Effective Scene Graph Generation by Statistical Relation Distillation

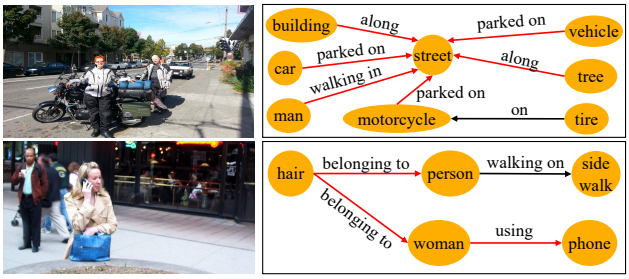
Son Nguyen[1] Hong Yang[1,2] Basura Fernando[1,2]
[1]Institute of High-Performance Computing, Agency for Science, Technology and Research, Singapore
[2]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research, Singapore
{nguyen_thanh_son, yang_hong,fernando_basura}@ihpc.a-star.edu.sg

Abstract

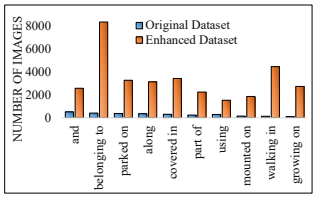
Annotating scene graphs for images is a time-consuming task, resulting in many instances of missing relations within existing datasets. In this paper, we introduce the Statistical Relation Distillation (SRD) method to enhance scene graph datasets. SRD leverages human-annotated relations alongside object-to-object and predicate-to-predicate similarities to reinforce the existence likelihood of scene graph relations. Moreover, SRD can augment relational frequency using relations of non-selected object and predicate categories that are usually omitted by scene graph generation (SGG) task. The output from SRD derives the prior probability which is combined with model-predicted probabilities to annotate missing relations in training images and subsequently re-train SGG models on the augmented dataset. We evaluate our proposed method on Visual Genome and GQA-200 datasets. Experimental results show that training on the augmented dataset enhances the performance of prominent scene-graph generation models. The implementation code at https://github.com/LUNAProject22/SRD.

1. Introduction

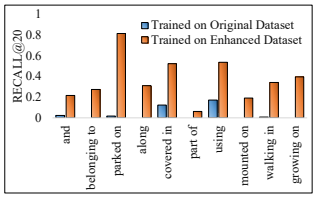
Scene graph generation (SGG) is an important computer vision problem in visual scene understanding [3, 5, 14, 19, 23, 25, 33, 54, 57]. A SGG model learns to predict objects in a scene and relations between pairs of objects [50]. The quality of the annotated scene graphs in the training set is an important factor affecting the generalization performance of SGG models [26, 32, 51]. However, recent works show that annotating all-important relations in an image are not an easy task, and the most common benchmarks such as Visual Genome (VG) [23, 26] have many missing relations in the training set [51]. For example, Figure 1a shows two images and their human-annotated relations (black arrows). Although there are many relations available, there is only one annotated relation of <*tire, on, motorcycle*> for the first image. Similarly, the second image only has one annotated relation of <*person, walking on, sidewalk*>. Moreover, to



(a) Example of enhanced scene graphs with newly-added relations (red arrows).



(b) Size changes of training samples for some predicates. Our method adds more relations to images, focusing on the infrequent predicates.



(c) Showing Recall@20 of the PredCls task for the selected predicates. Training on the enhanced dataset significantly improve the performance.

Figure 1. Illustration of our method. (a) We augment training image scene graphs by discovering new relations. (b) This results in an enhanced dataset having more training relation, specifically focused on infrequent predicates. (c) This leads to the improvement of the predicate classification task (PredCls for short) on rare predicates. This figure uses the results from Transformer as the base model, evaluated on VG dataset.

simplify the task, most methods work on a subset of selected object and predicate categories. For example, the SGG task using VG dataset commonly focuses on 150 object and 50 predicate categories although the dataset contains thousands of object and predicate categories. Consequently, a large amount of annotations which could potentially be used for training are omitted.

