

Supplementary Material: Towards Overcoming False Positives in Visual Relationship Detection

BMVC 2021 Submission # 332

1 Implementation Details

1.1 Overall Pipeline

The overall pipeline of SABRA consists of three major components: backbone, detector, and VRD classifier. Backbone is a feature extractor for object detection and relationship detection. Detector predicts all potential bounding boxes and produces a set of detections:

$$B = \{(x_1, y_1, x_2, y_2, cls, score)\}, \quad (1)$$

where (x_1, y_1) are the coordinates of the upper left corner, (x_2, y_2) are the coordinates of the bottom right corner, cls is the category of this bounding box, and $score$ is the confidence of this prediction. VRD classifier uses this set of detections to generate all relationship proposals S and categorize each proposal. In this way, we obtain a set of triplets associated with a $score$ from VRD classifier:

$$T = \{(b_1, b_2, cls, score)\}. \quad (2)$$

1.2 Backbone

To fully compare with other methods, we use several backbone setups in experiments including VGG-16, ResNet50, ResNet50-FPN, ResNet101, and ResNet152. In particular, when we use ResNet50-FPN for both object detection and relationship detection following [4], we use *ResNet50-FPN* to denote this setting. Meanwhile, some other works like [15] use ResNet50-FPN for object detection, but use only ResNet50 for relationship detection. Therefore, we build up this setting named *ResNet50* that a ResNet50 module is shared for both object detection and relationship detection, and an extra FPN is used only for object detection. In addition, FPN is used neither in object detection nor relationship detection process when we use VGG-16, ResNet101, and ResNet152 as the backbone.

Specifically, for V-COCO and HICO-DET datasets, we use ImageNet + MS COCO pre-trained ResNet50, ResNet50-FPN, and ResNet152. For VRD dataset, we respectively use VGG-16 pre-trained on ImageNet, ImageNet + MS COCO, and ImageNet + Visual Genome. We also use ImageNet pre-trained ResNet101 for VRD task.

1.3 VRD Classifier

The overall structure of the VRD classifier is shown in Fig. 2 in the main text. In this section, we will describe the implementation details of our VRD classifier, training strategy, and inference procedure.

Training Strategy. VRD classifier receives the detection set B and further divides it into two sets B_{subject} and B_{object} . In HOI task, B_{subject} contains all human bounding boxes and B_{object} contains all bounding boxes from detector. In general VRD task, both B_{subject} and B_{object} contain all detection results. Under this definition, we generate a proposal set S . For each image, we sample 64 relationship proposals from the set S . This hyper-parameter holds across all our model configurations, no matter whether we use BNPS. Moreover, we also keep the ratio of positive proposals as 0.25 for all experiments. Sigmoid activation is used to predict both the confidence of each category and the spatial mask. Correspondingly, binary cross-entropy loss is used for both supervisions.

Inference Procedure. For each relationship proposal (b_1, b_2) , we predict the score s_{cls} of each relationship category cls . The final score of triplet (b_1, b_2, cls) is calculated as below:

$$score = s_1 \times s_2 \times s_{cls}, \quad (3)$$

where s_1 and s_2 are the scores of bounding box b_1 and b_2 respectively.

1.4 Training Loss

During training, we jointly optimize the object detector and relationship classifier by:

$$L = L_{\text{RPN}}^{\text{cls}} + L_{\text{RPN}}^{\text{loc}} + L_{\text{RCNN}}^{\text{cls}} + L_{\text{RCNN}}^{\text{loc}} + L_{\text{VRD}}^{\text{cls}} + L_{\text{VRD}}^{\text{mask}}, \quad (4)$$

where $L_{\text{RPN}}^{\text{cls}}$, $L_{\text{RPN}}^{\text{loc}}$, $L_{\text{RCNN}}^{\text{cls}}$, $L_{\text{RCNN}}^{\text{loc}}$ are losses for object detection. Their definitions are the same as [8]. $L_{\text{VRD}}^{\text{cls}}$ is a classification loss of each relation proposal. $L_{\text{VRD}}^{\text{mask}}$ is an auxiliary loss used in our SMD model. These two losses are designed for the VRD branch. We opt for Binary Cross Entropy loss for both items. During inference, SABRA predicts all given relationship proposals without sampling.

