Fine-Grained Scene Graph Generation with Data Transfer

Ao Zhang[1]*, Yuan Yao[2]*, Qianyu Chen[2], Wei Ji[1]†, Zhiyuan Liu[2]†,
Maosong Sun[2], and Tat-Seng Chua[1]

[1] Sea-NExT Joint Lab, Singapore
School of Computing, National University of Singapore, Singapore
[2] Department of Computer Science and Technology
Institute for Artificial Intelligence, Tsinghua University, Beijing, China
Beijing National Research Center for Information Science and Technology, China
aozhang@u.nus.edu, yaoyuanthu@163.com

Abstract. Scene graph generation (SGG) is designed to extract (*subject*, `predicate`, *object*) triplets in images. Recent works have made a steady progress on SGG, and provide useful tools for high-level vision and language understanding. However, due to the data distribution problems including long-tail distribution and semantic ambiguity, the predictions of current SGG models tend to collapse to several frequent but uninformative predicates (*e.g.*, `on`, `at`), which limits practical application of these models in downstream tasks. To deal with the problems above, we propose a novel Internal and External Data Transfer (IETrans) method, which can be applied in a plug-and-play fashion and expanded to large SGG with 1,807 predicate classes. Our IETrans tries to relieve the data distribution problem by automatically creating an enhanced dataset that provides more sufficient and coherent annotations for all predicates. By applying our proposed method, a Neural Motif model doubles the macro performance for informative SGG. The code and data are publicly available at https://github.com/waxnkw/IETrans-SGG.pytorch.

Keywords: Scene graph generation, Plug-and-play, Large-scale

1 Introduction

Scene graph generation (SGG) aims to detect relational triplets (*e.g.*, (*man*, `riding`, *bike*)) in images. As an essential task for connecting vision and language, it can serve as a fundamental tool for high-level vision and language tasks, such as visual question answering [2,25,29,14], image captioning [33,7], and image retrieval [10,27,28]. However, existing SGG methods can only make correct predictions on a limited number of predicate classes (*e.g.*, 29 out of 50 pre-defined classes [39]), among which a majority of predicates are trivial and uninformative (*e.g.*, `on`, `and`, `near`). This undermines the application of SGG for

∗ indicates equal contribution.

† Corresponding author: jiwei@nus.edu.sg, liuzy@tsinghua.edu.cn

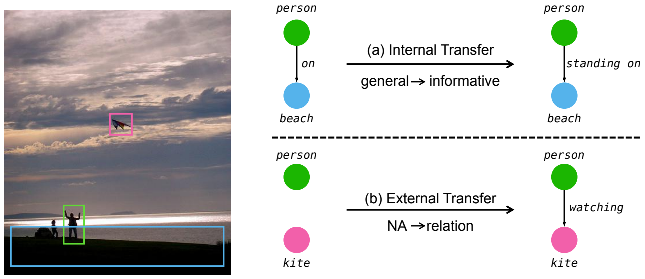


Fig. 1. Generate an enhanced dataset automatically for better model training with: (a) **Internal Transfer**: Specify general predicate annotations as informative ones. (b) **External Transfer**: Relabel missed relations from `NA`.

downstream tasks. To address the limitation, we first identify two main problems that need to deal with:

- **Long-tail problem:** the problem refers to the phenomenon that annotations mainly concentrate on a few head predicate classes, and are much sparse in most tail predicate classes. For example, in Visual Genome[12], there are over 100K samples for the top 5 predicate classes, while over 90% of predicate classes have less than 10 samples. As a result, the performance of tail predicate classes is poor due to the lack of effective supervision.
- **Semantic ambiguity:** many samples can be described as either general predicate class (*e.g.*, `on`) or an informative one (*e.g.*, `riding`). However, data annotators prefer some general (and thus uninformative) predicate classes to informative ones for simplicity. This causes conflicts in the widely adopted single-label optimization since different labels are annotated for the same type of instances. Thus, even when the informative predicates have enough training samples, the prediction will easily collapse to the general ones.

To address the problems mentioned above, recent works propose to use resampling [6,13], reweighting [32], and post-processing methods [23,8]. However, we argue that these problems can be better alleviated by enhancing the existing dataset into a reasonable dataset, that contains more abundant training samples for tail classes and also provides coherent annotations for different classes.

To this end, we propose a novel framework named **I**nternal and **E**xternal data **Trans**fer (**IETrans**), which can be equipped to different baseline models in a plug-and-play applied in a fashion. As shown in Figure 1, we automatically transfer data from general predicates to informative ones (**Internal Transfer**) and relabel relational triplets missed by annotators (**External Transfer**). (1) **For internal transfer**, we first identify the general-informative relational pairs based on the confusion matrix, and then conduct a triplet-level data transfer from general ones to informative ones. The internal transfer will not only alleviate the optimization conflict caused by semantic ambiguity but also provide more data for tail classes; (2) **For external transfer**, there exist many positive samples missed by annotators [12,19], which are usually treated as negative samples by current methods. However, this kind of data can be considered as a potential data source, covering a wide range of predicate categories. Inspired by Visual Distant Supervision [35] which employs `NA` samples for pre-training, we also consider the `NA` samples, which are the union of negative and missed annotated samples. The missed annotated samples can be relabeled to provide more training samples.

It is worth noting that both internal transfer and external transfer are indispensable for improving SGG performance. Without the internal transfer, the external transfer will suffer from the semantic ambiguity problem. Meanwhile, the external transfer can further provide training samples for tail classes, especially for those that have weak semantic connection with head classes.

Exhaustive experiments show that our method is both adaptive to different baseline models and expansible to large-scale SGG. We equip our data augmentation method with 4 different baseline models and find that it can significantly boost all models' macro performance and achieve SOTA performance for F@K metric, a metric for overall evaluation. For example, a Neural Motif Model with our proposed method can double the mR@100 performance and achieve the highest F@100 among all model-agnostic methods on predicate classification task of the widely adopted VG-50 [31] benchmark.