In this paper, we propose *Statistical Relation Distillation* (SRD) method to reinforce the existence likelihood of relations by accumulating frequency information from similar relations including the excluded object and predicate categories. In other words, SRD allows relations of un-selected object and predicate categories to contribute to the augmentation process. Specifically, SRD transfers the relational

information (i.e., frequency) based on object-to-object and predicate-to-predicate similarities. The assumption is that similar objects can have similar relations and similar predicates can form relations with similar object pairs. For example, the relation <*house, along, street*> can be used to infer that *building* can also be *along street* because *building* and *house* are similar objects. Similarly, <*sign, attached on, pole*> can be used to infer the relation of <*sign, mounted on, pole*> since *attached on* and *mounted on* are similar predicates. Subsequently, we employ *deductive reasoning* to add missing relations to enhance the scene graph annotations of training images. In particular, we compute statistical priors, namely Distilled Statistical Prior (DSP), based on the reinforced relational information obtained by SRD. DSP is combined with the self-labeling technique [51] to evaluate relation candidates and choose the most probable relations to add to the image. Returning to Figure 1a, the newly-added relations, depicted by red arrows, offer more comprehensive relational information among the objects in the images. Figure 1b shows the differences between the original dataset and the enhanced dataset regarding the number of images containing relations for some infrequent predicates. By adding more training samples for these predicates, SGG model can improve significantly for these predicates as shown in Figure 1c. Figure 2 illustrates our proposed method. On the top part of the figure, we first perform SRD on the original annotations to obtain distilled information which is then used to derive statistical priors (DSP) including prior pair probability and prior predicate probability. On the bottom part of the figure, we add new relations to an image based on the statistical priors and prediction of a pre-trained SGG model (self-labeling). As SRD utilizes all relational annotations, DSP encodes evidence from not only the relations of chosen object and predicate categories but also from those of non-selected ones. For example, predicate '*attached on*' is not selected, but its relational information is (partially) retained by the selected predicate '*mounted on*' because the two predicates are similar. In other words, the statistical prior of '*mounted on*' is enhanced by the presence of '*attached on*' annotations. DSP contains *conditioned pair probability* to answer questions such as "given a subject, how likely an object category can be the object of a relational triplet?", and *conditioned predicate probability* to answer the question of "what is the most likely predicate for a given pair of objects?". As statistical priors could vary for different types of images (due to the context and scene type), we derive *context-based DSP*. Specifically, we first cluster images into groups based on image context (e.g., animals, beach). Then, we apply SRD and obtain DSP for each cluster. During relation deduction, each image will use its corresponding cluster's DSP. Zhang et al. [51] proposes to add new relations (external transfer) and modify existing relations (internal transfer) based on trained SGG model's predictions. Similarly, we leverage a trained SGG model, but in addition, we take into account the enhanced prior distribution DSP. This allows us to take advantages of both the prior and conditional distributions to infer missing labels of triplets in the training set. Afterwards, we retrain the models with the enhanced scene graph dataset which includes both human-annotated and newly added annotations. Experimental results show that our method helps obtain better annotations leading to significant improvements over the self-labeling external transfer model of [51] using SGG models such as Motif [50], VCTree [44], and Transformer [43] for predicate classification, scene graph classification and scene graph detection tasks. Although such statistical priors have been used in prior scene graph generation models to aid conditional classifier predictions [43, 44], we are the first to use these priors to discover new relations for subject-object pairs that do not have predicate annotations. Furthermore, to the best of our knowledge, we are the first to introduce the concept of statistical distillation which transfers the occurrence statistics of a bigger set to the statistics of a smaller subset of entities. Then we use these enhanced statistics to discover relations for the missing pairs.

Our contributions are as follows. First, we introduce SRD method to transfer relational information among similar relations. SRD allows preserving relational information of object and predicate categories that are commonly omitted in the standard SGG setting. SRD strengthens statistical prior of triplets, creating more robust statistical priors (DSP). We further improve DSP by incorporating scene context information using foundational model features of images (i.e., CLIP [38]). Secondly, we use DSP to generate new relations for subject-object pairs lacking predicate labels in the training set. The integration of predictions from the trained SGG model (conditional probability) with DSP contributes to the improvement of the training set, subsequently leading to significant performance enhancements when being evaluated on VG and GQA-200.

