Towards Overcoming False Positives in Visual Relationship Detection

Daisheng Jin*[1][2]
jindaisheng@buaa.edu.cn
Xiao Ma*[3]
xiao-ma@comp.nus.edu.sg
Chongzhi Zhang[1][2]
chongzhizhang@buaa.edu.cn
Yizhuo Zhou[5]
zhou.yizhuo@bytedance.com
Jiashu Tao[3]
jiashut@comp.nus.edu.sg
Zhoujun Li[2]
lizj@buaa.edu.cn
Mingyuan Zhang†[4]
mingyuan001@e.ntu.edu.sg

[1] SenseTime, Inc.
Beijing, CN
[2] Beihang University
Beijing, CN
[3] National University of Singapore
Singapore, SG
[4] Nanyang Technological University
Singapore, SG
[5] ByteDance Ltd.
Beijing, CN

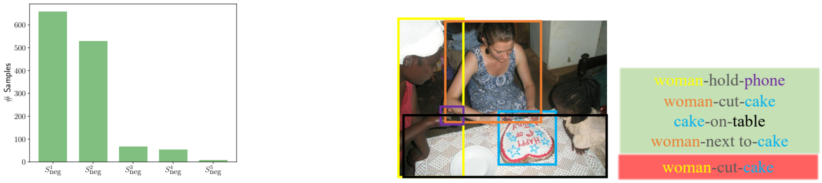
Abstract

In this paper, we investigate the cause of the high false positive rate in Visual Relationship Detection (VRD). We observe that during training, the relationship proposal distribution is highly imbalanced: most of the negative relationship proposals are easy to identify, e.g., the inaccurate object detection, which leads to the under-fitting of low-frequency difficult proposals. This paper presents *Spatially-Aware Balanced negative pRoposal sAmpling* (SABRA), a robust VRD framework as a proof of concept that alleviates the influence of false positives. To effectively optimize the model under imbalanced distribution, SABRA adopts *Balanced Negative Proposal Sampling* (BNPS) strategy for mini-batch sampling. BNPS divides proposals into 5 well-defined sub-classes and generates a balanced training distribution. To further resolve the low-frequency challenging false positive proposals with high spatial ambiguity, we adopt a spatial learning module that implicitly imposes the object-centric spatial configuration with a spatial mask decoder, using the global spatial features extracted with Graph Neural Networks. SABRA is conceptually simple and outperforms SOTA methods by a large margin on two human-object interaction (HOI) datasets and one general VRD dataset.

1 Introduction

Visual Relationship Detection (VRD) is an important visual task that bridges the gap between middle-level visual perception, e.g., object detection, and high-level visual understanding,

* equal contribution, † corresponding author



(a) Imbalanced negative proposal distribution (b) Influence of ambiguous context

Figure 1: (a) VRD normally gives an extremely imbalanced negative proposal distribution, where $S^{1:5}_{\text{neg}}$ denotes 5 different types of negative proposals. Detailed definition of $S^{1:5}_{\text{neg}}$ is given by Sec. 3.1. (b) Contextual objects introduce ambiguous information to the relationship classification, e.g., the model thinks the woman on the left is cutting the cake by mistake.

e.g., image captioning [7, 38], and visual question answering [28]. General VRD aims to understand the interaction between two arbitrary objects in the scene. Human-Object Interaction (HOI), as a specific case of VRD, focuses on understanding the interaction between humans and objects, e.g., woman-cut-cake.

Existing VRD methods focus on building powerful feature extractors for each [*subject*, *object*] pair, predicting the *predicate* between the *subject* and the *object*, and outputting [*subject*, *predicate*, *object*] triplet predictions. Some prior works model subject and object relationship independently [6, 32], which loses the global context and is susceptible to inaccurate detections. Some recent works incorporate the union bounding boxes of the [*subject*, *object*] as additional features to provide additional spatial information [31, 39] or use the graph neural networks (GNNs) to better extract global object relationships [16, 30, 48]. For all methods, they train the model with a dataset consists of both positive and negative [*subject*, *object*] proposal pairs. However, such a scheme often leads to a high false positive prediction rate, *i.e.*, many negative proposals are classified as positive (Fig. 1.b).

In this work, we aim to demystify the cause of the high false positive rate in VRD. We observe that the negative [*subject*, *object*] proposals form an imbalanced distribution which leads to a difficult optimization landscape [33]. This is because most of the negative samples are caused by simple yet prevalent inaccurate object detections, but the challenging incorrect [*subject*, *object*] associations, caused by ambiguous contexts, only contribute to a small portion of the negative proposals. As a result, the imbalance negative proposal distribution makes learning accurate predictions for the hard but rare negative proposals difficult.