To validate the scalability of our proposed method, we additionally propose a new benchmark with 1,807 predicate classes (VG-1800), which is more practical and challenging. To provide a reliable and stable evaluation, we manually remove unreasonable predicate classes and make sure there are over 5 samples for each predicate class on the test set. On VG-1800, our method achieves SOTA performance with significant superiority compared with all baselines. The proposed IETrans can make correct predictions on 467 categories, compared with all other baselines that can only correctly predict less than 70 categories. While the baseline model can only predict relations like (*cloud*, `in`, *sky*) and (*window*, `on`, *building*), our method enables to generate informative ones like (*cloud*, `floating through`, *sky*) and (*window*, `on exterior of`, *building*).

Our main contributions are summarized as follows: (1) To cope with the long-tail problem and semantic ambiguity in SGG, we propose a novel IETrans framework to generate an enhanced training set, which can be applied in a plug-and-play fashion. (2) We propose a new VG-1800 benchmark, which can provide reliable and stable evaluation for large-scale SGG. (3) Comprehensive experiments demonstrate the effectiveness of our IETrans in training SGG models.

2 Related Works

2.1 Scene Graph Generation

As an important tool of connecting vision and language, SGG [31,19,15] has drawn widespread attention from the community. SGG is first proposed as visual relation detection (VRD) [19], in which each relation is detected independently. Considering that relations are highly dependent on their context, [31] further proposes to formulate VRD as a dual-graph generation task, which can incorporate context information. Based on [31], different methods [18,24,39] are proposed to refine the object and relation representations in the scene graph. For example, [18] proposes a novel message passing mechanism that can encode edge directions into node representations. Recently, CPT [36] and PEVL [34] propose to employ pre-trained vision-language models for SGG. CPT shows promising few-shot ability and PEVL shows much better performance than models training from scratch.

2.2 Informative Scene Graph Generation

Although making steady progress on improving recall on SGG task, [24,3] point out that the predictions of current SGG models are easy to collapse to several general and trivial predicate classes. Instead of only focusing on recall metric, [24,3] propose a new metric named mean recall, which is the average recall of all predicate classes. [23] employs a causal inference framework, which can eliminate data bias during the inference process. CogTree [37] proposes to leverage the semantic relationship between different predicate classes, and design a novel CogTree loss to train models that can make informative predictions. In BGNN [13], the authors design a bi-level resampling strategy, which can help to provide a more balanced data distribution during the training process. However, previous works of designing new loss or conducting resampling, only focus on predicate-level adjustment, while the visual relation is triplet-level. For example, given the subject *man* and object *skateboard*, the predicate `riding` is an informative version of `standing on`, while given the subject *man* and object *horse*, `riding` will not be an informative alternative of `standing on`. Thus, instead of using less precise predicate-level manipulation, we employ a triplet-level transfer.

2.3 Large-scale Scene Graph Generation

In the last few years, there are some works [1,41,40,35] focusing on large-scale SGG. Then, how to provide a reliable evaluation is an important problem. [40] first proposes to study large-scale scene graph generation and makes a new split of Visual Genome dataset named VG80K, which contains 29,086 predicate categories. However, the annotations are too noisy to provide reliable evaluation. To cope with this problem, [1] further cleans the dataset, and finally reserves 2,000 predicate classes. However, only 1,429 predicate classes are contained in the test set, among which 903 relation categories have no more than 5 samples. To provide enough samples for each predicate class' evaluation, we re-split the Visual Genome to ensure each predicate class on the test set has more than 5 samples, and the total predicate class number is 1,807. For the proposed methods, [40] employs a triplet loss to regularize the visual representation with the constraint on word embedding space. RelMix [1] proposes to conduct data augmentation with the format of feature mixup. Visual distant supervision [35] pre-trains the model on relabeled `NA` data with the help of a knowledge base and achieve significant improvement on a well-defined VG setting without semantic ambiguity.

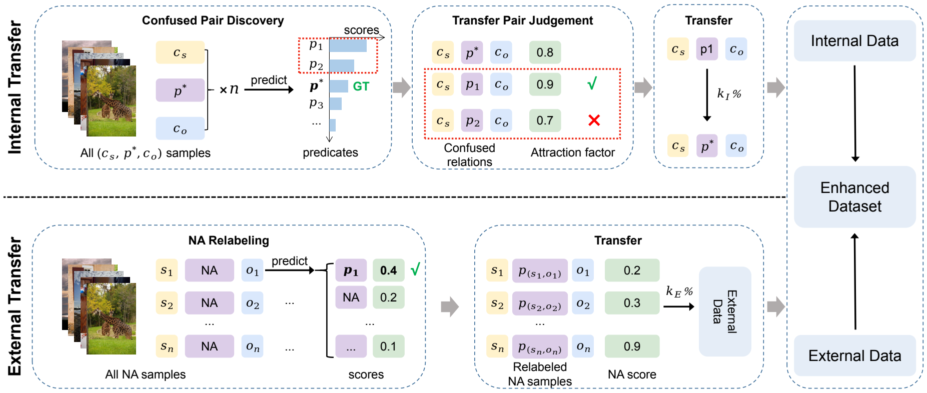


Fig. 2. Illustration of our proposed IETrans to generate an enhanced dataset. **Internal transfer** is designed to transfer data from general predicate to informative ones. **External transfer** is designed to relabel `NA` data. To avoid misunderstanding, (c_s, p^*, c_o) is a relational triplet class. $(s_i, p_{(s_i,o_i)}, o_i)$ represents a single relational triplet instance.

However, the data extension will be significantly limited by the semantic ambiguity problem. To deal with this problem, we propose an internal transfer method to generate informative triplets.

3 Method

In this section, we first introduce the internal data transfer and external data transfer, respectively, and then elaborate how to utilize them collaboratively. Figure 2 shows the pipeline of our proposed IETrans.

The goal of our method is to generate an enhanced dataset automatically, which should provide more training samples for tail classes, and also specify general predicate classes as informative ones. Concretely, as shown in Figure 1, the general relation `on` between (*person*, *beach*) need to be specified as more informative one `standing on`, and the missed annotations between (*person*, *kite*) can be labeled so as to provide more training samples.