2 Experimental Setup

2.1 Evaluation Metrics

We follow the convention in prior literature and used three different evaluation metrics for these datasets.

Following [8], we use Recall@ N as our evaluation metric on VRD. Recall@ N computes the recall rates using the top N predictions per image. To be consistent with prior works, we report Recall@50 and Recall@100 in our experiments. We evaluate two tasks: *relationship*

detection that outputs triple labels and evaluates bounding boxes of the subject and object separately; *phrase recognition* that takes a triplet as a union bounding box and predicts the triple labels. Besides, we additionally report the top- k predictions in Table 1 as [□], where $k = 1$ or $k = 70$. In Table 2 in the main text, $k = 1$ is set by default.

For V-COCO and HICO-DET, mean average precision (mAP) is used to estimate the performance. A triplet [human, verb, object] is considered a positive prediction if and only if there exists a triplet [human', verb', object'] in the ground truth satisfying: (1) $\text{IoU}(\text{human}, \text{human}') \geq 0.5$, (2) $\text{verb} = \text{verb}'$, (3) $\text{IoU}(\text{object}, \text{object}') \geq 0.5$

In HICO-DET, we calculate the mAP among all pairs of [verb, object]. In *Default* mode, we calculate AP on all images. In *Known Objects* mode, the category of the bounding box is known and we only calculate AP between humans and objects from a specific category. In V-COCO, we calculate the mAP for all categories of verbs, which is called AP_{role} .

2.2 Data Pre-processing

For V-COCO, the ground truth detection is obtained from COCO. As some relationships containing invisible objects, we fill the region of the object with the coordinates as the subject. In other words, we generate a ground truth triplet (b_1, b_1, cls) when the object is invisible.

For HICO-DET, we generate ground truth detection from all bounding boxes of triplets. It should be noted that, in HICO-DET, the annotation of each relationship is independent. Therefore, there may exist multiple bounding boxes for a single person or object. To generate a unique bounding box annotation for each instance, we merge those bounding boxes where the IoU values are no less than 0.5.

For general VRD, we have no extra pre-processing.

2.3 Training Procedure

For V-COCO, we first train the object detection on the COCO dataset. Then we freeze the backbone and jointly train the detector and VRD classifier on the corresponding dataset.

For HICO-DET, we first train the object detection on the COCO dataset and then finetune on the detection results from HICO-DET. After that, we jointly train the Detector and VRD classifier with a frozen backbone.

For general VRD, we use ImageNet pre-trained VGG-16 backbone to initialize our model and then train object detection on the general VRD dataset. Finally, we jointly train the detector and VRD classifier with the frozen backbone.

In each training process, we use 16 GPUs (1080TI) and train 25 epochs with the initial learning rate being 0.00125. The learning rate is decreased in the 17th and 23rd epoch with a 0.1 decay rate. When training solely the object detection, each batch contains four different images. For the joint training of detection and relationship detection, each batch contains two different images.

3 Additional Analysis for BNPS

3.1 Other reasons for inaccurate detections

Although inaccurate bounding boxes lead to a large number of easy negative proposals, we cannot simply ignore these detection results. We observe that decreasing the number of

Method	Backbone	Relationship Detection				Phrase Detection			
		k=1		k=70		k=1		k=70	
		R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
VRD [8]	VGG16	17.03	16.17	24.90	20.04	14.70	13.86	21.51	17.35
KL distillation [10]	VGG16	19.17	21.34	22.68	31.89	23.14	24.03	26.32	29.43
Zoom-Net [11]	VGG16	18.92	21.41	21.37	27.30	24.82	28.09	29.05	37.34
CAI + SCA-M [12]	VGG16	19.54	22.39	22.34	28.52	25.21	28.89	29.64	38.39
Hose-Net [13]	VGG16	20.46	23.57	22.13	27.36	27.04	31.71	28.89	36.16
RelDN [14]	VGG16	18.92	22.96	21.52	26.38	26.37	31.42	28.24	35.44
AVR [15]	VGG16	22.83	25.41	27.35	32.96	29.33	33.27	34.51	41.36
SABRA(Ours)	VGG16	24.47	29.16	27.27	33.99	30.57	36.80	33.39	41.79
GPS-Net [16]	VGG16 (C)	21.50	24.30	-	-	28.90	34.00	-	-
MCN [17]	VGG16 (C)	24.50	28.00	-	-	31.80	37.10	-	-
SABRA(Ours)	VGG16 (C)	26.29	31.08	29.44	36.44	32.01	38.48	35.45	44.07
UVTransE [18]	VGG16 (V)	25.66	29.71	27.32	34.11	30.01	36.18	31.51	39.79
SABRA(Ours)	VGG16 (V)	27.87	32.48	30.71	37.71	33.56	39.62	36.62	45.29
ATR-Net [19]	ResNet101	-	-	-	-	31.96	36.54	36.06	43.45
SABRA(Ours)	ResNet101	26.73	31.11	29.92	37.43	32.81	38.68	36.24	45.26

