between conditional training distribution $p^{train}(v|o)$, averaged model output $\hat{y}(v|o)$ and the ground-truth conditional verb distribution $p^{test}(v|o)$. (b) The Jensen-Shannon divergence [36] is utilized to compute the distribution distance. Note that the values are increased 100x for better illustration.

3.3 Comparison with Other Biases

Long-tail in HOI Detection We notice that some prior efforts [60,21,20] have studied the long-tail problem in HOI detection. Nevertheless, the *object bias problem* presented in this paper is intrinsically and technically different from the long-tail one. On the one hand, the object-conditional distribution can be distinct from the overall long-tail distribution. For example, `hold` is the most frequent verb across the whole dataset but less frequent in some objects (Fig. 2). On the other hand, the model prediction tends to conform with the object-conditional label distribution, rather than the overall one. Combing these two sides, we thus

introduce the *object bias problem* to the community and expect more insightful findings along this line.

Bias in Scene Graph Generation (SGG) It is worth noting that the bias problem in the sister task - SGG, is also different from the *object bias problem*. In fact, mainstream studies in SGG debiasing [58,7,46] mainly focus on the overall class imbalance, which is essentially same as the long-tail problem in HOI detection. The most relevant work to ours is [59]. It leverages the most frequent predicate under `subject-object` pair for relation prediction, which is shown to be a strong baseline on benchmark dataset. However, there are two key differences between [59] and our work: 1) [59] focuses on relational bias from the data's perspective only, while we provide a comprehensive study across the aspects of the dataset, model behavior and evaluation protocol. 2) [59] leverages the training set statistics to conduct prediction. However, when deploying the method to another dataset or other out-of-distribution settings, degraded performance is expected, as it severely overfits specific training set [6,46]. By contrast, we design a novel debiasing method to counteract the object bias during training, which is detailed in the following section.

4 Object Bias Alleviation

4.1 Problem Definition

Given an image, an HOI detection model is expected to detect each interactive triplet $\langle \texttt{human}, \texttt{verb}, \texttt{object} \rangle$ and output their interaction score $s^{h,v,o}$, which is calculated as $s^{h,v,o} = s^v \cdot s^h \cdot s^o$. s^h and s^o are the confidence scores for the detected person and object, respectively. They are often obtained from the confidence score output by the detector. s^v represents the verb score predicted by a classifier. In the following, we mainly consider the calculation of s^v and omit the upper-script v for notational convenience.

4.2 Base Model

In this work, we consider a generic HOI detection model, as shown in the left part of Fig. 4. It takes as input an image, detects all humans and objects, and links each human-object pair. Thereafter, with message passing or context aggregation, a set of human-object pair representations, i.e., the HOI features $\{\mathbf{x}_i^o\}_{i=1}^N$ are obtained, where N denotes the number of human-object pairs. Each feature $\mathbf{x}_i^o$ captures the interaction relation between a human and an object of class o.

We then feed these features into a classifier f_b to predict verb scores: $\mathbf{s}_i = \sigma(f_b(\mathbf{x}_i^o))$, where $\sigma(\cdot)$ is a sigmoid function. Note that there can be multiple or no interactions within one human-object pair. Thus, the verb recognition is usually formulated as a multi-label classification problem. The objective of the base model is formulated as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_b^{bce}(\mathbf{s}_i, \mathbf{v}_i^o) + \mathcal{L}_{aux}, \tag{1}$$

where $\mathbf{v}_i^o$ denotes the ground-truth label involving object o, $\mathcal{L}_b^{bce}$ is the binary cross entropy loss for verb classification and $\mathcal{L}_{aux}$ corresponds to other objectives

of the base model such as interactiveness prediction and object localization. As discussed before, the base model often severely suffers from the *object bias problem*. To overcome this issue, we design a novel Object-wise Debiasing Memory module which has minimal influence to the reasoning process of the base model and is plugable to any existing HOI detection methods.

