Alleviating Training Bias with Less Cost via Multi-expert De-biasing Method in Scene Graph Generation

Xuezhi Tong
xuezhitong899@gmail.com
College of Intelligence and Computing, Tianjin University
Tianjin, China

Rui Wang*
wangrui@iie.ac.cn
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China

Lihua Jing
jinglihua@iie.ac.cn
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China

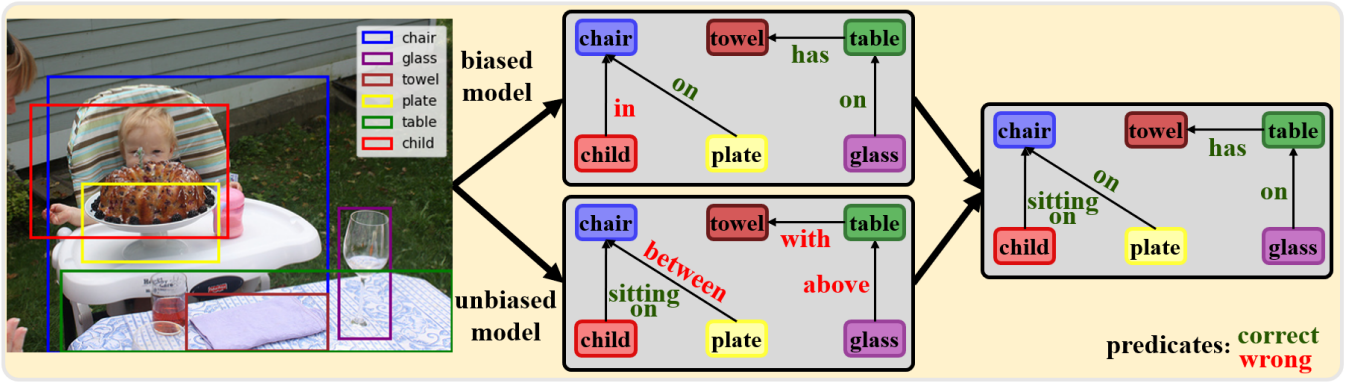

Figure 1: Illustration of the label ambiguity problem. We use MOTIFS [12] as a biased method, and RTPB [3] for unbiased predictions. While "child in chair" is corrected, other predicates are all wrongly predicted as less frequent classes because of the over-emphasizing of tail classes. With the proposed method, we can complement the biased and unbiased models and make correct predictions adaptively.

ABSTRACT

Scene graph generation (SGG) methods have suffered from a severe training bias towards frequent (head) predicate classes. Recent works owe it to the long-tailed distribution of predicates and alleviate the long-tailed problem to conduct de-biasing. However, the "unbiased" models are in turn biased to tail predicate classes, resulting in a significant performance loss on head predicate classes. The main cause of such a trade-off between head and tail predicates is the fact that multiple predicates from the head or tail ones can be labeled as the ground-truth. To this end, we propose a multi-expert de-biasing method (MED) for SGG that can produce unbiased scene graphs with minor influence on recognizing head predicates. We avoid the dilemma of balancing between head and tail predicates by adaptively classifying the predicates with multiple complementary models. Experiments on the Visual Genome dataset show that MED provides significant gains on mRecall@K without harming the performance on Recall@K, and achieves a state-of-the-art on the mean of Recall@K and mRecall@K.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**.

KEYWORDS

Scene graph generation, long-tailed distribution, multi-expert network, de-biasing

ACM Reference Format:
Xuezhi Tong, Rui Wang, and Lihua Jing. 2023. Alleviating Training Bias with Less Cost via Multi-expert De-biasing Method in Scene Graph Generation. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3607541.3616816

1 INTRODUCTION

Scene graph generation (SGG), aiming to recognize the objects and their relationships in images, is of great importance for high-level visual scene understanding. In this task, multiple relationship triplets are produced for each image and connected to form a scene graph, where objects are regarded as nodes and predicates are treated as edges in the graph. As the graphical representation of scenes, a scene graph not only presents spatial (localization) and

*Corresponding Author.