We present *Spatially-Aware Balanced negative pRoposal sAmpling* (SABRA), a robust and general VRD framework as a proof of concept that alleviates the influence of false positives for both HOI and general VRD tasks. According to the two sources of negative proposals, *i.e.*, inaccurate detections, and incorrect associations, we introduce a division of negative proposals into 5 sub-classes $S^{1:5}_{\text{neg}}$. For each single object detection, we consider (1) if it is an accurate detection, and (2) if it is in any relationship with a different object. Specifically, in Fig. 1.a, $S^{1:2}_{\text{neg}}$ cover the inaccurate detections, and $S^{3:5}_{\text{neg}}$ discuss the incorrect associations. From S^{1}_{neg} to S^{5}_{neg}, the sample size decreases, and the classification difficulty increases, because detecting the false positives according to the accuracy of detection is no longer sufficient, and careful understanding of object relationships becomes necessary for the task. As visualized in Fig. 1.a, the sample sizes of 5 negative proposal sub-classes give a highly imbalanced distribution, which degrades the performance of data-driven VRD algorithms.

Inspired by the learning under imbalanced distribution literature [3, 43], we alleviate the optimization difficulty by *Balanced Negative Proposal Sampling* scheme. BNPS computes the statistics of each class and performs a simple yet effective *Class Balanced Sampling* [34] for balanced data distribution. Balanced negative proposal sampling significantly reduces the number of false positive occurrences. For low-frequency difficult classes $S_{\text{neg}}^{3:5}$, e.g., Fig. 1.b, we further improve the spatial modeling of SABRA on two aspects. For the global context understanding, SABRA extends the existing GNN-based methods [30, 46] with a heterogeneous message passing scheme that effectively addresses the distribution divergence between different features. For local spatial configuration, SABRA learns a position-aware embedding vector by predicting the locations in each [*subject*, *object*] pair.

We evaluate SABRA on both HOI (V-COCO and HICO-DET) and general visual relationship detection (VRD). We show that SABRA significantly outperforms SOTA methods. We also visualize the results and show that SABRA effectively reduces the false positives in VRD and misclassification in terms of spatial ambiguity.

2 Related Works

Visual Relationship Detection. VRD is an important middle-level task bridging low-level visual recognition with high-level visual understanding. With the advances of deep learning, data-driven approaches are widely adopted for VRD. Specifically, *Convolutional Neural Networks* (CNNs) are used for automatic feature extraction and information fusion, which achieved great improvement in VRD [39, 42, 45]. *Graph Neural Networks* (GNNs) further improve the feature extraction process by explicitly modeling the instance-wise interactions between objects [30, 46]. Due to the nature of VRD tasks, additional information has been introduced as auxiliary training signals, such as language priors [26], prior interactiveness knowledge of objects [22], and action co-occurrence knowledge [20]. In comparison with the existing methods, SABRA is the first to identify the importance of false positives in VRD tasks and has significantly outperformed SOTA methods in our experiments.

Learning under Imbalanced Distribution. Real-world data is imbalanced by nature: a few high-frequency classes contribute to most of the samples, while a large number of low-frequency classes are under-represented in data. Standard imbalanced learning techniques include data re-balancing [1, 2, 3], loss function engineering [4, 14] and meta-learning [18]. VRD, as a common computer vision task, also suffers from the imbalanced problem [20, 25]. [20] considers the imbalance of relationship imbalance, i.e., the imbalance of positive samples, and uses action co-occurrence to provide additional labels. However, none of the existing methods consider the imbalance of the negative [*subject*, *object*] proposals, which commonly exists in VRD settings. By re-balancing the proposals, SABRA significantly improves the overall performance of VRD algorithms.

Spatial Information. Spatial information is key to understanding the relationship between objects. Prior methods fuse spatial information with positional embedding, which normalizes the absolute or relative coordinates of the subject, object, and union bounding boxes [52]. However, simple positional embedding implicitly captures spatial information with position coordinates as inputs to networks, which is unable to capture the explicit spatial configuration in the feature space. [11] introduces binary masks which explicitly specify the subject and object positions, implicitly specifying the spatial configuration by concatenating with union features. With the recent advances in graph neural networks, the relevant positional information can be captured by message passing between instances [54]. The implicit message passing, nevertheless, loses the contextual grounding of the [*subject*, *object*] pair. In contrast, the spatial mask decoder learns to predict the positional information of the [*subject*, *object*] , while capturing the relevant features by end-to-end learning.