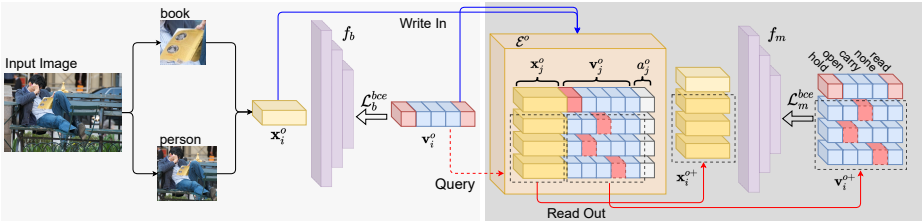

Fig. 4: Overview of the proposed method. Given an image, an HOI detection model extracts HOI features for each human-object pair. A memory cell $\mathcal{E}^o$ is maintained for each object o. During training, instances are conditionally read and written into its respective cell with label-awareness. We show one human object pair ($o = \texttt{book}$) for clearance.

4.3 Object-wise Debiasing Memory

It is widely accepted that instances from rare classes contain richer information for interaction understanding [45,55,23]. However, as discussed in Sec. 3, frequent verb classes under each object dominate the prediction results, while other informative but rare ones are often ignored. In view of this, we propose to re-sample HOI instances with a re-balancing strategy. In general, re-sampling has been shown to be an effective technique for class unbalance mitigation [55,43,28]. However, in HOI detection, it is infeasible to directly apply these techniques. On the one hand, the object bias problem is induced by object-conditional unbalance, rather than the overall one, which is distinct from the traditional class-imbalance scenario. On the other hand, there can be multiple human-object interactions within a single image, simply re-sampling one image with rare classes may lead to oversampling of non-rare ones, which may further exacerbates the *object bias problem*.

To circumvent this, we resort to the fine-grained feature-level re-sampling during model training. Accordingly, we maintain a memory for each object, on which an effective read and write strategy is devised to operate. We name this module Object-wise Debiasing Memory (ODM) and the framework is illustrated in Fig. 4. Specifically, a memory cell $\mathcal{E}^o$ is maintained for each object o, which has a fixed size n and stores three types of elements: the HOI feature $\mathbf{x}_j^o$, the verb label $\mathbf{v}_j^o$ and the feature generation time a_j^o. During training, each ODM cell is sampled (read out) with label awareness, followed by a dynamic update (write in) operation to ensure feature consistency. The pseudo-code for read and write strategy is shown in Alg. 1 and detailed as follows.

Read Strategy To achieve verb balance under each object, it makes sense to assign high sampling priority to rare class instances. At each training step,

Algorithm 1 Read and Write Strategy for $\mathcal{E}^o$

// Read Strategy
Input: HOI instance $\{\mathbf{x}, \mathbf{v}\}$, number of required samples k
Output: k HOI instances and labels
  features = [$\mathbf{x}$]; labels = [$\mathbf{v}$]
  while number of sampled features $< k$ **do**
    // pick rare class entries when not selected
    $j = argmax_j \ \sum_i dist(\text{labels[i]}, \mathcal{E}^o[j])$ (Eq. 2)
    Append $\mathbf{x}_j, \mathbf{v}_j$ to features, labels
  return features, labels
// Write Strategy
Input: HOI instance {$\mathbf{x}$, $\mathbf{v}$ }, generation time a
  if $\mathcal{E}^o$ is not full **then**
    Append $\{\mathbf{x}, \mathbf{v}, a\}$ to $\mathcal{E}^o$
  else:
    if $score(\mathbf{v}^o) \geq \tau^o$ (Eq. 3) **then**
      Replace entry of the longest duration with $\{\mathbf{x}, \mathbf{v}, a\}$.