2. Related Work

The scene graph, introduced for image retrieval in 2015 [18], has since been applied to various tasks such as visual question answering [37, 45], image captioning [47, 55, 56], and image generation [17, 28]. Numerous SGG approaches have emerged, including TransE-based SGG [13, 53], CNN-based SGG [29, 48], RNN/LSTM-based SGG [44, 46, 50], GNN-based SGG [30, 36], Transformer-based SGG [6, 7], and methods addressing long-tail issues [8]. Our work is primarily related to the methods focusing on unbiased scene graph generation using data enhancement techniques and dataset curation [16, 22, 34]. Li *et al*. [26] argue that there are noisy labels in annotated predicates for both positive and negative samples. They propose to detect noisy samples

and then correct them by treating the scene graph generation problem as a noisy label learning problem. Lyu *et al*. [32] focus on hard-to-distinguish samples (predicate triplets) and proposes a fine-grained categorization loss. Li *et al*. [27] use similarity between two predicates quantified using probability predictions of a biased model and uses this measure to develop a new reweighting loss function to eliminate the biased predictions. In contrast, our computation of similarity (among images, objects, and predicates) relies on embeddings generated from external foundational models such as CLIP [38] and BERT [9]. Specifically, we make use of image-visual similarity (CLIP) to cluster image contexts, and linguistic (BERT) to obtain object and predicate similarity. More recently, Yu *et al*. [49] propose to make use of large language models to generate open-world scene graph generation. To solve this task, they use a visually prefixed prompt and a language prompt learning with a hybrid template. Authors also perform predicate clustering based on the GloVe [35] vectors to obtain highly correlated predicates. A similar approach is also presented in [54]. Min *et al*. [33] propose to address the issue of subject-object pair distributions. Our method also addresses both subject-object imbalances as well as predicate imbalance issues in a data and semantic-driven manner.

Zhang *et al*. [51] developed a method to transfer triplet information from uninformative ones to informative ones such as a triplet with "on" predicate is replaced with a more informative "sitting on". Secondly, they also propose a method to infer missing triplets (those ones that are valid, yet not annotated). The authors use all non-annotated overlapping object pairs and pseudo-labeled the missing relation with a trained model. Even if this strategy could label some of the missing labels correctly, it has to rely on the accuracy of the predictor. We use both the predictor and the prior to obtain more accurate relations. In the same spirit, Goel *et al*. [14] propose a method to replace explicit relations (on, under) with implicit relations (riding on) using a model-based relabelling approach. Lin *et al*. [31] use a graph convolutional approach to obtain refined feature representations for each object, subject, and relation features. Li *et al*. [25] propose using feature augmentation to enrich the feature diversity of original relation triplets by replacing and mixing up their intrinsic or extrinsic features from other samples. Similarly, a semantic embedding augmentation strategy to eradicate bias is used in [1]. Our strategy is to acquire information from all samples and use the infused statistics to improve on the predicate and subject-object biases by self-labelling missing rare predicates. Good priors have significantly improved many other problems in Computer Vision and Machine Learning [11, 12, 42]. We discover prior for the relational data using SRD.

Furthermore, there are more recent advancements in scene graph detection extending DETR [2] with specialized query learning [20]. Large language model-based weakly supervised scene graph generation has been studied to address labelling issues [21] and Few-shot scene graph generation has been studied in [4]. The diverse visual appearance within the same predicate and the lack of patterns in tail predicates have been addressed in [24]. External domain knowledge has been used to improve scene graphs in [52]. Scene graphs were extended to represent situations in [41].

3. Methodology

The overview of our proposed method is shown in Figure 2. Many object pairs in the existing dataset are valid, yet not annotated with the correct relation triplet categories– see also the Figure 5. Therefore, our objective is to obtain a predicate label for each pair of objects that have object class annotation but no predicate label in the training set of a scene graph dataset. Specifically, we make use of a trained predicate classification model ($\phi()$) and our distilled statistical priors as explained in Section 3.3 to infer missing predicate labels. Given a pair of object instances annotated in an image having subject and object classes o_i and o_j that is not annotated with a predicate label, we infer a new probability $P(t|I)$ for the triplet $t =< o_i, r, o_j >$ where r is the predicate. We use pretrained predicate classification model $\phi()$ to obtain $P_\phi(r|I, o_i, o_j)$ and multiply it with the prior triplet probability ($P_\mathcal{G}(t)$) as follows:

$$P(t|I) = P_\mathcal{G}(t) \times P_\phi(r|I, o_i, o_j) \tag{1}$$

The prior triplet probability ($P_\mathcal{G}(t)$) is an enhanced prior obtained using the statistical relation distillation. In contrast to traditional statistical priors, the new prior $P_\mathcal{G}(t)$ is estimated using the scene-summary graph of the entire dataset ($\mathcal{G}_\mathcal{D}$) and relation-to-relation and object-to-object semantic similarities. Using the output from Equation (1), we predict the missing relation (selecting only the maximum value) for each pair of objects that lack a predicate label in the training set images. The enhanced dataset is used to retrain SGG models. The technique of using a pretrained model to enhance the training set is referred to as *self-labeling* in the literature and has been extensively employed in previous studies [51]. In the following sections, we will explain the scene summary graphs and the process of statistical relation distillation in Section 3.2. The method for obtaining prior triplet probability ($P_\mathcal{G}(t)$) is detailed in Section 3.3, and the approach for adding new relations into an image is discussed in Section 3.5.

3.1. Scene-Summary Graph (SSG)

A dataset of scene graphs $\mathcal{D} = \{I, G, \mathcal{O}, \mathcal{R}\}$ contains q images $I = \{I_1, \cdots, I_q\}$ and the corresponding scene graphs $G = \{G_1, \cdots, G_q\}$. Each $G_i = O_i \times R_i \times O_i$, where $O_i \subseteq \mathcal{O}$ is a set of objects appearing in image I_i and

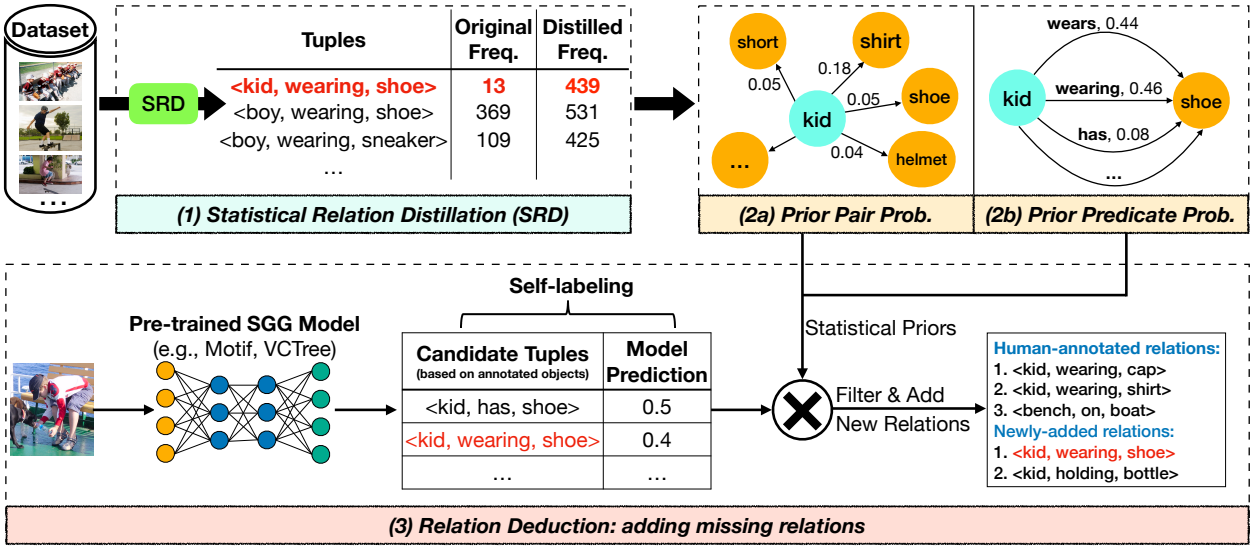


Figure 2. Illustration of the method. (1) *Statistical Relation Distillation* (SRD) enhances the original frequency of relation-triplets (tuples) by accumulating frequency from similar relations. (2) The new frequencies are used to compute *pair probability* (2a) and *predicate probability* (2b). (3) priors are combined with model-generated probabilities to identify and add missing relations. 'Freq.' and 'Prob.' are the short forms of 'frequency' and 'probability'.