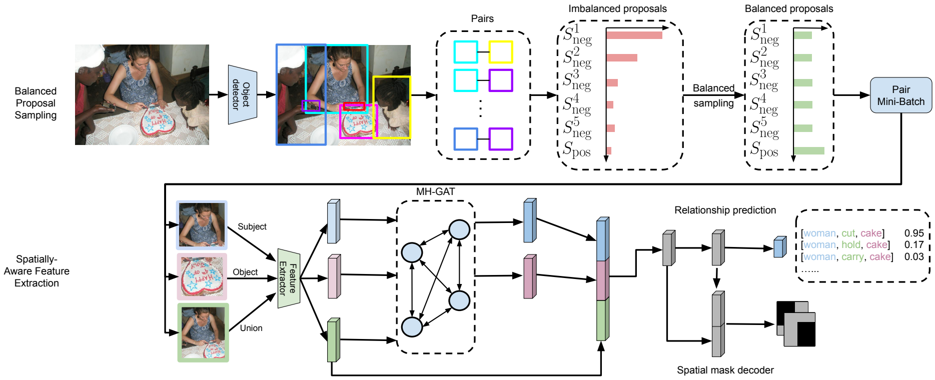


Figure 2: SABRA classifies the negative [subject, object] proposals into 5 sub-categories $S^{1:5}_{\text{neg}}$ and use Balanced Sampling to re-balance the categorical distribution $\text{Cat}(S_{\text{pos}}, S^{1:5}_{\text{neg}})$. To reduce the false positive relationship predictions caused by spatial ambiguity, SABRA first uses MH-GAT to capture the global context and then learns to predict the spatial mask of [subject, object] in the ROI pooled feature space with a spatial mask decoder.

3 SABRA

We observe that there are two major causes to the hard false positive predictions. Firstly, the relationship classifier is trained under an extremely imbalanced distribution, which contains few hard negative samples but numerous easy negative samples. As a result, it gives less training signal to hard examples and hinders the final performance. Besides, these samples are often caused by spatial ambiguity, which requires us to embed spatial relations into the proposed classifier. To this end, we propose a balanced negative proposal sampling strategy and spatially-aware embedding modules to enhance visual feature learning.

The pipeline of SABRA is shown in Fig. 2. The input image is fed into an object detector to predict all bounding boxes, which are exhaustively paired to generate all relationship proposals. In the proposal sampling stage, given the imbalanced relationship proposals, SABRA constructs a balanced pair mini-batch by the proposed BNPS that samples data point i proportionally to its inverse-frequency. The features of subjects, objects, and union bounding boxes are fed into a graph neural network to extract spatial object interactions. The mask decoder discovers the spatial configuration of [*subject*, *object*] pairs in the learned embeddings. SABRA is robust to the imbalanced negative proposal distribution and also reduces the number of false positive predictions caused by spatial ambiguity.

3.1 Balanced Negative Proposal Sampling

Negative Relationship Proposals. Despite abundant researches in improving feature extraction for positive [*subject*, *predicate*, *object*] triplets in VRD, effective learning under a large number of negative proposals is rarely explored. Previous works [24] only consider the imbalance between positive and negative proposals but ignore the inner imbalance of negative proposals, which has a great impact on the performance of the VRD task.

Suppose we have two sets of bounding boxes $B_{\text{subject}}, B_{\text{object}}$, where B_{subject} contains all the bounding boxes for subjects, and B_{object} is for objects. A relationship proposal is defined as a tuple (b_1, b_2). We define the set of proposals S as $S = \{(b_1, b_2) \mid b_1 \in B_{\text{subject}}, b_2 \in B_{\text{object}}\}$. S can be divided into two disjoint subsets S_{pos} and S_{neg}. S_{pos} denotes the set of positive proposals that correspond to the ground truth, and S_{neg} stands for the wrong proposals.

In VRD task, negative proposals are caused mainly by two reasons: inaccurate detections and incorrect associations. Inaccurate detections lead to negative proposals generated by inaccurate bounding boxes. This type of negative proposals can be easily identified by the visual appearance feature of single object detections alone, but contribute a large portion of S as Fig.1.a shows, due to the exhaustively pairing. On the other hand, incorrect associations cause negative proposals in a more complicated way. For detections with at least one positive relationship, $D_{\text{rel}} = \{b \mid \exists b', (b, b') \in S_{\text{pos}} \vee (b', b) \in S_{\text{pos}}\}$, empirically we have $|D_{\text{rel}}|/|D_{\text{pos}}| < 0.2$, where D_{pos} stands for accurate detections. Moreover, consider a proposal (b_1, b_2) where $b_1 \in D_{\text{rel}}$, i.e., $\exists b'_2, (b_1, b'_2) \in S_{\text{pos}}$. The relationship prediction of (b_1, b_2) will be influenced by the positive proposal (b_1, b'_2), which introduces extra confusion. Thus, the proposals with bounding boxes from D_{rel} are not only under-represented but more confusing. In summary, an imbalanced proposal distribution, where the difficulty of each proposal is negatively correlated with its population size, commonly exists in VRD tasks and would degrade the overall performance of VRD algorithms.

Balanced Negative Proposal Sampling. Our solution is motivated by learning with data imbalance [1, 2]. We propose a balanced negative proposal sampling strategy considering 1 positive class and 5 negative classes.

We first define two helper functions, f_{box} and f_{rel}:

$$f_{\text{box}}(b) = \begin{cases} 1, & \text{if } \exists g \in \text{GT}_{\text{box}}, \text{IoU}(b, g) \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

$$f_{\text{rel}}(b) = \begin{cases} 1, & \text{if } \exists (g_1, g_2) \in \text{GT}_{\text{rel}}, \text{maxIoU}(b, (g_1, g_2)) \geq 0.5, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where GT_{box} denotes the set of ground truth bounding boxes, GT_{rel} denotes the set of ground truth relationships, and $\text{maxIoU}(b, (g_1, g_2)) = \max(\text{IoU}(b, g_1), \text{IoU}(b, g_2))$. Intuitively, $f_{\text{box}}(b)$ indicates if a bounding box b is a positive bounding box, and $f_{\text{rel}}(b)$ denotes if bounding box b belongs to a positive relationship.