given an interactive HOI feature $\mathbf{x}_i^o$ with verb label $\mathbf{v}_i^o$, we take $\mathbf{v}_i^o$ as query and sample a set of k HOI instances $\{\mathbf{x}_j^o, \mathbf{v}_j^o\}_{j=1}^k$ from the memory $\mathcal{E}^o$ such that the label distribution after sampling is less skewed. To that end, we select from the memory with the largest weighted hamming distance, which is calculated as:

$$dist(\mathbf{v}_1^o, \mathbf{v}_2^o) = \sum_{t=1}^{c} w_t^o \cdot (\mathbf{v}_1^o[t] \oplus \mathbf{v}_2^o[t]), \tag{2}$$

where $\oplus$ means XOR operation, $[\cdot]$ is subscription and w_t^o is a weighting coefficient of the t-th class associated with object o. Firstly, the hamming distance is employed to consider absent classes with respect to selected instance [14]. Secondly, the weighting mechanism ensures dynamic control over certain classes. Specifically, we calculate w_t^o as $N_o/N_{v,o}$, where N_o and $N_{v,o}$ denotes the number of object o and interaction $\langle v, o \rangle$ in the training set. By designing w_t^o as inverse interaction frequency within object o, rare class instances are prioritized and thus more frequently sampled from the memory. In addition, we perform iterative sampling to avoid all selected samples are from the same class.

Write Strategy During the writing stage, it is expected to store more rare class instances to ensure the sample complexity for memory reading. Specifically, we treat one instance as write-feasible if its hamming score for a multi-hot label is greater than a threshold τ^o. The hamming score is given by:

$$score(\mathbf{v}_j^o) = \sum_{t=1}^{c} w_t^o \cdot \mathbf{v}_j^o[t], \tag{3}$$

where c is the number of verb classes and w_t^o is the same weighting coefficient as that in Eq. 2. With this strategy, non-rare instances will not be written into the memory, thereby alleviating the risk of their dominance for model training.

When the memory is full, we replace the feature of the longest duration with write-feasible instances, so as to ensure timely update of memory contents.

4.4 Training and Inference

The proposed memory operations serve as an ad-hoc re-sampling approach to ensure more balanced training at each iteration. After reading from the memory, we then leverage another classifier f_m to perform more balanced interaction classification. Inspired by recent work on class-imbalanced learning [66,51,23], we combine f_m with the base classifier f_b to achieve a trade-off between the debiasing and representation capability. The overall objective is defined as follows:

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_b^{bce}(\mathbf{s}_i, \mathbf{v}_i^o) + \mathcal{L}_m^{bce}(\mathbf{s}_i^+, \mathbf{v}_i^{o+}) + \mathcal{L}_{aux}, \tag{4}$$

where $\mathbf{x}_i^{o+} = [\mathbf{x}_i^o; \{\mathbf{x}_j^o\}_{j=1}^k]$ and $\mathbf{v}_i^{o+} = [\mathbf{v}_i^o; \{\mathbf{v}_j^o\}_{j=1}^k]$ are obtained after the read operation and $\mathbf{s}_i^+ = f_m\big(\mathbf{x}_i^{o+}\big)$.

During inference, given an HOI feature $\mathbf{x}_i$, we take the weighted combination of these two classifiers' output as the final prediction:

$$\hat{\mathbf{v}}_i = \sigma\big(\lambda f_b(\mathbf{x}_i) + (1-\lambda) f_m(\mathbf{x}_i)\big), \tag{5}$$

where λ is a hyper-parameter balancing the two classifiers.

5 Experiments

5.1 Experimental Setting

Dataset We conducted experiments on two benchmarks: HICO-DET [4] and HOI-COCO [21]. **HICO-DET** is the most widely employed benchmark in HOI detection. It consists of 38,118 and 9,658 images in the training and test set, respectively. HICO-DET covers the whole 80 object classes in MS-COCO [37] and 117 verb classes, resulting in a total of 600 HOI categories in the form of $\langle$`person`, `verb`, `object`$\rangle$. **HOI-COCO** is a recently introduced dataset based on V-COCO [17]. It has a total of 9,915 images, with 4,969 for training and 4,946 for test. There are 222 HOI categories composed of 21 verb classes from V-COCO and 80 MS-COCO object classes.