3.1 Problem Definition

Scene Graph Generation. Given an image I, a scene graph corresponding to I has a set of objects $O = \{(b_i, c_i)\}_{i=1}^{N_o}$ and a set of relational triplets $E = \{(s_i, p_{(s_i,o_i)}, o_i)\}_{i=1}^{N_e}$. For each object (b_i, c_i), it consists of an object bounding box $b_i \in \mathbb{R}^4$ and an object class c_i which belongs to the pre-defined object class set $\mathcal{C}$. With $s_i \in O$ and $o_i \in O$, p_i is defined as relation between them and belongs to the pre-defined predicate class set $\mathcal{P}$.

Inference. SGG is defined as a joint detection of objects and relations. Generally, an SGG model will first detect the objects in the image I. Based on the detected objects, a typical SGG model will conduct a feature refinement for objects and relation representation, and then classify the objects and relations.

3.2 Internal Data Transfer

The key insight of internal transfer is to transfer samples from general predicate classes to their corresponding informative ones, like the example shown in Figure 1. We split the process into 3 sub-steps, including **(1) Confusion Pair Discovery**: specify confused predicate pairs as potential general-informative pairs for given subject and object classes. **(2) Transfer Pair Judgement**: judge whether the candidate pair is valid. **(3) Triplet Transfer**: transfer data from the selected general predicate class to the corresponding informative one.

Confusion Pair Discovery. To find general predicate classes and corresponding informative ones, a straightforward way is to annotate the possible relation transitions manually. However, relations are highly dependent on the subject and object classes, i.e. relation is triplet-level rather than predicate-level. For example, given the entity pair *man* and *bike*, `riding` is a sub-type of `sitting on`, while for *man* and *skateboard*, `riding` shares different meaning with `sitting on`. In this condition, even under 50 predicate classes settings, the possible relation elements will scale up to an infeasible number for human annotation. Another promising alternative is to employ pre-defined knowledge bases, such as WordNet [20] and VerbNet [11]. However, existing knowledge bases are not specifically designed to cope with visual relation problems, which result in a gap between visual and textual hierarchies [26].

Thus, in this work, we try to specify the general-informative pairs by taking advantage of information within the dataset, and leave the exploration of external knowledge sources for future work. A basic observation is that **informative predicate classes are easily confused by general ones**. Thus, we can first find confusion pairs as candidate general-informative pairs. By observing the predictions of the pre-trained Motif [39] model, we find that the collapse from informative predicate classes to general ones, appears not only on the test set but also on the training set. As shown in Figure 3, the predicate classes `riding` and `sitting on` are significantly confused by a more general predicate class `on`.

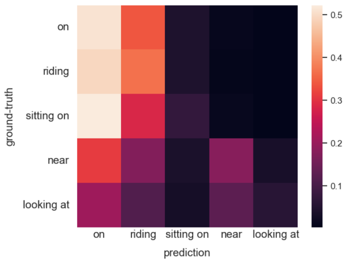


Fig. 3. Confusion matrix of Motif [39]'s prediction score on all entity pairs in VG training set with the subject *man* and the object *motorcycle*.

On the training set, given a relational triplet class (c_s, p, c_o), we use a pre-trained baseline model to predict predicate labels of all samples belonging to (c_s, p, c_o), and average their score vectors. We denote the aggregated scores for all predicates as $S = \{s_{p_i} | p_i \in \mathcal{P}\}$. From S, we select all predicate classes with higher prediction scores than the ground-truth annotation p, which can be formulated as $\mathcal{P}_c = \{p_i | s_{p_i} > s_p\}$. $\mathcal{P}_c$ can be considered as the most confusing predicate set for (c_s, p, c_o), which can serve as candidate transfer sources.

Transfer Pair Judgement. However, a confused predicate class does not equal to a general one. Sometimes, a general predicate can also be confused by an informative predicate. For example, in Figure 3, under the constraint of $c_s = \mathit{man}$ and $c_o = \mathit{motorcycle}$, the less informative predicate `sitting on` is confused by the more informative predicate `riding`. In this condition, it is not a good choice to transfer from `riding` to `sitting on`. Thus, we need to further select the truly general predicates from the candidate set $\mathcal{P}_c$.

To select the most possible general predicate classes from $\mathcal{P}_c$, we first introduce an important feature that is useful to recognize general predicate classes. As observed by [37], **the general predicate classes usually cover more diverse relational triplets**, while informative ones are limited. Based on this observation, we can define the attraction factor of a triplet category (c_s, p, c_o) as:

$$A(c_s, p, c_o) = \frac{1}{\sum_{c_i, c_j \in \mathcal{C}} \mathcal{I}(c_i, p, c_j)}, \tag{1}$$

where $\mathcal{C}$ is the object categories set and $\mathcal{I}(t)$ indicates whether the triplet category t exists in the training set, which can be formulated as:

$$\mathcal{I}(t) = \begin{cases} 1, & \text{if } t \in \text{training set} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The denominator of $A(c_s, p_i, c_o)$ is the number of relational triplet types containing p_i. Thus, $A(c_s, p_i, c_o)$ with smaller value means p_i is more likely to be a general predicate. Concretely, when $A(c_s, p_i, c_o) < A(c_s, p, c_o)$, we transfer data from (c_s, p_i, c_o) to (c_s, p, c_o).

However, only considering the number of relational triplet types also has drawbacks: some relational triplets with very limited number of samples (*e.g.*, only 1 or 2 samples) might be annotation noise, while these relational triplets are easily selected as transfer targets. Transferring too much data to such uncertain relational triplets will significantly degenerate models' performance. Thus, we further consider the number of each relational triplet and modify the attraction factor as:

$$A(c_s, p, c_o) = \frac{N(c_s, p, c_o)}{\sum_{c_i, c_j \in \mathcal{C}} \mathcal{I}(c_i, p, c_j) \cdot N(c_i, p, c_j)}, \tag{3}$$

where $N(t)$ denotes the number of instances with relational type t in the training set. With the attraction factor, we can further filter the candidate confusion set $\mathcal{P}_c$ to the valid transfer source for (c_s, p, c_o):

$$\mathcal{P}_s = \{p_i | (p_i \in \mathcal{P}_c) \wedge (A(c_s, p_i, c_o) < A(c_s, p, c_o))\}, \tag{4}$$

where $\wedge$ denotes the logical conjunction operator.