McGE '23, October 29, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0278-5/23/10.
https://doi.org/10.1145/3607541.3616816

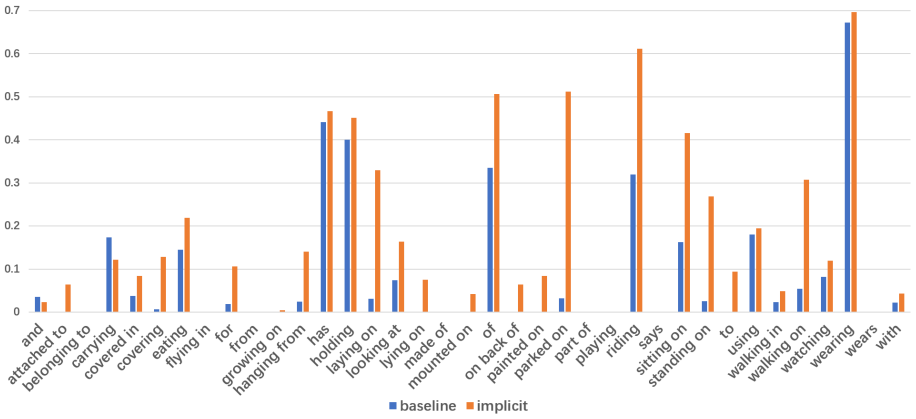

Figure 2: The per-class Recall@100 for a baseline method [19] and the implicit model finetuned based on it. With less label ambiguity, the performance on implicit predicates is significantly improved, which shows the correlation between label ambiguity and training bias.

semantic (recognition) information of objects, but also consists of an interactive formulation describing the scene. However, early methods show a severe training bias towards naive predicate classes like preferring *on* rather than *sitting on*, making the generated scene graph less informative.

Most following works owe such training bias to the long-tailed distribution of predicate classes. They either re-sample the image data to provide more training instances for tail classes or re-weight the predicate classes to emphasize tail classes, based on the frequency of predicate classes. For example, re-sampling methods can re-sample the predicates from both image-level and instance-level to provide more training data for tail classes [11] or use different sampling strategies for different training stages with a simple model to avoid over-fitting [1]. Re-weighting methods can directly balance the classes using the softened frequency [16], or further train a sub-model to learn a more balanced set of weights for the predicate classes [6]. However, while these methods have achieved state-of-the-art performance on mean Recall@K, they have also severely degraded the model performance on Recall@K, which is dominated by the head predicates. That means as a cost of correcting a relatively small number of tail predicates, a large number of head predicates have been recognized wrongly, as shown in Fig. 1. In fact, the trade-off of head predicates and tail predicates is due to the diversity of predicate labels in the dataset, *i.e.* for a certain object pair, there may exist multiple options for the annotators. As SGG asks the model to predict a single relationship for each pair of objects, instances that are similar visually may have different labels. To conclude, there exists a contradiction between the labeling rule and the nature of relationship in the SGG datasets. We call it the label ambiguity problem for SGG.

In this paper, we explore alleviating the training bias by solving the label ambiguity problem. To validate the correlation between label ambiguity and training bias, we design a toy experiment. As in [2], we divide the label set of predicates into explicit and implicit sets. Explicit predicates like *on*, *in*, and *in front of*, only describe the spatial layout of the objects, while implicit predicates are more about semantic interactions like *belonging to*, *made of*, and *wearing*. We finetune an SGG model with the implicit subset of the Visual Genome (VG) dataset [10], namely the implicit model. As shown in Fig. 2, we evaluate the model on the original test set and see significant improvement in recognizing implicit predicates. With the explicit classes excluded, there is less label ambiguity in the dataset. As a result, the model performance on implicit predicates is improved without any structural modification or training strategy adjustment.