Baselines As our goal is to prove the superiority and versatility of the proposed method, we applied our approach to three existing methods: HOID [50], SCG [62] and QPIC [44]. **HOID** generates human-centric object proposal for interactive objects only. **SCG** is a recently proposed two-stage method leveraging spatial information for graph-based message propagation. It achieves state-of-the-art performance with both fine-tuned and ground-truth detection among two-stage methods. **QPIC** is an advanced one-stage method, which utilizes Transformer architecture to perform query-based detection and classification.

Standard Evaluation Metrics We followed the standard evaluation setting [4] and reported mean average precision (mAP) for both datasets, where the mAP on rare (less than 10 training instances), non-rare and full classes are

Table 1: Performance comparison under object-bias setting on HICO-DET. OR and ONR denote **O**bject-**R**are and **O**bject-**N**on**R**are, respectively.

Detector	*Pre-trained Detector*				*Fine-tuned Detector*				*Oracle Detector*	
Method	HOID [50]	+Ours	SCG [62]	+Ours	QPIC [44]	+Ours	SCG	+Ours	SCG	+Ours
OR	17.05	**19.02**	18.38	**19.47**	26.29	**26.96**	28.67	**30.21**	51.03	**52.48**
ONR	24.24	**24.33**	**25.06**	25.01	34.64	**34.65**	40.72	**41.08**	73.97	**75.43**
AVE	20.65	**21.17**	21.72	**22.24**	30.47	**30.81**	34.69	**35.64**	62.50	**63.95**

reported. For both settings, a prediction is regarded as positive if (1) the HOI classification is correct and (2) the detected human and object bounding boxes have IoUs greater than 0.5 with the ground-truth bounding box.

Object-bias Evaluation Metric To quantitatively study how much object bias has been alleviated by our method, we propose a new *object bias* evaluation setting. Specifically, we treated an interaction class as *object rare* (*object non-rare*) if $N_{v,o}/N_o < (\geq)\ \alpha$. On HICO-DET dataset, we set α to 0.3 based on its statistics. Note that an originally non-rare class in the whole training set can be *object rare* under this setting. For each object, we computed the mean of Average Precision (AP) for *object rare* and *object non-rare* classes, respectively. After that, we averaged across all objects to obtain mean Average Precision for **O**bject-**R**are (OR) and **O**bject-**N**on**R**are (ONR), respectively. Besides, their **AVE**rage is also reported. Different from the traditional evaluation, the *object bias* evaluation protocol considers the performance within each object and thus offers a better test bed for quantifying a model's ability to overcome the *object bias* problem.

5.2 Object Bias Evaluation

The results under the new *object bias* setting are shown in Tab. 1. With SCG as baseline, our method significantly improves *object rare* classes by a clear margin of +**1.09** mAP, +**1.54** mAP and +**1.45** mAP under three detection settings. Besides, the proposed method also boosts HOID and QPIC by +**0.97** mAP and +**0.67** mAP under OR setting. This provides evidence that our method can effectively alleviate the object bias problem. Notably, the proposed module can also improve ONR classes in most cases.

5.3 Standard Evaluation

Results on HICO-DET We followed [62] to report the results with *detector pre-trained on MS-COCO [37]* (HOID and SCG), *detector find-tuned on HICO-DET* (SCG and QPIC) and *oracle detector* (SCG). The results can be found in Tab. 2, 3 and 4, respectively.

Our method improves the performance of all three baseline methods across all detection settings. For instance, with pre-trained detector, our method promotes HOID and SCG by +**0.89** and +**1.29** mAP on rare classes, respectively, which amounts to **6**% and **8**% relative improvements. When leveraging fine-tuned detector, the proposed approach can improve QPIC and SCG on rare classes by +**0.52** and +**0.81** mAP. In particular, with the detection quality improved,

Table 2: Results on HICO-DET with *pre-trained* detector.