Triplet Transfer. Given the transfer source $\mathcal{P}_s$, we collect all samples in the training set satisfying:

$$T = \{(o_i, p_k, o_j) | (c_{o_i} = c_s) \land (p_k \in \mathcal{P}_s) \land (c_{o_j} = c_o)\}. \tag{5}$$

Then, we sort T by model's prediction score of p, and transfer the top k_I% samples to the target triplet category (c_s, p, c_o). Note that, a triplet instance may need to be transferred to more than one relational triplets. To deal with the conflict, we choose the target predicate with the highest attraction factor.

3.3 External Data Transfer

The goal of our external transfer is to relabel unannotated samples to excavate missed relational triplets, as the example shown in Figure 1.

NA Relabeling. `NA` samples refer to all unannotated samples in the training set, including both truly negative samples and missed positive samples. In external transfer, `NA` samples are directly considered as the transfer source and are relabeled as existing relational triplet types.

To get the `NA` samples, we first traverse all unannotated object pairs in images. However, considering that data transfer from all `NA` samples to all possible predicate classes will bring heavy computational burden, and inevitably increase the difficulty of conducting precise transfer, so as to sacrifice the quality of transferred data. Thus, we only focus on object pairs whose bounding boxes have overlaps and limit the possible transfer targets to existing relational triplet types. The exploration of borrowing zero-shot relational triplets from `NA` is left for future work.

Given a sample $(s, \texttt{NA}, o)$, we can get its candidate target predicate set as:

$$\mathrm{Tar}(s, \texttt{NA}, o) = \{p | (p \in \mathcal{P}) \land (N(c_s, p, c_o) > 0) \land (\mathrm{IoU}(b_s, b_o) > 0)\}, \tag{6}$$

where $\mathcal{P}$ denotes pre-defined predicate classes, b_s and b_o denote bounding boxes of s and o, and IoU denotes the intersection over union.

Given a triplet $(s, \texttt{NA}, o)$, the predicate class with the highest prediction score except for `NA` is chosen. The label assignment can be formulated as:

$$p_{(s,o)} = \mathop{\arg\max}_{p \in \mathrm{Tar}(s, \texttt{NA}, o)} (\phi^p(s, o)), \tag{7}$$

where $\phi^p(\cdot)$ denotes the prediction score of predicate p.

NA Triplet Transfer. To decide transfer or not, we rank all chosen $(s, \texttt{NA}, o)$ samples according to `NA` scores in an ascending order. The lower `NA` score means the sample is more likely to be a missed positive sample. Similar with internal transfer, we simply transfer the top k_E% data.

3.4 Integration

Internal transfer is conducted on annotated data and external transfer is conducted on unannotated data, which are orthogonal to each other. Thus, we can simply merge the data without conflicts. After obtaining the enhanced dataset, we re-train a new model from scratch and use the new model to make inferences on the test set.

4 Experiments

In this section, we first show the generalizability of our method with different baseline models and the expansibility to large-scale SGG. We also make ablation studies to explore the influence of different modules and hyperparameters. Finally, analysis is conducted to show the effectiveness of our method in enhancing the current dataset.

4.1 Generalizability with Different Baseline Models

We first validate the generalizability of our method with different baseline models and its effectiveness when compared with current SOTA methods.

Datasets. Popular VG-50 [31] benchmark is employed, which consists of 50 predicate classes and 150 object classes.

Tasks. Following previous works [31,23,39], we evaluate our model on three widely used SGG tasks: (1) **Predicate Classification (PREDCLS)** provides both localization and object classes, and requires models to recognize predicate classes. (2) **Scene Graph Classification (SGCLS)** provides only correct localization and asks models to recognize both object and predicate classes. (3) In **Scene Graph Detection (SGDET)**, models are required to first detect the bounding boxes and then recognize both object and predicate classes.

Metrics. Following previous works [37,24], we use Recall@K (**R@K**) and mean Recall@K (**mR@K**) as our metrics. However, different trade-offs between R@K and mR@K are made in different methods, which makes it hard to make a direct comparison. Therefore, we further propose an overall metric **F@K** to jointly evaluate R@K and mR@K, which is the harmonic average of R@K and mR@K.

Baselines. We categorize several baseline methods into two categories: (1) **Model-agnostic baselines.** They refers to methods that can be applied in a plug-and-play fashion. For this part, we include Resampling [13], TDE [23], CogTree [37], EBM [22], DeC [9], and DLFE [5]. (2) **Specific models.** We also include some dedicated designed models with strong performance, including KERN [4], KERN [4], GBNet [38], BGNN [13], DT2-ACBS [6], and PCPL [32].

Implementation Details. Following [23], we employ a pre-trained Faster-RCNN [21] with ResNeXt-101-FPN [16,30] backbone. In the training process, the parameters of the detector are fixed to reduce the computation cost. The batch size is set to 12, and the learning rate is 0.12, except for Transformer. We optimize all models with an SGD optimizer. Specifically, to better balance the data distribution, the external transfer will not be conducted for the top 15 frequent predicate classes. To avoid deviating too much from the original data

Table 1. Performance (%) of our method and other baselines on VG-50 dataset. **IETrans** denotes different models equipped with our IETrans. **Rwt** denotes using the reweighting strategy.