Inspired by the multi-expert network [5], we explore stepping out of the one-model-training scheme that existing methods usually adopt, *i.e.* using multiple models for predicates that vary in semantic depth. To this end, we propose a novel general framework that adds a multi-expert network to SGG models, dubbed the multi-expert de-biasing method (MED). We divide the network of a pre-trained SGG model into two modules according to their functionality. The object module is used to make spatial and semantic predictions of the objects, while the predicate module takes its features as part of the input and predicts the relationships of object pairs. For two models that are designed for a full label set and an implicit label set respectively, we call them the full model and implicit model. To complement these two models, we combine the predicate module of the full and implicit models to be a multi-expert predicate module. Compared to a single-model framework that is usually biased to head/tail predicates, MED switches from class-level de-biasing to instance-level de-biasing, *i.e.* determines to prefer head or tail predicates based on the confidence of two models. By taking label ambiguity into account, the framework is able to maintain the performance of the two biased models for corresponding classes as much as possible. The contributions of this paper are as follows:

(1) We demonstrate that the trade-off between Recall@K and mRecall@K in current unbiased SGG models is caused by the label ambiguity problem: The predicate label of a relationship may have multiple plausible options while only a single category is regarded as the ground-truth;

(2) We propose a simple yet effective novel framework leveraging multi-expert architectures to alleviate the label ambiguity problem in SGG.

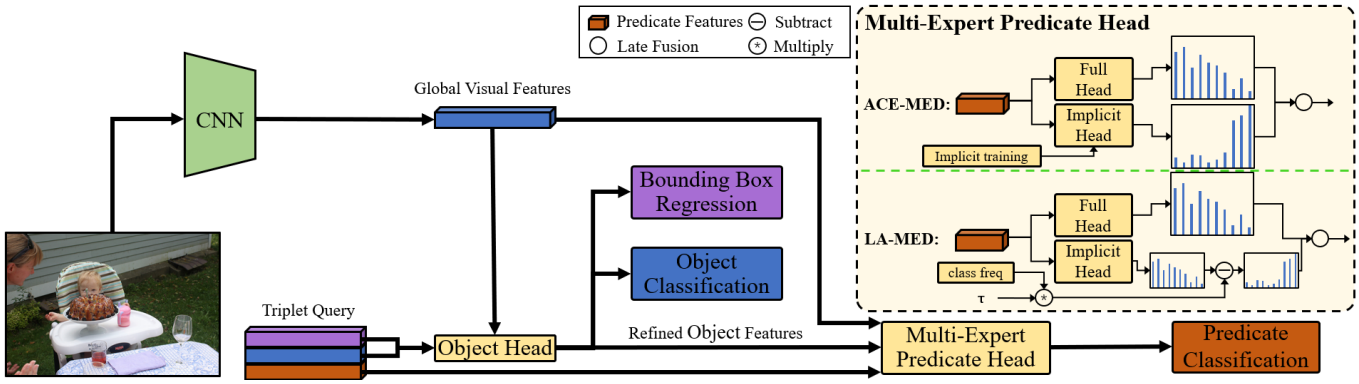

Figure 3: Illustration of the framework of the proposed method. For bounding box regression and object classification, we adopt the general pipeline of one-stage SGG methods. We discuss the predicate head in detail in Sec. 2.3.

(3) The proposed method is evaluated on the VG dataset [10], achieving significant improvement on mRecall@K while maintaining the performance on Recall@K, and resulting in state-of-the-art results on the mean of Recall@K and mRecall@K.

2 METHOD

Overview. Fig. 3 illustrates the proposed framework of MED. The global visual features are extracted from the image by a convolution neural network (CNN) and fed into the object and predicate heads. Triplet queries are randomly initiated and split into three parts for bounding box regression, object classification, and predicate classification respectively. The two predicate heads are designed for the full set and implicit set of classes, respectively. The results are then fused to get the final predictions. Note that MED is a model-independent method that is applicable to any biased model. In this paper, we take a one-stage SGG generator as an example.

2.1 Problem Formulation

A *scene graph* $G = (U, E)$, is a graphical representation of the visual contents in a scene. The objects in the scene are represented by the node set $U = (B, O)$ of the scene graph, where B and O are the set of bounding boxes and labels for the objects, respectively. The edge set E consists of predicates describing the relationships of connected objects. Object i, j, and the predicate between them forms a relationship triplet $e_{ij} = (o_i, r_{ij}, o_j)$, where $r_{ij} \in \mathcal{R}$ means the class of predicate.