Method	Full	Rare	Non-rare
iCAN [13]	14.84	10.45	16.15
TIN [32]	17.03	13.42	18.11
DRG [12]	19.26	17.74	19.71
VCL [20]	19.43	16.55	20.29
ACP [26]	20.59	15.92	21.98
DJ-RN [30]	21.34	**18.53**	22.18
HOID* [50]	19.58	15.29	20.96
+Ours	20.45	16.18	21.73
SCG* [62]	20.99	16.30	22.40
+Ours	**21.50**	17.59	**22.67**

Table 3: Results on HICO-DET with *fine-tuned* detector.

Method	Full	Rare	Non-rare
PPDM [35]	21.73	13.78	24.10
HOI-Trans [68]	23.46	16.91	25.41
ATL [21]	27.68	20.31	29.89
AS-Net [5]	28.87	24.25	30.25
FCL [22]	29.12	23.67	30.75
QPIC* [44]	29.04	21.55	31.27
QPIC + Ours	29.26	22.07	31.41
SCG* [62]	31.08	24.14	33.15
SCG + Ours	**31.65**	**24.95**	**33.65**

our method also enhances non-rare classes by a noticeable margin. Lastly, with the oracle detector, the proposed method can advance SCG on both rare (+**1.22** mAP) and non-rare classes (+**1.26** mAP). These results demonstrate the superiority of the proposed method. As a side product, we achieve new state-of-the-art on the HICO-DET dataset.

Table 4: Results on HICO-DET with *oracle detector*.

Method	Full	Rare	Non-rare
iCAN [13]	33.38	21.43	36.95
TIN [32]	34.26	22.90	37.65
Peyre *et al.* [40]	34.35	27.57	36.38
FCL [22]	44.26	35.46	46.88
SCG* [62]	51.03	38.93	54.65
SCG + Ours	**52.29**	**40.15**	**55.91**

Table 5: Results on the HOI-COCO. * indicates reproduced baseline.

Method	Full	Rare	Non-rare
Baseline [21]	22.86	6.87	35.27
+VCL [20]	23.53	8.29	35.36
+ATL [21]	23.40	8.01	35.34
Baseline*	22.87	6.98	35.21
+CDN [60]	23.15	7.25	**35.49**
+Ours	**23.73**	**8.58**	**35.49**

Results on HOI-COCO We followed [21] to provide results with MS-COCO pre-trained detector, which is the most typical setting for two-stage methods. The results are shown in Tab. 5. For fair comparison, we reproduced the baseline method used in ATL [21]. We also compared with the debiasing technique applied in CDN [60], which aims to alleviate the general long-tail problem. It can be observed that our method outperforms these debiasing methods on this relatively small scale dataset, especially for rare classes.

5.4 Ablation Studies

We studied the effectiveness of our proposed method. All experiments are conducted on HICO-DET dataset with the SCG [62] baseline, and evaluated under both standard protocol and the proposed *object bias* setting.

Comparison with other Debiasing Methods We compared our method with various debiasing methods in Tab. 6. The competitors include loss reweighting methods, general debiasing methods and Scene Graph Generation

Table 6: Comparsion with debiasing methods.

Type	Method	**Full**	**Rare**	**Non-rare**	**OR**	**ONR**	**AVE**
	Baseline	20.99	16.30	22.40	18.38	**25.06**	21.70
Reweighting	+inv. freq.	17.58	14.15	18.61	9.77	21.01	15.39
	+CB-Loss(0.9999) [9]	14.30	13.54	14.53	9.93	21.48	15.71
	+CB-Loss(0.999) [9]	13.34	12.96	13.45	9.02	20.73	14.88
	+CB-Loss(0.99) [9]	13.98	13.20	14.21	9.46	20.98	15.22
General Debiasing	+AT [54]	20.49	16.22	21.77	18.12	24.46	21.29
	+DIT [54]	18.13	16.99	18.47	17.35	23.05	20.20
SGG Debiasing	+TDE [46]	20.44	14.89	22.10	18.30	24.44	21.37
	+PCPL [58]	16.93	12.95	18.12	15.04	24.27	19.65
	+Ours	**21.50**	**17.59**	**22.67**	**19.47**	25.01	**22.24**

Table 7: Performance of different classifiers on HICO-DET.