	Models	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
		R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100	R@50 / 100	mR@50 / 100	F@50 / 100
Specific	KERN [4]	65.8 / 67.6	17.7 / 19.2	27.9 / 29.9	36.7 / 37.4	9.4 / 10.0	15.0 / 15.8	27.1 / 29.8	6.4 / 7.3	10.4 / 11.7
	GBNet [38]	66.6 / 68.2	22.1 / 24.0	33.2 / 35.5	37.3 / 38.0	12.7 / 13.4	18.9 / 19.8	26.3 / 29.9	7.1 / 8.5	11.2 / 13.2
	BGNN [13]	59.2 / 61.3	30.4 / 32.9	40.2 / 42.8	37.4 / 38.5	14.3 / 16.5	20.7 / 23.1	31.0 / 35.8	10.7 / 12.6	15.9 / 18.6
	DT2-ACBS [6]	23.3 / 25.6	35.9 / 39.7	28.3 / 31.1	16.2 / 17.6	24.8 / 27.5	19.6 / 21.5	15.0 / 16.3	22.0 / 24.0	17.8 / 19.4
	PCPL [32]	50.8 / 52.6	35.2 / 37.8	41.6 / 44.0	27.6 / 28.4	18.6 / 19.6	22.2 / 23.2	14.6 / 18.6	9.5 / 11.7	11.5 / 14.4
Model-Agnostic	Motif [39]	64.0 / 66.0	15.2 / 16.2	24.6 / 26.0	38.0 / 38.9	8.7 / 9.3	14.2 / 15.0	31.0 / 35.1	6.7 / 7.7	11.0 / 12.6
	-TDE [23]	46.2 / 51.4	25.5 / 29.1	32.9 / 37.2	27.7 / 29.9	13.1 / 14.9	17.8 / 19.9	16.9 / 20.3	8.2 / 9.8	11.0 / 13.2
	-CogTree [37]	35.6 / 36.8	26.4 / 29.0	30.3 / 32.4	21.6 / 22.2	14.9 / 16.1	17.6 / 18.7	20.0 / 22.1	10.4 / 11.8	13.7 / 15.4
	-EBM [22]	- / -	18.0 / 19.5	- / -	- / -	10.2 / 11.0	- / -	- / -	7.7 / 9.3	- / -
	-DeC [9]	- / -	35.7 / 38.9	- / -	- / -	18.4 / 19.1	- / -	- / -	13.2 / 15.6	- / -
	-DLFE [5]	52.5 / 54.2	26.9 / 28.8	35.6 / 37.6	32.3 / 33.1	15.2 / 15.9	20.7 / 21.5	25.4 / 29.4	11.7 / 13.8	16.0 / 18.8
	-IETrans (ours)	54.7 / 56.7	30.9 / 33.6	39.5 / 42.2	32.5 / 33.4	16.8 / 17.9	22.2 / 23.3	26.4 / 30.6	12.4 / 14.9	16.9 / 20.0
	-IETrans+Rwt (ours)	48.6 / 50.5	**35.8** / **39.1**	**41.2** / **44.1**	29.4 / 30.2	**21.5** / **22.8**	**24.8** / **26.0**	23.5 / 27.2	**15.5** / **18.0**	**18.7** / **21.7**
	VCTree [24]	64.5 / 66.5	16.3 / 17.7	26.0 / 28.0	39.3 / 40.2	8.9 / 9.5	14.5 / 15.4	30.2 / 34.6	6.7 / 8.0	11.0 / 13.0
	-TDE [23]	47.2 / 51.6	25.4 / 28.7	33.0 / 36.9	25.4 / 27.9	12.2 / 14.0	16.5 / 18.6	19.4 / 23.2	9.3 / 11.1	12.6 / 15.0
	-CogTree [37]	44.0 / 45.4	27.6 / 29.7	33.9 / 35.9	30.9 / 31.7	18.8 / 19.9	23.4 / 24.5	18.2 / 20.4	10.4 / 12.1	13.2 / 15.2
	-EBM [22]	- / -	18.2 / 19.7	- / -	- / -	12.5 / 13.5	- / -	- / -	7.7 / 9.1	- / -
	-DLFE [5]	51.8 / 53.5	25.3 / 27.1	34.0 / 36.0	33.5 / 34.6	18.9 / 20.0	**24.2** / 25.3	22.7 / 26.3	11.8 / 13.8	15.5 / 18.1
	-IETrans (ours)	53.0 / 55.0	30.3 / 33.9	38.6 / 41.9	32.9 / 33.8	16.5 / 18.1	22.0 / 23.6	25.4 / 29.3	11.5 / 14.0	15.8 / 18.9
	-IETrans+Rwt (ours)	48.0 / 49.9	**37.0** / **39.7**	**41.8** / **44.2**	30.0 / 30.9	**19.9** / **21.8**	23.9 / **25.6**	23.6 / 27.8	**12.0** / **14.9**	**15.9** / **19.4**
	GPS-Net [18]	65.1 / 66.9	15.0 / 16.0	24.4 / 25.8	36.9 / 38.0	8.2 / 8.7	13.4 / 14.2	30.3 / 35.0	5.9 / 7.1	9.9 / 11.8
	-Resampling [13]	64.4 / 66.7	19.2 / 21.4	29.6 / 32.4	37.5 / 38.6	11.7 / 12.5	17.8 / 18.9	27.8 / 32.1	7.4 / 9.5	11.7 / 14.7
	-DeC [9]	- / -	**35.9** / 38.4	- / -	- / -	17.4 / 18.5	- / -	- / -	11.2 / 15.2	- / -
	-IETrans (ours)	52.3 / 54.3	31.0 / 34.5	38.9 / 42.2	31.8 / 32.7	17.0 / 18.3	22.2 / 23.5	25.9 / 28.1	14.6 / 16.5	18.7 / 20.8
	-IETrans+Rwt (ours)	47.5 / 49.4	34.9 / **38.6**	**40.2** / **43.3**	29.3 / 30.3	**19.8** / **21.6**	**23.6** / **25.2**	23.1 / 25.0	**16.2** / **18.8**	**19.0** / **21.5**
	Transformer [23]	63.6 / 65.7	17.9 / 19.6	27.9 / 30.2	38.1 / 39.2	9.9 / 10.5	15.7 / 16.6	30.0 / 34.3	7.4 / 8.8	11.9 / 14.0
	-CogTree [37]	38.4 / 39.7	28.4 / 31.0	32.7 / 34.8	22.9 / 23.4	15.7 / 16.7	18.6 / 19.5	19.5 / 21.7	11.1 / 12.7	14.1 / 16.0
	-IETrans (ours)	51.8 / 53.8	30.8 / 34.7	38.6 / 42.2	32.6 / 33.5	17.4 / 19.1	22.7 / 24.3	25.5 / 29.6	12.5 / 15.0	16.8 / 19.9
	-IETrans+Rwt (ours)	49.0 / 50.8	**35.0** / **38.0**	**40.8** / **43.5**	29.6 / 30.5	**20.8** / **22.3**	**24.4** / **25.8**	23.1 / 27.1	**15.0** / **18.1**	**18.2** / **21.7**

distribution, the frequency bias item calculated from the original dataset is applied to our IETrans in the inference stage. For internal and external transfer, the k_I is set to 70% and k_E is set to 100%. Please refer to the Appendix for more details.