2.2 One-stage Scene Graph Generation

Describing the details of MED involves the mechanism of backbone methods. To this end, we briefly recap the pipeline of one-stage SGG models.

Inspired by the success of DETR [4] in Object Detection, recent works [19] propose to detect the whole relationship triplets at the same time for each triplet query. This procedure is formulated as follows:

$$\boldsymbol{g} = CNN(\boldsymbol{I}), \tag{1}$$

$$\boldsymbol{s}_i, \boldsymbol{b}_i^s, \boldsymbol{o}_i, \boldsymbol{b}_i^o = OH(\boldsymbol{q}_i^b, \boldsymbol{q}_i^o, \boldsymbol{g}), \tag{2}$$

$$\boldsymbol{p}_i = PH(\boldsymbol{q}_i^p, \boldsymbol{s}_i, \boldsymbol{o}_i, \boldsymbol{g}), \tag{3}$$

$$\boldsymbol{q}_i = [\boldsymbol{b}_i^s, \boldsymbol{b}_i^o, \boldsymbol{s}_i, \boldsymbol{o}_i, \boldsymbol{p}_i], \tag{4}$$

$$\boldsymbol{c}_i^s, \boldsymbol{c}_i^o = FC^o(\boldsymbol{s}_i, \boldsymbol{o}_i), \tag{5}$$

$$\boldsymbol{c}_i^p = FC^p(\boldsymbol{p}_i), \tag{6}$$

where CNN is a convolution neural network used to extract global features $\boldsymbol{g}$ from input image $\boldsymbol{I}$. The query $\boldsymbol{q}_i$ for a triplet i consists of a bounding box query $\boldsymbol{q}^b$, an object query $\boldsymbol{q}^o$, and a predicate query $\boldsymbol{q}^p$. The object head OH conducts object detection for the subject and object, with refined features $\boldsymbol{s}_i$, $\boldsymbol{o}_i$ for the two objects, and their bounding box $\boldsymbol{b}_i^s, \boldsymbol{b}_i^o$ as the results. Similarly, $\boldsymbol{p}_i$ represents the predicate features from the predicate head PH. Afterward, $\boldsymbol{q}_i$ is updated with the outputs of the object and predicate head. The procedure from Eq. 2 to Eq. 4 is repeated N times and followed by two classifiers (a series of fully connected layers) FC^o, FC^p to get the classification results.

2.3 Multi-expert De-biasing Method

Although one-stage SGG methods provide better bounding boxes and result in more precise SGG results, they still suffer from the aforementioned label ambiguity problem. As this problem mainly involves predicate labels, we fix the object head for a pre-trained biased model and explore reinforcing the predicate head. As illustrated in Fig. 3, we explored two variances of MED by conducting de-biasing from two aspects: (1) changing training strategy; (2) directly correcting logits using class frequency.

ACE-MED. Inspired by ally complementary experts (ACE) [5], we regard a predicate head as an expert that is most knowledgeable in its training set. Some of the experts can be trained with a narrower subset of labels so that they are not disturbed by the unseen frequent classes. The experts are designed to be complementary to each other and can help alleviate the training bias with a simple ensemble mechanism. Two experts are trained for the full set and the implicit set, respectively. Different from [5], we train the whole

Method	SGDet					
	Recall@50	Recall@100	mRecall@50	mRecall@100	M@50	M@100
PCPL [13]	14.6	18.6	9.5	11.7	12.1	15.2
TDE [7]	16.9	20.3	8.2	9.8	12.6	15.1
CogTree [6]	20.0	22.1	10.4	11.8	15.2	17.0
BA-SGG [18]	23.0	26.9	13.5	15.6	18.3	21.3
RTPB [3]	19.0	22.5	13.1	15.5	16.1	19.0
GCL [15]	18.4	22.0	16.8	19.3	17.6	20.7
PPDL [14]	21.2	23.9	11.4	13.5	16.3	18.7
SSR-CNN-LA [19]	23.7	27.3	**18.6**	**22.5**	21.2	24.9
SSR-CNN* [19]	33.5	38.4	8.6	10.3	21.1	24.4
ACE-MED	**33.0**	**37.8**	10.3	12.5	21.7	25.1
LA-MED	32.8	37.6	11.1	13.5	**21.9**	**25.6**