Detector	**Classifier**	**Full**	**Rare**	**Non-rare**	**OR**	**ONR**	**AVE**
Pre-trained on MS-COCO	f_b	21.08	16.66	22.40	19.00	24.38	21.69
	f_m	20.79	16.50	22.08	18.59	24.87	21.73
	$full$	**21.50**	**17.59**	**22.67**	**19.47**	**25.01**	**22.24**
Fine-tuned on HICO-DET	f_b	31.24	24.77	33.17	30.04	40.62	35.33
	f_m	30.60	23.30	32.77	29.43	40.79	35.11
	$full$	**31.65**	**24.95**	**33.65**	**30.21**	**41.08**	**35.64**
Oracle	f_b	51.24	39.27	54.82	51.81	74.73	63.27
	f_m	51.09	37.94	55.01	51.28	75.15	63.21
	$full$	**52.29**	**40.15**	**55.91**	**52.48**	**75.43**	**63.95**

(SGG) debiasing methods. We observed all these methods degrade the original baselines. This may be related to the strong interference with the original training process. Besides, some methods are designed to tackle the globally long-tail problem and single-label classification, thus incapable of resolving the object-conditional long-tail problem in HOI detection.

Efficacy of Classifiers The distinctive importance of the verb classifier in the base model (f_b), the one trained with ODM (f_m) and the full classifier ($\lambda f_b + (1-\lambda) f_m$) are explored in this experiment. From the results in Tab. 7, we see that for all three detectors, f_m is inferior to f_b on both evaluation protocols. However, when combining these two together, the final performance can be further promoted. This is mainly because these two classifiers focus on different classes and are in fact complementary to each other.

5.5 Visualizations

Effects of Memory We studied how the proposed ODM alleviates the distribution imbalance and illustrated the evolution of verb distribution after reading from the memory in Fig. 5. With these examples, we can conclude that our method can effectively address the label imbalance problem under each object.

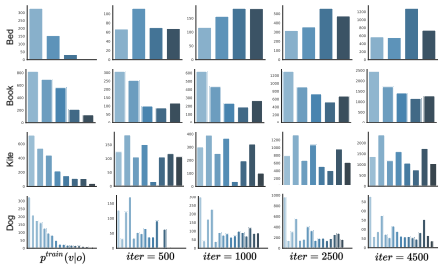

Fig. 5: The evolution of accumulated verb distribution after reading from the proposed ODM for 4 randomly selected objects. The leftmost column shows $p^{train}(v|o)$ and the other 4 columns represent the sampled verb distribution at different iterations.

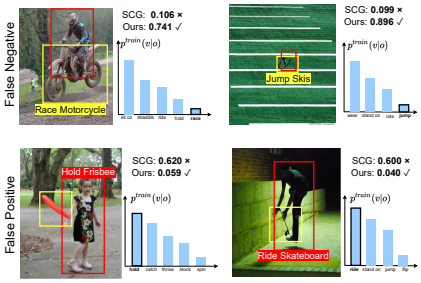

Fig. 6: False negative (top two) and false positive (bottom two) instances from the SCG baseline on HICO-DET test set. For each instance, the $p^{train}(v|o)$ is also shown, where the involved verb is bold by a rectangular.

Especially, at the 2500-*th* iteration, the verb distribution is already less skewed, which remains stable till the end of this epoch (∼4.5k iterations).

Qualitative Results We show some qualitative results in Fig. 6. For the two false negative instances (top two), the baseline model assigns low score to ground-truth interactions, wherein both involved verbs are conditionally rare in the training set (`race` for `motorcycle`, `jump` for `skis`). For the false positive instances (bottom two), the baseline favors more frequent verbs (`hold` for `frisbee`, `ride` for `skateboard`), though the interaction prediction is incorrect. In contrast, our method can overcome these two kinds of errors and achieve better performance.