Comparison with SOTAs. We report the results of our IETrans and baselines for VG-50 in Table 1. Based on the observation of experimental results, we have summarized the following conclusions:

Our IETrans is adaptive to different baseline models. We equip our method with 4 different models, including Motif [39], VCTree [24], GPS-Net [18], and Transformer [23]. The module architectures range from conventional CNN to TreeLSTM (VCTree) and self-attention layers (Transformer). The training algorithm contains both supervised training and reinforcement learning (VCTree). Despite the model diversity, our IETrans can boost all models' mR@K metric and also achieve competitive F@K performance. For example, our IETrans can double mR@50/100 and improve the overall metric F@50/100 for over 9 points across all 3 tasks for GPS-Net.

Compared with other model-agnostic methods, our method outperforms all of them in nearly all metrics. For example, when applying IETrans to Motif on PREDCLS, our model can achieve the highest R@50/100 and mR@50/100 among all model-agnostic baselines except for DeC. After adding the reweighting strategy, our IETrans can outperform DeC on mR@K.

Compared with strong specific baselines, our method can also achieve competitive performance on mR@50/100, and best overall performance on F@50/100. Considering mR@50/100, our method with reweighting strategy is slightly lower than DT2-ACBS on SGCLS and SGDET tasks, while our method performs much better than them on R@50/100 (*e.g.*, 24.3 points of VCTree on PREDCLS task). For overall comparison considering F@50/100 metrics, our VCTree+IETrans+Rwt can achieve the best F@50/100 on PREDCLS and Motif+IETrans+Rwt achieves the best F@50/100 in SGCLS and SGDET task.

4.2 Expansibility to Large-Scale SGG

We also validate our IETrans on VG-1800 dataset to show its expansibility to large-scale scenarios.

Datasets. We re-split the Visual Genome dataset to create a VG-1800 benchmark, which contains 70,098 object categories and 1,807 predicate categories. Different from previous large-scale VG split [40,1], we clean the misspellings and unreasonable relations manually and make sure all 1,807 predicate categories appear on both training and test set. For each predicate category, there are over 5 samples on the test set to provide a reliable evaluation. Detailed statistics of VG-1800 dataset are provided in Appendix.

Tasks. In this work, we mainly focus on the predicate-level recognition ability and thus compare models on PREDCLS in the main paper. For SGCLS results, please refer to the Appendix.

Metrics. Following [40,1], we use accuracy (**Acc**) and mean accuracy upon all predicate classes (**mAcc**). Similar to VG-50, the harmonic average of two metrics is reported as **F-Acc**. In addition, we also report the number of predicate classes that the model can make at least one correct prediction, denoted as **Non-Zero**.

Baselines. We also include model-agnostic baselines including Focal Loss [17], TDE [23], and RelMix [1], and a specific model BGNN [13].

Implementation Details. Please refer to the Appendix for details.

Table 2. Performance of our method and baselines on VG-1800 dataset. **IETrans** denotes the Motif [39] model trained using our IETrans. To better compare with baselines, we show different Acc and mAcc trade-offs by setting different k_I.

Models	Top-1				Top-5				Top-10			
	Acc	mAcc	F-Acc	Non-Zero	Acc	mAcc	F-Acc	Non-Zero	Acc	mAcc	F-Acc	Non-Zero
BGNN [13]	**61.55**	0.59	1.16	37	**85.64**	2.33	4.5	111	**90.07**	3.91	7.50	139
Motif [39]	59.63	0.61	1.21	47	84.82	2.68	5.20	112	89.44	4.37	8.33	139
-Focal Loss	54.65	0.26	0.52	14	79.69	0.79	1.56	27	85.21	1.36	2.68	41
-TDE [23]	60.00	0.62	1.23	45	85.29	2.77	5.37	119	89.92	4.65	8.84	152
-RelMix [1]	60.16	0.81	1.60	65	85.31	3.27	6.30	134	89.91	5.17	9.78	177
-IETrans (k_I = 10%) (ours)	56.66	1.89	3.66	202	83.99	8.23	14.99	419	89.71	13.06	22.80	530
-IETrans (k_I = 90%) (ours)	27.40	**4.70**	**8.02**	**467**	72.48	**13.34**	**22.53**	**741**	83.50	**19.12**	**31.12**	**865**

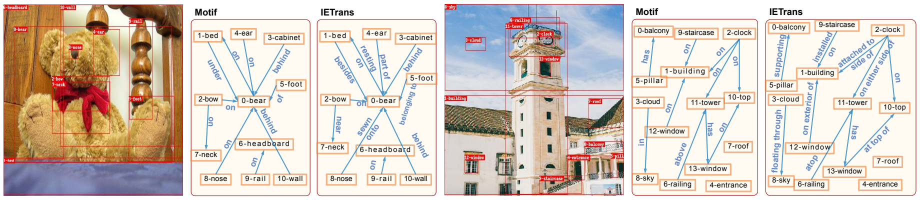


Fig. 5. Visualization of raw Motif model and Motif equipped with our IETrans.

Comparison with SOTAs. Performance of our method and baselines are shown in Table 2. Based on the observation of experimental results, we have summarized the following conclusions:

Our model can successfully work on large-scale settings. On VG-1800 dataset, the long-tail problem is even exacerbated, where hundreds of predicate classes have only less than 10 samples. Simply increasing loss weight (Focal Loss) on tail classes can not work well. Different from these methods, our IETrans can successfully boost the performance on mAcc while keeping competitive results on Acc. For quantitative comparison, our **IETrans ($k_I = 10\%$)** can significantly improve the performance on top-10 mAcc (*e.g.*, 19.12% vs. 4.37%) while maintaining comparable performance on Acc.