Table 1: SGDet performance of different de-biasing SGG methods on VG dataset [10]. * means a biased method as the baseline. For a fair comparison, all the methods are based on MOTIFS [12] if possible. The mean values are calculated from precise values.

Method	R@50	R@100	mR@50	mR@100	M@50	M@100
α=0	33.5	38.4	8.6	10.3	21.1	24.4
α=0.05	33.4	38.3	9.5	11.5	21.4	24.9
α=0.075	33.4	38.2	9.3	11.5	21.3	24.8
α=0.1	31.9	36.6	11.6	14.0	21.7	25.3
α=0.125	32.8	37.6	11.1	13.5	**21.9**	**25.6**

Table 2: Parameter analysis on the of α in Eq. 9. "α=0" means the SSR-CNN [19] model used as the baseline method to show the degradation of model performance on Recall@K.

predicate head rather than the classifier only, for more informative representations. We finetune the implicit set expert based on the full set expert for two reasons: (1) They can share the same object head, which is necessary if we want to fuse their results. (2) Training on explicit predicates can help the model learn some general patterns, so as to produce better features for implicit predicates. We combine the two experts as illustrated in Fig. 3. We fuse the logits of the two predicate heads by their weighted sum:

$$z_i^{full} = f^{full}(\boldsymbol{p}_i^{full}), \tag{7}$$

$$z_i^{im} = f^{im}(\boldsymbol{p}_i^{im}), \tag{8}$$

$$z_i = z_i^{full} + \alpha * z_i^{im}, \tag{9}$$

where f^{full}, f^{im} are the fully-connected networks for the two predicate heads, and z_i^{full}, z_i^{im} are the corresponding logits. α is an empirical factor ranging from 0 to 1.

LA-MED. MED is inspired by the multi-expert network, which makes experts different by switching the training set. However, we find that the implicit set expert is not limited to a fixed method. We extend the aforementioned method and utilize a more powerful implicit set expert, by adopting logit adjustment (LA) [19] on the resulting logits. After calculating the frequency of each predicate class, we use the log of it as a bias for predictions. After multiplying a tuning factor τ, we get the residual of logits. The final logits of the implicit set expert are then calculated using the original logits minus such residual. The remaining procedure is the same as ACE-MED.

3 EXPERIMENTS

3.1 Dataset and Settings

Visual Genome (VG) dataset [10] is the most popular dataset for SGG. We follow [12] to get a subset of the VG dataset (VG150), which has 150 object categories and 50 predicate categories. Limited by the structure of the one-stage model [19], the proposed method is only evaluated on SGDet.

3.2 Implementation Details

All the compared methods use ResNeXt-101-FPN [8, 17] as the CNN backbone. For ACE-MED, we optimize the network by AdamW [9] and set the initial learning rate and batch size to be $5 \times 10\text{-}7$ and 4, respectively. We finetune the implicit model with 800 queries for 20k iterations. For LA-MED, the τ of LA is set to be 0.65. Other settings are the same as in [19].

3.3 Quantitative Results and Parameter Analysis

The comparison results with the state-of-the-art methods are shown in Table 1. "ACE-MED" and "LA-MED" means MED variances using the methods described in Sec. 2.3. We can see that MED achieves significant improvements on mRecall@K while making minor losses on Recall@K. As a result, LA-MED achieves superior performance over previous methods (4%,3%) on the mean of Recall@K and mRecall@K (M@K). Compared to SSR-CNN-LA [19], which achieves the highest Recall@K, LA-MED significantly improves on Recall@K (38%,38%). This result shows the superior ability of the proposed on maintaining the capability of the model on recognizing head predicates while producing unbiased scene graphs. Specifically, the

values of Recall@100 on *on* of SSR-CNN, LA-MED and SSR-CNN-LA are 42.7, 41.8 and 13.9 respectively. The values of Recall@100 on *sitting on* are 16.5, 23.8 and 30.5 respectively. These results show the significant efficacy of MED in alleviating the label ambiguity problem.