6 Conclusion and Future Work

In this work, we systematically studied the *object bias* problem in Human-Object Interaction detection. We demonstrated the recognition of this problem from the aspects of unbalanced label distribution and biased model learning, and advocated a new protocol to comprehensively evaluate model performance. To reduce the heavily skewed label distribution under each object, we proposed an Object-wise Debiasing Memory to facilitate balanced sampling of HOI instances. Extensive experiments validate the effectiveness of the proposed method, demonstrating that it can significantly alleviate the *object bias problem* and outperform advanced baselines with large margins. Due to the universal existence of the *bias problem*, in the future, we plan to explore identifying bias factors in other related tasks such as visual relation detection and scene graph generation.

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: ICCV (2015) 1, 5
2. Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. In: AAAI (2020) 4
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV (2020) 2, 4
4. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: WACV (2018) 3, 4, 5, 6, 10
5. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating hoi detection as adaptive set prediction. In: CVPR (2021) 2, 4, 12
6. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: CVPR (2019) 7
7. Chiou, M.J., Ding, H., Yan, H., Wang, C., Zimmermann, R., Feng, J.: Recovering the unbiased scene graphs from the biased ones. In: ACM MM (2021) 7
8. Choi, K., Grover, A., Singh, T., Shu, R., Ermon, S.: Fair generative modeling via weak supervision. In: ICML (2020) 5
9. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: CVPR (2019) 13
10. Dong, Q., Tu, Z., Liao, H., Zhang, Y., Mahadevan, V., Soatto, S.: Visual relationship detection using part-and-sum transformers with composite queries. In: ICCV (2021) 2
11. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: ICML (2015) 5
12. Gao, C., Xu, J., Zou, Y., Huang, J.B.: DRG: Dual relation graph for human-object interaction detection. In: ECCV (2020) 2, 12
13. Gao, C., Zou, Y., Huang, J.B.: iCAN: Instance-centric attention network for human-object interaction detection. In: BMVC (2018) 4, 12
14. Gordo, A., Perronnin, F., Gong, Y., Lazebnik, S.: Asymmetric distances for binary embeddings. IEEE TPAMI (2013) 9
15. Guo, Y., Cheng, Z., Nie, L., Liu, Y., Wang, Y., Kankanhalli, M.: Quantifying and alleviating the language prior problem in visual question answering. In: SIGIR (2019) 1
16. Guo, Y., Nie, L., Cheng, Z., Tian, Q., Zhang, M.: Loss re-scaling vqa: revisiting the language prior problem from a class-imbalance view. IEEE TIP (2021) 5
17. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015) 10
18. Gupta, T., Schwing, A., Hoiem, D.: No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: ICCV (2019) 2, 4
19. He, T., Gao, L., Song, J., Li, Y.F.: Exploiting scene graphs for human-object interaction detection. In: ICCV (2021) 4
20. Hou, Z., Peng, X., Qiao, Y., Tao, D.: Visual compositional learning for human-object interaction detection. In: ECCV (2020) 2, 6, 12
21. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for human-object interaction detection. In: CVPR (2021) 2, 3, 4, 6, 10, 12
22. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Detecting human-object interaction via fabricated compositional learning. In: CVPR (2021) 4, 12
23. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: ICLR (2020) 8, 10