Next, we divide S_{neg} following these two principles: (1) simple proposals are caused by inaccurate bounding boxes; (2) difficult proposals are introduced by incorrect [*subject*, *object*] associations. For clarity, we formulate the 5 sub-classes:

$$\begin{aligned} S^1_{\text{neg}} =& \{(b_1, b_2) | \neg f_{\text{box}}(b_1) \wedge \neg f_{\text{box}}(b_2)\}, \\ S^2_{\text{neg}} =& \{(b_1, b_2) | (\neg f_{\text{box}}(b_1) \wedge f_{\text{box}}(b_2)) \vee (f_{\text{box}}(b_1) \wedge \neg f_{\text{box}}(b_2))\}, \\ S^3_{\text{neg}} =& \{(b_1, b_2) | f_{\text{box}}(b_1) \wedge \neg f_{\text{rel}}(b_1) \wedge f_{\text{box}}(b_2) \wedge \neg f_{\text{rel}}(b_2)\}, \\ S^4_{\text{neg}} =& \{(b_1, b_2) | (f_{\text{rel}}(b_1) \wedge f_{\text{box}}(b_2) \wedge \neg f_{\text{rel}}(b_2)) \vee (f_{\text{box}}(b_1) \wedge \neg f_{\text{rel}}(b_1) \wedge f_{\text{rel}}(b_2))\}, \\ S^5_{\text{neg}} =& \{(b_1, b_2) | f_{\text{rel}}(b_1) \wedge f_{\text{rel}}(b_2) \wedge (b_1, b_2) \notin S_{\text{pos}}\}. \end{aligned} \tag{3}$$

This divides the negative proposals into 5 sub-classes: (1) S^1_{neg}, both detections are incorrect; (2) S^2_{neg}, one detection is incorrect; (3) S^3_{neg}, both detections are correct, but they belong to no [*subject*, *predicate*, *object*] triplet; (4) S^4_{neg}, both detections are correct, but only one of them appears in [*subject*, *predicate*, *object*] triplets; (5) S^5_{neg}, both detections are correct, but they appear in two disjoint sets of [*subject*, *predicate*, *object*] triplets, which is the minority of the negative proposals but is the most difficult to learn. Together with the S_{pos}, we divide S into 6 classes, $S = \bigcup_{i=1}^{5} S^i_{\text{neg}} \cup S_{\text{pos}}$.

We adopt the most standard *Balanced Sampling* scheme [34] in learning with imbalance literature to address the imbalance of negative proposals S_{neg}, and we introduce Balanced Negative Proposal Sampling (BNPS). During training time, for each image, we find the top-K object detections, construct $K \times K$ relationship proposals, and count the number of samples for each class. For each proposal s_i, we assign a weight w_i:

$$w_i = \begin{cases} 0.25/|S_{\text{pos}}|, & \text{if } s_i \in S_{\text{pos}}, \\ 0.15/|S^j_{\text{neg}}|, & \text{if } s_i \in S^j_{\text{neg}}, j = 1, \ldots, 5 \end{cases} \tag{4}$$

where we keep the weight, 0.25, of positive proposals to retain the ratio of S_{pos} as [24], while re-balancing the weight of negative proposals, which improves the prediction of low-frequency difficult classes.

3.2 Spatially-Aware Embedding Learning

SABRA adopts a GNN-based paradigm for spatial modeling. Nevertheless, existing methods [22, 46] make no clear separation between the subject, object, and the union box features. **Global Spatial Modeling.** SABRA addresses the divergence in the feature distributions using a simple yet effective method, heterogeneous message passing in GNNs. We construct a heterogeneous graph by explicitly separating the subject nodes x_i and object nodes x_j, and connect them by edges with features e_{ij} extracted from their union bounding box. We process each feature type with different embedding functions, such that a single embedding function fits for only a specific type of feature distribution. Specifically, we build upon the Graph Attention Networks [37], and introduce the heterogeneous message passing as follows:

$$x'_i = x_i + \sum_j \alpha_{ij} f_m([x_i, x_j, e_{ij}]), \alpha_{ij} = \frac{\exp(g_{ij})}{\sum_k \exp(g_{ik})}, g_{ij} = [f_s(x_i), f_t(x_j), f_e(e_{ij})], \tag{5}$$

where α_{ij} is the attention weight for edge (i, j), f_m, f_s, f_t and f_e are functions for message embedding, source node feature embedding, target node feature embedding, and edge feature embedding respectively. k represents all the neighbors of node i. Specifically, in our method, every pair of nodes are connected, no matter which categories are them. Besides, each node is connected to itself. We can extend such a formulation with the multi-head attention mechanism from Transformer Networks [36]. For simplicity, we denote the used multi-head heterogeneous GNN as MH-GAT.