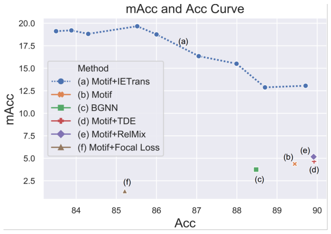


Fig. 4. The mAcc and Acc curve. (a) is our IETrans method. k_I is tuned to generate the blue curve. (b-f) are baselines.

Compared with different baselines, our method can outperform them for overall evaluation. As shown in Table 2, our **IETrans ($k_I = 90\%$)** can achieve best performance on F-Acc, which is over 3 times of the second highest baseline, RelMix, a method specifically designed for large-scale SGG. To make the visualized comparison, we plot a curve of IETrans with different Acc and mAcc trade-offs by tuning k_I, and show the performance of other baselines as points. As shown in Figure 4, all baselines drawn as points are under our curve, which means our method can achieve better performance than them. Moreover, our IETrans ($k_I = 90\%$) can make correct predictions on 467 predicate classes for top-1 results, while the **Non-Zero** value of all other baselines are less than 70.

Case Studies. To show the potential of our method for real-world application, we provide some cases in Figure 5. We can observe that our IETrans can help to generate more informative predicate classes while keeping faithful to the image content. For example, when the Motif model only predict relational triplets like (*foot*, `of`, *bear*), (*nose*, `on`, *bear*) and (*cloud*, `in`, *sky*), our IETrans can generate more informative ones as (*foot*, `belonging to`, *bear*), (*nose*, `sewn onto`, *bear*), and (*cloud*, `floating through`, *sky*).

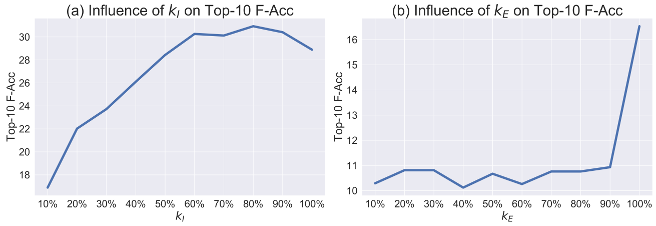


Fig. 7. (a) The influence of k_I in the Top-10 F-Acc with only internal transfer. (b) The influence of k_E in the Top-10 F-Acc with only external transfer.

4.3 Ablation Studies

In this part, we analyse the influence of internal transfer, external transfer, and corresponding parameters, k_I and k_E.

Influence of Internal Transfer. As shown in Figure 6, only using external transfer (yellow cube) is hard to boost the mAcc performance as much as IETrans. The reason is that although introducing samples for tail classes, they will still be suppressed by corresponding general ones. However, by introducing internal transfer (green point) to cope with semantic ambiguity problem, the performance (red cross) can be improved significantly on mAcc, together with minor performance drop on Acc.

Influence of External Transfer. Although internal transfer can achieve huge improvement on mAcc compared with Motif, its performance is poor compared with IETrans, which shows the importance of further introducing training data by external transfer. Integration of two methods can maximize the advantages of data transfer.

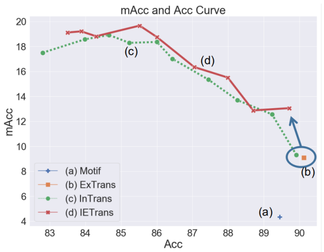
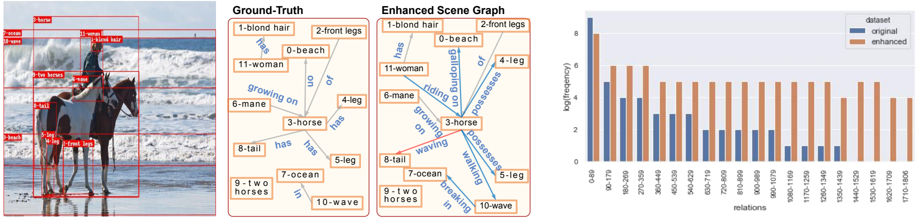


Fig. 6. The mAcc and Acc curve. (a) Normally trained Motif. (b) ExTrans: external transfer. (c) InTrans: internal transfer. (d) Our proposed IETrans. k_I is tuned to generate a curve. The blue circle and arrow mean that combining ExTrans and InTrans can lead to the pointed result.

Influence of k_I. As shown in Figure 7 (a), with the increase of k_I, the top-10 F-Acc will increase until $k_I = 80\%$, and begin to decrease when $k_I > 80\%$. The phenomenon indicates that a large number of general predicates can be interpreted as informative ones. Moving these predicates to informative ones will boost the overall performance. However, there also exists some predicates that can not be interpreted as informative ones or be modified suitably by current methods, which is harmful to the performance of models.

Influence of k_E. As shown in Figure 7 (b), the overall performance increases slowly with the initial 90% transferred data, but improves significantly with the rest 10%. Note that, the data is ranked according to the `NA` score, which means that the last 10% data is actually



(a) Enhanced Scene Graph Visualization

(b) Distribution Change

Fig. 8. (a) **Enhanced Scene Graph Visualization.** Gray line denotes unchanged relation. Blue line denotes changed and reasonable relation. Red line denotes changed but unreasonable relation. (b) **Distribution Change.** The comparison between distributions of original dataset and enhanced dataset for VG-1800. The x-axis is the relation id intervals from head to tail classes. The y-axis is the corresponding log-frequency.

what the model considered as most likely to be truly negative. The phenomenon indicates that the model may easily classify tail classes as negative samples, while this part of data is of vital significance for improving the model's ability of making informative predictions.

4.4 Analysis of Enhanced Dataset

Enhanced Scene Graph Correctness. We investigate the correctness of enhanced scene graphs from the instance level. An example is shown in Figure 8(a). We can see that the IETrans is less accurate on VG-1800, which indicates that it is more challenging to conduct precise data transfer on VG-1800.