24. Kim, B., Choi, T., Kang, J., Kim, H.J.: UnionDet: Union-level detector towards real-time human-object interaction detection. In: ECCV (2020) 4
25. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: HOTR: End-to-end human-object interaction detection with transformers. In: CVPR (2021) 2, 4
26. Kim, D.J., Sun, X., Choi, J., Lin, S., Kweon, I.S.: Detecting human-object interactions with action co-occurrence priors. In: ECCV (2020) 4, 12
27. Kim, D.J., Sun, X., Choi, J., Lin, S., Kweon, I.S.: Acp++: Action co-occurrence priors for human-object interaction detection. IEEE TIP (2021) 2
28. Li, Y., Vasconcelos, N.: Repair: Removing representation bias by dataset resampling. In: CVPR (2019) 3, 8
29. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017) 1
30. 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) 2, 4, 12
31. Li, Y.L., Liu, X., Wu, X., Li, Y., Lu, C.: HOI analysis: Integrating and decomposing human-object interaction. In: NeurIPS (2020) 4
32. 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) 2, 4, 12
33. Li, Y., Liu, X., Wu, X., Huang, X., Xu, L., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. IEEE TPAMI (2021) 2, 4
34. Li, Z., Xu, C.: Discover the unknown biased attribute of an image classifier. In: ICCV (2021) 5
35. 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) 2, 4, 12
36. Lin, J.: Divergence measures based on the shannon entropy. IEEE Trans. Inf. Theory (1991) 6
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 10, 11
38. Liu, Y., Chen, Q., Zisserman, A.: Amplifying key cues for human-object-interaction detection. In: ECCV (2020) 4
39. Manjunatha, V., Saini, N., Davis, L.S.: Explicit bias discovery in visual question answering models. In: CVPR (2019) 5
40. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: ICCV (2019) 12
41. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018) 2, 4
42. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015) 2
43. Shen, L., Lin, Z., Huang, Q.: Relay backpropagation for effective learning of deep convolutional neural networks. In: ECCV (2016) 3, 8
44. Tamura, M., Ohashi, H., Yoshinaga, T.: QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In: CVPR (2021) 2, 4, 10, 11, 12
45. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: NeurIPS (2020) 3, 8
46. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: CVPR (2020) 7, 13

47. Ulutan, O., Iftekhar, A., Manjunath, B.S.: VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In: CVPR (2020) 2, 4
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017) 4
49. Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: ICCV (2019) 2
50. Wang, S., Yap, K.H., Yuan, J., Tan, Y.P.: Discovering human interactions with novel objects via zero-shot learning. In: CVPR (2020) 10, 11, 12
51. Wang, T., Li, Y., Kang, B., Li, J., Liew, J., Tang, S., Hoi, S., Feng, J.: The devil is in classification: A simple framework for long-tail instance segmentation. ECCV (2020) 10
52. Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J.: Deep contextual attention for human-object interaction detection. In: ICCV (2019) 2
53. Wang, T., Yang, T., Danelljan, M., Khan, F.S., Zhang, X., Sun, J.: Learning human-object interaction detection using interaction points. In: CVPR (2020) 2, 4
54. 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) 5, 13
55. Wu, T., Huang, Q., Liu, Z., Wang, Y., Lin, D.: Distribution-balanced loss for multi-label classification in long-tailed datasets. In: ECCV (2020) 3, 8
56. Xu, B., Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Interact as you intend: Intention-driven human-object interaction detection. IEEE TMM (2019) 2
57. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: CVPR (2019) 2, 4
58. 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) 7, 13
59. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR (2018) 7
60. Zhang, A., Liao, Y., Liu, S., Lu, M., Wang, Y., Gao, C., Li, X.: Mining the benefits of two-stage and one-stage hoi detection. In: NeurIPS (2021) 2, 4, 6, 12
61. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: AIES (2018) 5
62. Zhang, F.Z., Campbell, D., Gould, S.: Spatially conditioned graphs for detecting human-object interactions. In: ICCV (2021) 2, 4, 6, 10, 11, 12
63. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: EMNLP (2017) 5
64. Zhong, X., Ding, C., Qu, X., Tao, D.: Polysemy deciphering network for human-object interaction detection. In: ECCV (2020) 2
65. Zhong, X., Qu, X., Ding, C., Tao, D.: Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In: CVPR (2021) 2
66. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: CVPR (2020) 10
67. Zhou, P., Chi, M.: Relation parsing neural network for human-object interaction detection. In: ICCV (2019) 2

68. 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) 2, 4, 12