$R_i \subseteq \mathcal{R}$ is a set of predicates (relations) between the objects. $\mathcal{O} = \{o_1, \cdots, o_M\}$ and $\mathcal{R} = \{r_1, \cdots, r_L\}$ are the set of all the objects and relations appeared in G, respectively. A scene graph consists of relationship triplets in the form of $<$*subject*, *predicate*, *object*$>$ where *subject* and *object* are objects from $\mathcal{O}$ and *predicate* is from $\mathcal{R}$. A scene graph contains relations between *object instances*. For example, in an image of two cats sitting on a table. The corresponding scene graph contains object instances (e.g., table, cat1, and cat2) and the relations (described by predicates) between them. We define a *scene summary graph* (SSG) to store the frequency of relationship triplets in scene graph at the *object-category* level. In the above-mentioned example, the corresponding image SSG contains $<$*cat*, *on*, *table*$>$ with the frequency value of two because it has two relations $<$cat1, on, table1$>$ and $<$cat2, on, table1$>$. The SSG for an image I_i is denoted as $\mathcal{G}_i$ and is represented as a tensor of size $L \times M \times M$ for *predicate*, *subject*, *object*, respectively. where $L = |\mathcal{R}|$ is the number of all the predicates and $M = |\mathcal{O}|$ is the number of all objects. The value at index (j, k, l) is the frequency of triplet $<o_k, r_j, o_l>$ obtained from G_i. The SSG of a dataset is obtained by aggregating SSG of all the images in the dataset as follows:

$$\mathcal{G}_\mathcal{D} = \sum_{i=1}^{q} \mathcal{G}_i \tag{2}$$

where $\mathcal{G}_\mathcal{D}$, represented as a tensor of size $L \times M \times M$ is the SSG of dataset $\mathcal{D}$. The $\mathcal{G}_\mathcal{D}$ contains all the relations between objects and the corresponding frequency. The frequency information in the $\mathcal{G}_\mathcal{D}$ along with additional similarity information is used to deduce new relations and new statistical prior.

3.2. Statistical Relation Distillation (SRD)

Similar objects might have similar relations and a relation can be described using semantically similar predicates. Inspired by these observations, we propose **S**tatistical **R**elation **D**istillation (SRD) method to *distill* relational information among triplets having similar objects and/or similar predicates. We base on similarity matching to obtain new frequencies for all possible relations in the dataset given the SSG of the dataset $\mathcal{G}_\mathcal{D}$. Specifically, the frequency of a relationship triplet is passed to other triplets of similar objects and/or predicates, and in turn, receiving from those as well. Next, we explain the SRD in detail.

SRD takes the original SSG ($\mathcal{G}_\mathcal{D}$, obtained by Equation 2), object similarities, and predicate similarities as input, and computes the distilled SSG, denoted as $\mathcal{G}'_\mathcal{D}$. The distillation can take two paths: 1) *self-distillation* and 2) *subset-distillation*. In self-distillation, $\mathcal{G}_\mathcal{D}$ and $\mathcal{G}'_\mathcal{D}$ have the same sets of objects and predicates. Whereas, in subset-distillation, the objects and predicates in $\mathcal{G}'_\mathcal{D}$ are subsets of those in $\mathcal{G}_\mathcal{D}$. In other words, we distill the information from bigger sets of objects and predicates to smaller sets. For generalizability, we represent $\mathcal{G}'_\mathcal{D}$ as a tensor of $L' \times M' \times M'$ where $L' = |\mathcal{R}'|$ and $M' = |\mathcal{O}'|$. Here, $\mathcal{R}' \subseteq \mathcal{R}$ and $\mathcal{O}' \subseteq \mathcal{O}$ are the sets of relations and objects encoded in $\mathcal{G}'_\mathcal{D}$, respectively. When $\mathcal{R}' = \mathcal{R}$ and $\mathcal{O}' = \mathcal{O}$, it is self-distillation. When $\mathcal{R}' \subset \mathcal{R}$ and $\mathcal{O}' \subset \mathcal{O}$, it is

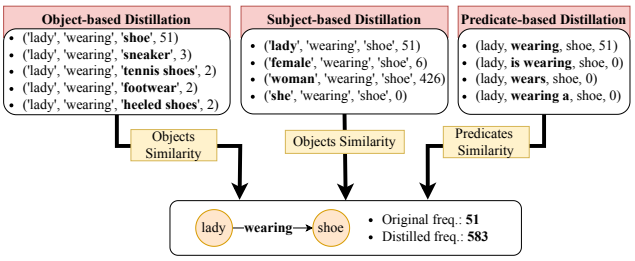


Figure 3. Example of SRD that enhances the frequency (freq.) of triplet <lady, wearing, shoe> from 51 to 583. The contributing triplets (similar relations) are shown with their original frequency.