Local Spatial Modeling. To learn the local spatial configuration, we present Spatial Mask Decoder (SMD) which predicts the locations of [*subject*, *object*] . SMD predicts a $2 \times l_p \times l_p$ mask using the VRD feature vectors, where l_p represents the pooling size and the two channels represent the spatial location of subjects and objects in the pooled feature map.

The detailed architecture of Spatial Mask Decoder (SMD) is shown in Fig. 3. The relationship feature is originally used for classification. To ensure that the learned features contain sufficient spatial information during the entire forward pass, we fuse and concatenate two intermediate features from the classifier. This 2048-d feature will be passed to an extra fully connect layer to predict the object mask and subject mask, which describe the position of corresponding bounding boxes related to the union bounding box. Specifically, the subject mask is a matrix $M_{\text{subject}} \in \{0,1\}^{l_p \times l_p}$. In M_{subject}, the value of an entry is equal to 1 if and only if at least half of this rectangular area belongs to the subject bounding box. The object mask $M_{\text{object}} \in \{0,1\}^{l_p \times l_p}$ uses a similar definition as the subject mask.

This guarantees that the local spatial configuration is tightly embedded in the feature vector for relationship prediction. We scale the absolute coordinates of the [*subject*, *object*] pair to get the relative position in the pooled feature map. Compared to reconstruction in the image space, the ROI feature space preserves the locality of the union bounding box but requires fewer parameters. Different from the standard positional embedding [52] which implicitly utilizes the spatial information with the absolute positions as an input, SMD explicitly learns a structured embedding. Compared to [11] which concatenates binary mask as position feature, SMD explicitly imposes the spatial structure in the embedding vector and gives better spatially-aware embeddings.

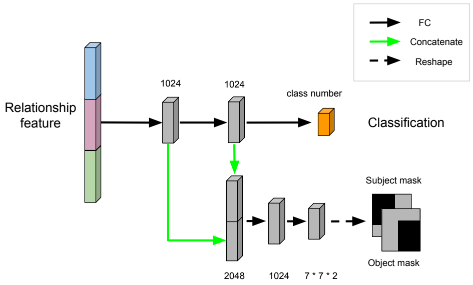


Figure 3: Structure of Spatial Mask Decoder.

4 Experiments

We evaluate SABRA on three commonly used datasets: V-COCO [13], HICO-DET [6], and VRD [27], covering both human-object interaction (V-COCO and HICO-DET) and visual relationship detection (VRD). We compared SABRA with over 20 SOTA methods. Specifically, we trained SABRA with different backbones and perform a comprehensive and fair comparison with existing methods. We incrementally add our proposed components to a baseline model. The baseline directly concatenates appearance features of subjects, objects, and their union features, which are fed into a 3-layer Multi-Layer Perceptron (MLP) for classification. We repeated all experiments more than 3 times, and the standard deviation is smaller than 0.3 for all datasets.

As a brief conclusion, we show that: 1) SABRA generally outperforms other methods by a large margin; 2) the balanced negative proposal sampling strategy can reduce the number of false positive predictions; 3) spatial mask decoder successfully reduces the number of false positives caused by spatial ambiguity.

4.1 Datasets and Evaluation Metrics

V-COCO is based on the 80-class object detection annotations of COCO [23]. It has 10,346 images (2,533 for training, 2,867 for validating and 4,946 for testing). HICO-DET has a total of 47,774 images, covering 600 categories of human-object interactions over 117 common actions on 80 common objects. VRD dataset contains 4,000 images in the train split and 1,000 in the test split. It has 100 different types of objects and 70 types of relationships.

		V-COCO	HICO-DET					
			Default			Known Objects		
Method	Backbone	AP$_{role}$	Full	Rare	Non-Rare	Full	Rare	Non-Rare
iCAN [8]	ResNet50	45.30	14.84	10.45	16.15	16.26	11.33	17.73
Contextual Attention [41]	ResNet50	47.30	16.24	11.16	17.75	17.73	12.78	19.21
In-GraphNet [47]	ResNet50	48.90	17.72	12.93	13.91	-	-	-
VCL [15]	ResNet50	48.30	19.43	**16.55**	20.29	22.00	19.09	22.87
PD-Net [53]	ResNet50	52.30	20.76	15.68	22.28	25.58	19.93	27.28
SABRA(Ours)	ResNet50	**53.57**	**23.48**	16.39	**25.59**	**28.79**	**22.75**	**30.54**
InteractNet [12]	ResNet50-FPN	40.00	9.94	7.16	10.77	-	-	-
PMFNet [39]	ResNet50-FPN	52.00	17.46	15.65	18.00	20.34	17.47	21.20
DRG [9]	ResNet50-FPN	51.00	19.26	**17.74**	19.71	23.40	21.75	23.89
IP-Net [42]	ResNet50-FPN	51.00	19.56	12.79	21.58	22.05	15.77	23.92
Contextual HGNN [40]	ResNet50-FPN	52.70	17.57	16.85	17.78	21.00	20.74	21.08
SABRA(Ours)	ResNet50-FPN	**54.69**	**24.12**	15.91	**26.57**	**29.65**	**22.92**	**31.65**
VSGNet [35]	ResNet152	51.76	19.80	16.05	20.91	-	-	-
PD-Net [53]	ResNet152	52.20	22.37	**17.61**	23.79	26.89	21.70	28.44
ACP [19]	ResNet152	52.98	20.59	15.92	21.98	-	-	-
SABRA(Ours)	ResNet152	**56.62**	**26.09**	16.29	**29.02**	**31.08**	**23.44**	**33.37**