Table 1: Complete results on the VRD dataset.

Detection Top- N	RS	BNPS
100	51.95	54.29
50	52.82	54.69
40	52.53	54.43
30	53.05	54.52
20	52.92	53.68

Table 2: AP_{role} performance of the different numbers of detections from the detector on the V-COCO test set. We use ResNet50-FPN as our backbone and modify the sampling strategy and selection of the detector’s top- N while keeping others unchanged.

top- N detections may have a negative influence on VRD algorithms. Table 2 is a thorough comparison of the top- N detections we keep from the detector and whether we use BNPS or random sampling (RS) against the model performance. The results suggest that decreasing the number of top- N has no evident improvement in the final performance (Top-50 w/o BNPS V.S. Top-30 w/o BNPS, 52.82 V.S. 53.05). The major reason is that we still need many inaccurate bounding boxes in inference for higher recall values. Besides, we find that our proposed Balanced Negative Proposal Sampling has remarkable improvement no matter which top- N we select. These results strongly prove the effectiveness and robustness of our proposed method.

3.2 Analysis of the improvement

The improvement of BNPS comes from two major sources. Firstly, BNPS reduces the easy negative proposals caused by inaccurate detections; secondly, BNPS balances the difficult negative proposals, considering whether the subject and the object are involved in any triplet from the ground truth. These two parts are both significant and essential, which is partially proven in the ablation study. In this section, we add some extra quantitative and qualitative results.

We include an additional experiment to test if the performance improvement of SABRA

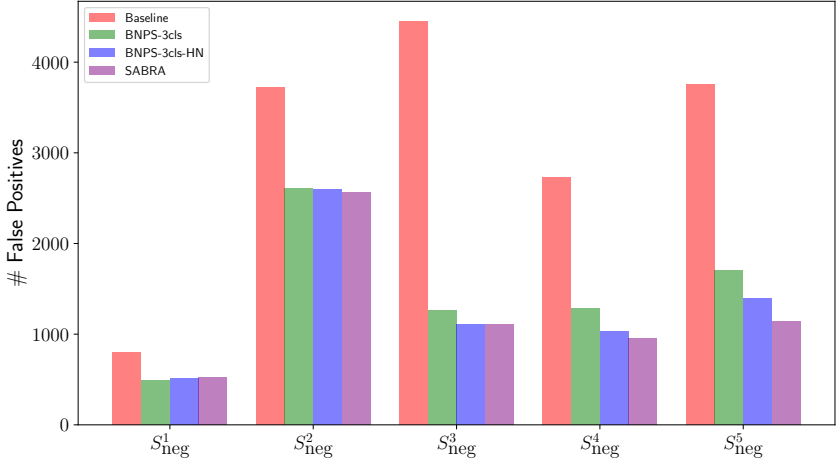


Figure 1: Qualitative analysis of different sampling strategies.

comes only from the increase in the number of difficult negative proposals $S^{3:5}_{neg}$. We compare a BNPS-3cls variant, BNPS-3cls-HN with BNPS. In the original BNPS-3cls, we use a sample rate $[0.25, 0.25, 0.25]$ for S^1_{neg} , S^2_{neg} and $S^{3:5}_{neg}$. However, in BNPS, each class receives a sample rate of 0.15, which gives a 0.45 sample rate for the negative classes $S^{3:5}_{neg}$. We design BNPS-3cls-HN such that it assigns the same sample rate to the hard negative classes, $[0.15, 0.15, 0.45]$ for S^1_{neg} , S^2_{neg} and $S^{3:5}_{neg}$. We train both models on V-COCO and report the results briefly here: BNPS-3cls-HN gives 54.20 AP_{role} while SABRA achieves 54.69. This suggests that the improvement from BNPS-3cls to SABRA is not simply because of the increased sample rate of hard negative proposals. The balance among $S^{3:5}_{neg}$ also plays a critical role in this improvement.