subset-distillation. We define a similarity function $S(\cdot,\cdot)$ which returns a similarity matrix between the terms (object-object or predicate-predicate). The similarity between two terms is computed as the *cosine similarity* of their embedding generated by pretrained large language model (e.g., the cosine similarity between the embeddings of *table* and *desk*). The main idea of the distillation process is to estimate the new (weighted) frequency of triplets by propagating the frequencies from similar triplets. The weights are determined by how similar they are. SRD consists of three main steps, *object-based distillation*, *subject-based distillation*, and *predicate-based distillation* (Fig. 3). We first perform the object-based distillation by computing the matrix multiplication between the original SSG and the object similarity matrix $S(\mathcal{O},\mathcal{O}')$.

$$T_1 = \mathcal{G}_{\mathcal{D}} \times S(\mathcal{O},\mathcal{O}') \tag{3}$$

where T_1 is a tensor of size $L \times M \times M'$ (for predicate, subject, obj*[1]) and similarity matrix $S(\mathcal{O},\mathcal{O}')$[2] has the size of $M \times M'$. Next we perform the subject-based distillation.

$$T_2 = S(\mathcal{O},\mathcal{O}')^{\mathrm{T}} \times \text{permute}(T_1,(1,2,0)) \tag{4}$$

where T_2 is a tensor of size $M' \times M' \times L$ (for subject*, object*, predicate), permute$(\cdot,\cdot)$ is a function which returns a view of the original tensor input (the first parameter) with its dimensions permuted with the desired ordering of dimensions (the second parameter). Lastly, we perform the predicate-based distillation which is based on the similarity between predicates $S(\mathcal{R},\mathcal{R}')$.

$$T_3 = T_2 \times S(\mathcal{R},\mathcal{R}') \tag{5}$$

where T_3 is a tensor of size $M' \times M' \times L'$ (for subject*, object*, predicate*) and the similarity matrix $S(\mathcal{R},\mathcal{R}')$ has the size of $L \times L'$. Finally, we permute T_3 to obtain the distilled SSG.

$$\mathcal{G}'_{\mathcal{D}} = \text{permute}(T_3,(2,0,1)) \tag{6}$$

[1] "*" is used to indicate the dimensions that has done the distillation

[2] The similarity matrix returned by the similarity function $S(\cdot,\cdot)$.

where $\mathcal{G}'_{\mathcal{D}}$ has the size of $L' \times M' \times M'$ (for predicate*, subject*, object*). $\mathcal{G}'_{\mathcal{D}}$ encodes not only the original frequency information for each triple, but also the distilled information aggregated from triples of similar objects and predicates.

3.3. Obtaining Statistical Priors from SSG

Given a SSG, we derive the statistical priors including the *conditioned pair probability* and *conditioned predicate probability*. The conditioned pair probability can be used to answer for the question of "given a subject, how likely an object category can be the object of a triplet?", and the conditioned predicate probability is used to answer for the question of "what is the most likely predicate that can be used to describe the relation for a given pair of subject and object?". The conditioned pair probability for a given subject o_i denoted by $P_{\mathcal{G}}(\mathbf{o}=o_j|\mathbf{s}=o_i)$ (where o_i is the subject and o_j is the object) is computed as follows:

$$P_{\mathcal{G}}(\mathbf{o}=o_j|\mathbf{s}=o_i) = \frac{\sum_{k=0}^{L}\mathcal{G}(r_k,o_i,o_j)}{\sum_{l=0}^{M}\sum_{m=0}^{L}\mathcal{G}(r_m,o_i,o_l)} \tag{7}$$

where $\mathcal{G}(r_k,o_i,o_j) = \mathcal{G}[k,i,j]$ (value at index $[k,i,j]$ of tensor $\mathcal{G}$) is the frequency of triplet $<o_i,r_k,o_j>$ in SSG. The probability of having r_k as the predicate for the pair $(\mathbf{s}=o_i,\mathbf{o}=o_j)$ is derived using the following equation.

$$P_{\mathcal{G}}(\mathbf{p}=r_k|\mathbf{s}=o_i,\mathbf{o}=o_j) = \frac{\mathcal{G}(r_k,o_i,o_j)}{\sum_{m=0}^{L}\mathcal{G}(r_m,o_i,o_j)} \tag{8}$$