Table 1: Results on V-COCO and HICO-DET datasets

		Rel.		Phr.	
Method	Backbone	R@50	R@100	R@50	R@100
VRD [27]	VGG16	17.03	16.17	14.70	13.86
KL distillation [50]	VGG16	19.17	21.34	23.14	24.03
Zoom-Net [49]	VGG16	18.92	21.41	24.82	28.09
CAI + SCA-M [49]	VGG16	19.54	22.39	25.21	28.89
Hose-Net [47]	VGG16	20.46	23.57	27.04	31.71
RelDN [52]	VGG16	18.92	22.96	26.37	31.42
AVR [29]	VGG16	22.83	25.41	29.33	33.27
SABRA(Ours)	VGG16	**24.47**	**29.16**	**30.57**	**36.80**
GPS-Net [22]	VGG16(C)	21.50	24.30	28.90	34.00
MCN [51]	VGG16(C)	24.50	28.00	31.80	37.10
SABRA(Ours)	VGG16(C)	**26.29**	**31.08**	**32.01**	**38.48**
UVTransE [17]	VGG16(V)	25.66	29.71	30.01	36.18
SABRA(Ours)	VGG16(V)	**27.87**	**32.48**	**33.56**	**39.62**
ATR-Net [11]	ResNet101	-	-	31.96	36.54
SABRA(Ours)	ResNet101	**26.73**	**31.11**	**32.81**	**38.68**

Table 2: Results on the VRD dataset

	Sampling	Spatial Learning	GNN	AP$_{role}$
1	-	-	-	50.20
2	BNPS	-	-	52.24
3	-	SMD	-	50.84
4	-	-	MH-GAT	51.65
5	-	SMD	MH-GAT	52.82
6	BNPS-2cls	SMD	MH-GAT	53.74
7	BNPS-3cls	SMD	MH-GAT	53.93
8	BNPS	-	MH-GAT	53.67
9	BNPS	Binary [illegible]	MH-GAT	53.90
10	BNPS	PE [illegible]	MH-GAT	54.29
11	BNPS	SMD	-	52.53
12	BNPS	SMD	M-GAT	51.65
13	BNPS	SMD	MH-GAT(NE)	53.80
14	BNPS	SMD	MH-GAT	**54.69**

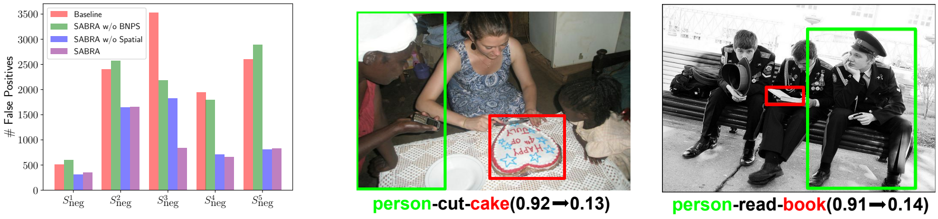
Table 3: Ablation Study on V-COCO

As for evaluation metrics, we followed the convention in prior literature[27, 50]: mean average precision (mAP) is used to estimate the performance of V-COCO and HICO-DET; Recall@K is used for VRD, where K denotes the number of top K predictions.

4.2 Quantitative Results

We present our results in Table 1 for HOI datasets (V-COCO and HICO-DET) and Table 2 for the VRD dataset. For HOI, we cluster our results according to the backbones, including ResNet50, ResNet50-FPN, and ResNet152, with increasing feature extraction ability. For VRD datasets, VGG-16 and ResNet101 pretrained on ImageNet are used. To align with the setup of some prior works [17, 24, 51], we use MS COCO [23] and Visual Genome [21] as additional datasets for VGG16, denoted by VGG (C) and VGG (V) respectively.

We observe that SABRA generally improves the SOTA methods significantly on all datasets for both HOI and VRD. For example, on V-COCO with ResNet152 backbone, SABRA achieves 56.62 mAP while the SOTA model ACP [20], gives an mAP of 52.98. Specifically, we want to highlight the performance gain of SABRA on the V-COCO dataset.



(a) False positive comparison (b) False positive reduction examples of SABRA

Figure 4: Qualitative analysis of SABRA on V-COCO dataset. (a) The number of false positive predictions under each class by Baseline model, SABRA w/o BNPS, SABRA w/o spatial module (MH-GAT + SMD), and SABRA. BNPS and spatial module significantly reduce the number of false positive predictions. (b) Predictions of SABRA. In sub-predicate-obj(v_1, v_2), v_1 denotes the prediction confidence of the baseline model and v_2 denotes the confidence of SABRA. SABRA successfully reduced the low-frequency difficult false positives.