Distribution Change. The distribution change is shown in Figure 8(b). We can see that our IETrans can effectively supply samples for non-head classes.

5 Conclusion

In this paper, we design a data transfer method named IETrans to generate an enhanced dataset for the SGG. The proposed IETrans consists of an internal transfer module to relabel general predicate classes as informative ones and an external transfer module to complete missed annotations. Comprehensive experiments are conducted to show the effectiveness of our method. In the future, we hope to extend our method to other large-scale visual recognition problems (*e.g.*, image classification, semantic segmentation) with similar challenges.

Acknowledgements. This research is funded by Sea-NExT Joint Lab, Singapore. The research is also supported by the National Key Research and Development Program of China (No. 2020AAA0106500).

References

1. Abdelkarim, S., Agarwal, A., Achlioptas, P., Chen, J., Huang, J., Li, B., Church, K., Elhoseiny, M.: Exploring long tail visual relationship recognition with large vocabulary. In: Proceedings of ICCV. pp. 15921–15930 (2021)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Proceedings of ICCV. pp. 2425–2433 (2015)
3. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of CVPR. pp. 6163–6171 (2019)
4. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of CVPR. pp. 6163–6171 (2019)
5. Chiou, M.J., Ding, H., Yan, H., Wang, C., Zimmermann, R., Feng, J.: Recovering the unbiased scene graphs from the biased ones. In: Proceedings of ACM Multimedia. pp. 1581–1590 (2021)
6. Desai, A., Wu, T.Y., Tripathi, S., Vasconcelos, N.: Learning of visual relations: The devil is in the tails. In: Proceedings of ICCV. pp. 15404–15413 (2021)
7. Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G.: Unpaired image captioning via scene graph alignments. In: Proceedings of ICCV. pp. 10323–10332 (2019)
8. Guo, Y., Gao, L., Wang, X., Hu, Y., Xu, X., Lu, X., Shen, H.T., Song, J.: From general to specific: Informative scene graph generation via balance adjustment. In: Proceedings of ICCV. pp. 16383–16392 (2021)
9. He, T., Gao, L., Song, J., Cai, J., Li, Y.F.: Semantic compositional learning for low-shot scene graph generation. In: Proceedings of ICCV. pp. 2961–2969 (2021)
10. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of CVPR. pp. 3668–3678 (2015)
11. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: Extending VerbNet with novel verb classes. In: LREC. pp. 1027–1032 (2006)
12. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. IJCV pp. 32–73 (2017)
13. Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: Proceedings of CVPR. pp. 11109–11119 (2021)
14. Li, Y., Wang, X., Xiao, J., Ji, W., Chua, T.S.: Invariant grounding for video question answering. In: Proceedings of CVPR. pp. 2928–2937 (2022)
15. Li, Y., Yang, X., Shang, X., Chua, T.S.: Interventional video relation detection. In: Proceedings of ACM Multimedia. pp. 4091–4099 (2021)
16. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of CVPR. pp. 2117–2125 (2017)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of ICCV. pp. 2980–2988 (2017)
18. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-Net: Graph property sensing network for scene graph generation. In: Proceedings of CVPR. pp. 3746–3753 (2020)
19. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Proceedings of ECCV. pp. 852–869 (2016)
20. Miller, G.A.: WordNet: a lexical database for english. Communications of the ACM pp. 39–41 (1995)

21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of NIPS. pp. 91–99 (2015)
22. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.: Energy-based learning for scene graph generation. In: Proceedings of CVPR. pp. 13936–13945 (2021)
23. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of CVPR. pp. 3716–3725 (2020)
24. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of CVPR. pp. 6619–6628 (2019)
25. Teney, D., Liu, L., van Den Hengel, A.: Graph-structured representations for visual question answering. In: Proceedings of CVPR. pp. 1–9 (2017)
26. Wan, A., Dunlap, L., Ho, D., Yin, J., Lee, S., Jin, H., Petryk, S., Bargal, S.A., Gonzalez, J.E.: NBDT: Neural-backed decision trees pp. 1027–1032 (2021)
27. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph matching for relationship-aware image-text retrieval. In: Proceedings of WACV. pp. 1508–1517 (2020)
28. Wei, M., Chen, L., Ji, W., Yue, X., Chua, T.S.: Rethinking the two-stage framework for grounded situation recognition. In: Proceedings of AAAI. pp. 2651–2658 (2022)
29. Xiao, J., Yao, A., Liu, Z., Li, Y., Ji, W., Chua, T.S.: Video as conditional graph hierarchy for multi-granular question answering. In: Proceedings of AAAI. pp. 2804–2812 (2022)
30. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of CVPR. pp. 1492–1500 (2017)
31. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of CVPR. pp. 5410–5419 (2017)
32. 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: Proceedings of ACM Multimedia. pp. 265–273 (2020)
33. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of CVPR. pp. 10685–10694 (2019)
34. Yao, Y., Chen, Q., Zhang, A., Ji, W., Liu, Z., Chua, T.S., Sun, M.: PEVL: Position-enhanced pre-training and prompt tuning for vision-language models. arXiv preprint arXiv:2205.11169 (2022)
35. Yao, Y., Zhang, A., Han, X., Li, M., Weber, C., Liu, Z., Wermter, S., Sun, M.: Visual distant supervision for scene graph generation. In: Proceedings of ICCV. pp. 15816–15826 (2021)
36. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: CPT: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021)
37. Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q.: CogTree: Cognition tree loss for unbiased scene graph generation pp. 1274–1280 (2021)
38. Zareian, A., Karaman, S., Chang, S.F.: Bridging knowledge graphs to generate scene graphs. In: Proceedings of ECCV. pp. 606–623 (2020)
39. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural Motifs: Scene graph parsing with global context. In: Proceedings of CVPR. pp. 5831–5840 (2018)
40. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: Proceedings of the AAAI. pp. 9185–9194 (2019)
41. Zhuang, B., Wu, Q., Shen, C., Reid, I., van den Hengel, A.: HCVRD: a benchmark for large-scale human-centered visual relationship detection. In: Proceedings of AAAI. pp. 7631–7638 (2018)