We also visualize the number of negative predictions of each negative type in Fig. 1. We observe that: (1) the total number of false positives in low-frequency difficult classes, $S^{3:5}_{neg}$, of SABRA is lower than BNPS-3cls because we increase the total ratio of difficult negative proposals. (2) the reduction ratio of S^5_{neg} is higher than that of S^4_{neg} . Meanwhile, the reduction ratio of S^4_{neg} is higher than that of S^3_{neg} , which suggests that our balance among S^3_{neg} , S^4_{neg} , S^5_{neg} clearly improves the ability to identify the negative proposals, especially on the difficult proposals.

3.3 BNPS compared with other sampling methods

We have proved the rationality and effectiveness of our proposed BNPS in sampling ideal relationship pairs among a vast number of proposals. To further examine the superiority of BNPS, we implement several alternative sampling methods including *online hard example mining (OHEM)* [9] and *focal loss* [9] and compare their capacities.

OHEM was first proposed in the object detection area and it prefers to sample harder examples than easier ones [9]. Typically, for a mini-batch of samples, only the top- k most

difficult proposals, i.e., proposals with top- k highest loss values, will contribute to model optimization. In practice, the mini-batch is constructed to enforce a 1:3 ratio between positives and negatives, i.e., proposals in S_{pos} and S_{neg} , to help ensure each mini-batch has enough positives. Since positive proposals are usually insufficient, we only apply OHEM on S_{neg} to ensure this claim. We set mini-batch size $B \in \{48, 64, 96\}$ in our experiment.

The focal loss modifies the standard cross-entropy loss to dynamically down-weight the contribution of easy examples:

$$\begin{aligned} \text{FL}(p_t) &= -\alpha(1 - p_t)^\gamma \log(p_t), \\ p_t &= \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise.} \end{cases} \end{aligned} \quad (5)$$

In the above $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0, 1]$ is the model’s estimated probability for the class with label $y = 1$. α and γ are the balance coefficient and the tunable focusing parameter respectively. In our experiment, we use $\alpha = 0.25$ with $\gamma \in \{1, 2\}$.

Besides the above methods, Random Sampling and BNPS are also implemented as relationship pair sampling methods for comparison. We train models using these methods on V-COCO and keep all other settings consistent with setups in Section 2. Additionally, we assign $N = 50$ for top- N detections.

Method	Setup	AP _{role}
RS	-	52.82
	$B = 48$	45.81
OHEM	$B = 64$	45.75
	$B = 96$	45.73
Focal Loss	$\gamma = 1$	52.55
	$\gamma = 2$	51.10
BNPS	-	54.69

Table 3: Performance of the different relationship pair sampling methods. We use ResNet50-FPN as our backbone and modify the sampling strategy while keeping others unchanged.

As results shown in Table 3, we observe that BNPS achieves the best performance, which reconfirms its superiority. Meanwhile, focal loss and OHEM in all settings behave worse than the basic random sampling method, which draws our attention. Focal loss and OHEM both emphasize the importance of samples with high loss values. However, in the scenario of relationship detection, these methods may focus too much on the hard samples and overlook the role of easy ones. Compared with them, our proposed BNPS uses the predefined sample classification method and pay equal attention to 5 classes of negative proposals, which makes our sampling strategy more robust towards the complicated distribution of samples in this scenario and less likely to be influenced by outliers.

References

- [1] 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.

- [2] 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.
- [3] 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.
- [4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [5] Xin Lin, Changxing Ding, Jinqian 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.
- [6] 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.
- [7] Jianming Lv, Qinzhe Xiao, and Jiajie Zhong. Avr: Attention based salient visual relationship detection. *arXiv preprint arXiv:2003.07012*, 2020.
- [8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015.
- [9] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [10] Meng Wei, Chun Yuan, Xiaoyu Yue, and Kuo Zhong. Hose-net: Higher order structure embedded network for scene graph generation. *CoRR*, 2020.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.

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