V-COCO uses the object annotations of the COCO dataset, which are accurate and include all objects, regardless of the relationships. This gives a more imbalanced distribution than HICO-DET, where only the objects in relationships are considered. Furthermore, accurate annotations also give rise to better object detection, which amplifies the significance of spatial information for a good performance. Specifically, we noticed that for PD-Net [53], using a more powerful backbone (ResNet152) gives no performance gain than the smaller one (ResNet50). SABRA, on the contrary, can fully exploit the power of ResNet152 and give a large performance improvement compared with ResNet50 (56.62 V.S. 53.57).

4.3 Ablation Studies

We conduct a comprehensive ablation study on the V-COCO dataset to understand each proposed component: Balanced Negative Proposal Sampling (BNPS), Multi-head Heterogeneous Graph Attention (MH-GAT), and Spatial Mask Decoder (SMD). We present the results in Table 3, where row 1 denotes the baseline and row 14 denotes the complete SABRA.

Each proposed component improves the baseline. We perform incremental analysis in rows 1-4. Specifically, we observe that by simply improving the optimization process with BNPS, baseline + BNPS (row 2) gains 2.04 improvement on AP$_{role}$. This suggests that imbalanced proposal distribution significantly hinders the model performance, and BNPS addresses this issue effectively.

Balancing negative proposal distribution generally improves the VRD performance. In rows 5-7 and row 14, we compare BNPS with 3 other alternatives. (1) BNPS-3cls (row 7): balanced sampling over $\{S^1_{\text{neg}}, S^2_{\text{neg}}, S^3_{\text{neg}} \cup S^4_{\text{neg}} \cup S^5_{\text{neg}}\}$, i.e., ignoring the difference between negative proposals when detections are correct; (2) BNPS-2cls (row 6): balanced sampling over $\{S^1_{\text{neg}} \cup S^2_{\text{neg}}, S^3_{\text{neg}} \cup S^4_{\text{neg}} \cup S^5_{\text{neg}}\}$, i.e., further ignoring the differences of negative proposals when detections are incorrect; (3) None (row 5): we remove the BNPS. We fix the positive sample rate to be 25% and balance the rest 75% samples over the given distributions. Comparing BNPS with BNPS-2cls and BNPS-3cls, we conclude BNPS improves prediction accuracy for both inaccurate detections and incorrect associations.

Understanding spatial information is crucial to VRD. In rows 8-10 and row 14, we compare the proposed Spatial Mask Decoder (SMD) with 3 alternatives. (1) Binary (row 9) [11]:

a binary mask over the union feature to specify the position of [*subject*, *object*] ; (2) PE (row 10) [52]: naive positional embedding; (3) None (row 8): no spatial learning module. We conclude that: (1) adding spatial information learning generally improves the performance; (2) implicitly imposing spatial information with PE or binary masks gives worse performance than SMD which enforces an explicit constraint over the feature space.

MH-GAT is more effective in heterogeneous VRD graphs. In rows 11-13 and row 14, we compare MH-GAT with other 4 alternatives: (1) Multi-head GAT [37] (row 12), standard homogeneous message passing scheme which has the same number of heads and edge features used in MH-GAT, (2) MH-GAT without edge feature (row 13), and (3) None (row 11). We observe that: (1) GNNs generally improve the VRD performance; (2) by separating the parameters for objects, subjects, and union features, the heterogeneous message passing scheme in MH-GAT improves the M-GAT; (3) edge feature is important to MH-GAT.

4.4 Qualitative Analysis

Qualitatively, we provide extra statistics and visualizations on the V-COCO dataset to better understand the performance improvement of SABRA in Fig. 4.

To verify the ability of SABRA on reducing the number of false positives, we compute the per-image false positive predictions for $S_{\text{neg}}^{1:5}$ by thresholding the prediction confidence at 0.5 in Fig. 4. Compared with the baseline, SABRA w/o BNPS, SABRA w/o Spatial, and SABRA have reduced the total number of low-frequency difficult $S_{\text{neg}}^{3:5}$ by 14.7%, 58.5%, and 71.1%. BNPS gave a sharp decrease on $S_{\text{neg}}^{3:5}$, which suggests the necessity of learning from a balanced proposal distribution. The spatial module successfully filtered the irrelevant objects in S_{neg}^{3} and significantly reduced the number of false positives in total. Combining both of them, SABRA gives the best performance among all alternatives. However, we also noticed that on $S_{\text{neg}}^{4:5}$ and $S_{\text{neg}}^{1:2}$, the spatial module has no clear improvement, because the spatial module potentially overexploited the correct detection of the positive [*subject*, *predicate*, *object*] triplet. Auxiliary constraints on the relationship assignment could be considered to address this issue. We leave it for future study.