For parameter analysis, we investigate the influence of the values of α in Eq. 9. In particular, we use "LA-MED" described in 2.3. The experimental results are summarized in Table 2. We choose the final value of α to be 0.125 according to M@K.

4 CONCLUSION

This paper revisits the training bias problem in scene graph generation and reveals another critical cause of it: the label ambiguity problem. We find that single-model method will inevitably be biased to either head or tail classes. To this end, we propose a multi-expert de-biasing method (MED) that fuses multiple models to ensure the model performance on both parts of the label set. Experimental results show that the proposed method provides significant gains on mRecall@K while giving a minor influence on Recall@K, and achieves a state-of-the-art on the mean of Recall@K and mRecall@K.

5 ACKNOWLEDGEMENTS

This work is supported in part by the National Natural Science Foundation of China Under Grants No.62176253 and No.U20B2066.

REFERENCES

[1] Desai A., Wu T. Y., Tripathi S., and Vasconcelos N. 2021. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE International Conferenceon Computer Vision*. IEEE, 6163–6171.

[2] Goel A., Fernando B., Keller F., and H. Bilen. 2022. Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 15596–15606.

[3] Chao C., Yibing Z., Baosheng Y., Liu L., Yong L., and Bo D. 2022. Resistance trainingusing prior bias: toward unbiased scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[4] Nicolas C., Francisco M., Gabriel S., Nicolas U., Alexander K., and Sergey Z. 2020. Endto-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. 213–229.

[5] Cai J., Wang Y., and Hwang J. N. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE International Conferenceon Computer Vision*. IEEE, 112–121.

[6] Tang K., Zhang H., Wu B., Luo W., and Liu W. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6619–6628.

[7] Tang K., Niu Y., Huang J., Shi J., and Zhang H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3716–3725.

[8] Tsung-Yi L., Piotr D., Ross B. G., Kaiming H., Bharath H., and Serge J. B. 2017. Feature pyramidnetworks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2117–2125.

[9] llya L. and Frank H. 2019. Decoupled weight decayregularization. In *The International Conference on Learning Representations*. IEEE.

[10] Krishna R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., Kalantidis Y., Li L.-J., Shamma D. A., and et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal on Computer Vision*, Vol. 123. 32–73.

[11] Li R., Zhang S., Wan B., and He X. 2021. Bipartite Graph Network with Adaptive Message Passing for Unbiased Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 11109–11119.

[12] Zellers R., Yatskar M., Thomson S., and Choi Y. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5831–5840.

[13] Yan S., Shen C., Jin Z., Huang J., Jiang R., Chen Y., and Hua X. 2020. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 265–273.

[14] Li W., Zhang H., Bai Q., Zhao G., Jiang N., and Yuan X. 2022. PPDL: Predicate Probability Distribution Based Loss for Unbiased Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 19447–19456.

[15] Dong X., Gan T., Song X., Wu J., Cheng Y., and Nie L. 2022. Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 19427–19436.

[16] Lin X., Ding C., Zeng J., and Tao D. 2020. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3746–3753.

[17] Saining X, Ross B. G., Piotr D., Zhuowen T., and Kaiming H. 2017. Aggregated residual transformations fordeep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5987–5995.

[18] Guo Y., Gao L., Wang X., Hu Y., Xu X., Lu X., and Song J. 2021. From general to specific: Informative scene graph generation via balance adjustment. In *Proceedings of the IEEE International Conferenceon Computer Vision*. IEEE, 16383–16392.

[19] Teng Y. and Wang L. 2022. Structured sparse r-cnn for direct scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 19437–19446.