We visualize two examples of SABRA successfully reducing the low-frequency difficult false positives where both detections are correct in Fig. 4.b. Without the spatial module, the VRD algorithm assigns high confidence (0.92) to the [*subject*, *object*] pair that the woman is cutting the cake. This confidence was reduced to 0.13 by SABRA.

5 Conclusion

We present SABRA for alleviating false positives in VRD. We divided the negative [*subject*, *object*] proposals into 5 sub-classes with imbalanced data distribution, and addressed the data imbalance by Balanced Negative Proposal Sampling. SABRA incorporates the global contextual information with MH-GAT and local spatial configuration by SMD. SABRA significantly outperforms the SOTA methods on V-COCO, HICO-DET, and VRD datasets.

As the first paper to consider the data imbalance in the negative proposal distribution, SABRA used a relatively simple strategy, balanced sampling. More advanced techniques could be considered to further improve the performance in future studies.

Acknowledgments: We thank the anonymous reviewers for their invaluable comments and suggestions. This work is partially supported by Sensetime, Grant Nos.U1636211, 61672081, 61370126, and Grant No.SKLSDE2019ZX-17.

References

[1] Ricardo Barandela, E Rangel, José Salvador Sánchez, and Francesc J Ferri. Restricted decontamination for the imbalanced training sample problem. In *Iberoamerican congress on pattern recognition*, pages 424–431. Springer, 2003.

[2] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259, 2018.

[3] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881, 2019.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019.

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. IEEE Computer Society, 2018.

[6] Carolina Galleguillos, Andrew Rabinovich, and Serge Belongie. Object categorization using co-occurrence, location and appearance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[7] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017.

[8] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018.

[9] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. DRG: dual relation graph for human-object interaction detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, volume 12357 of *Lecture Notes in Computer Science*, pages 696–712. Springer, 2020. doi: 10.1007/978-3-030-58610-2_41. URL https://doi.org/10.1007/978-3-030-58610-2_41.

[10] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, and Petros Maragos. Attention-translation-relation network for scalable scene graph generation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[11] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, Athanasia Zlatintsi, and Petros Maragos. Deeply supervised multimodal attentional translation embeddings for visual relationship detection. In *2019 IEEE International Conference on Image Processing, ICIP 2019, Taipei, Taiwan, September 22-25, 2019*, pages 1840–1844. IEEE, 2019.

[12] Georgia Gkioxari, Ross B. Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018.

[13] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *CoRR*, 2015.

[14] Munawar Hayat, Salman Khan, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Gaussian affinity for max-margin class imbalanced learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6469–6479, 2019.

[15] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. *CoRR*, abs/2007.12407, 2020. URL https://arxiv.org/abs/2007.12407.

[16] Yue Hu, Siheng Chen, Xu Chen, Ya Zhang, and Xiao Gu. Neural message passing for visual relationship detection. In *ICML Workshop on Learning and Reasoning with Graph-Structured Representations, Long Beach, CA*, 2019.

[17] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik. Contextual translation embedding for visual relationship detection and scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[18] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.

[19] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. *CoRR*, 2020.

[20] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. *arXiv preprint arXiv:2007.08728*, 2020.

[21] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

[22] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019.

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, 2014.

[24] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020.

[25] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020.

[26] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[27] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, 2016.

[28] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016.

[29] Jianming Lv, Qinzhe Xiao, and Jiajie Zhong. Avr: Attention based salient visual relationship detection. *arXiv preprint arXiv:2003.07012*, 2020.

[30] Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2020.

[31] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, 2018.

[32] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Charles Rosenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1100–1109, 2015.

[33] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *arXiv preprint arXiv:2007.10740*, 2020.

[34] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision ECCV*, 2016.

[35] Oytun Ulutan, A. S. M. Iftekhar, and B. S. Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13614–13623. IEEE, 2020.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

[37] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016.

[39] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *International Conference on Computer Vision, ICCV*, 2019.

[40] Hai Wang, Wei-Shi Zheng, and Yingbiao Ling. Contextual heterogeneous graph network for human-object interaction detection. *CoRR*, abs/2010.10001, 2020. URL https://arxiv.org/abs/2010.10001.

[41] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5693–5701. IEEE, 2019. doi: 10.1109/ICCV.2019.00579. URL https://doi.org/10.1109/ICCV.2019.00579.

[42] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020.

[43] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[44] Meng Wei, Chun Yuan, Xiaoyu Yue, and Kuo Zhong. Hose-net: Higher order structure embedded network for scene graph generation. *CoRR*, 2020.

[45] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *IEEE Transactions on Multimedia*, 22(6):1423–1432, 2019.

[46] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[47] Dongming Yang and Yuexian Zou. A graph-based interactive reasoning for human-object interaction detection. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. ijcai.org, 2020.

[48] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018.

[49] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, 2018.

[50] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017.

[51] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. Multi-task compositional network for visual relationship detection. *International Journal of Computer Vision*, 128(8): 2146–2165, 2020.

[52] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019.

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

[54] Hao Zhou, Chongyang Zhang, and Chuanping Hu. Visual relationship detection with relative location mining. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. ACM, 2019.