We obtain DSP by applying these equations on the distilled SSG ($\mathcal{G}'_{\mathcal{D}}$). Prior triplet probability ($P_{\mathcal{G}}(t)$) is as follows:

$$P_{\mathcal{G}}(t) = P_{\mathcal{G}}(\mathbf{o}=o_j|\mathbf{s}=o_i)\cdot P_{\mathcal{G}}(\mathbf{p}=r|\mathbf{s}=o_i,\mathbf{o}=o_j) \cdot P_{\mathcal{G}}(\mathbf{s}=o_i) \tag{9}$$

where the probability of selecting an object as subject, $P_{\mathcal{G}}(\mathbf{s}=o_i)$, is uniform.

3.4. Context-based SRD

The likelihood of having a relation between a pair of objects depends on the context the two objects appear. For example, in the context of 'beach', a *man* is more likely to have a relation with a *surfboard* than with a *laptop* (see supplementary). Similarly, the exact predicate to describe the relation between a pair of objects could also vary in different contexts. For example, *man* is likely to *wear jacket* if the context is under snow, but *man* is more likely to *hold jacket* if the context is indoor. Therefore, our hypothesis is that relation distillation by deduction should be performed on the images belonging to the same context to avoid passing irrelevant information. Our objective is not to know the exact context of each image, but rather which images belong to the same context. Therefore, we group the images

into different clusters and consider the images belonging to the same cluster to have the same context. Specifically, we apply the K-Means algorithm to cluster the images into K clusters based on the CLIP (ViT-B-32) [38] model image features. To decide the number of clusters, we based on the Silhouette coefficient [40]. For each cluster, we compute the context-based SSG and apply SRD on the SSG to obtain context-based distilled SSG. Then, the context-based distilled SSG is used to derive the statistical priors (DSP) for the cluster using the method described in Section 3.3.

3.5. Relation Deduction: Adding Missing Relations

For each image in the training set, we identify and add relations for pairs of annotated objects that lack human annotations. When adding new relations, we filter out triplet candidates that do not have overlapping subject-object bounding boxes. Following the methodology of [51], we select only those triplets that appeared in human annotations, focusing on the top 50% infrequent predicates for candidate pairs. This strategy enables us to add more new rare triplets into the training set compared to the frequent ones. Additionally, we label a pair of object instances (without a predicate label) with a predicate for the subject and object categories only if that triplet appears at least 15 times in the training set. This approach helps us to add rare but high-quality triplets rather than those annotated by mistake (i.e., those that appear only a few times). This strategy prevents the introduction of noise or incorrect relations. For each valid candidate pair, we choose a valid predicate (i.e., belonging to the less frequent 25 predicates) based on the score computed as described in Equation 1.

4. Experiments

Datasets. We evaluate our proposed approach on two benchmark datasets: Visual Genome(VG)-150 [23] and GQA-200 [15]. For the VG dataset, we use the default splits of 57,723, 5,000, and 26,446 images for training, validation, and testing, respectively [43]. The dataset has 150 selected object classes and 50 selected predicate classes [49]. Similarly, for the GQA dataset, we adopt the default experimental procedures [10, 33, 49]. This entails employing 57,623 images for training, 5,000 for validation, and 8,209 for testing. Within this dataset, the number of selected object classes and predicate classes are 200 and 100, respectively. The base models are implemented based on the code from [43][3]. During SRD process, we leverage the entire set of object classes and predicate classes annotated within the datasets. Specifically, for VG, this includes 13,053 objects and 5,232 predicates, while for GQA-200, it includes 1,684 objects and 310 predicates. However, during the training and evaluation stages, only the respective selected objects and predicates are utilized.

[3] https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch

Method	Pred.Cls	SGCls	SGDet
	mR@20 /50 /100	mR@20 /50 /100	mR@20 /50 /100
Motif (Base)	12.1 / 15.7/ 17.4	7.2 / 8.7 / 9.3	5.1 / 6.5 / 7.8
–NICE [26]	– / 29.9 / 32.3	– / 16.6 / 17.9	– / 12.2 / 14.4
–IETrans [51]	30.2 / 35.8 / 39.1	18.2 / 21.5 / 22.8	12.0 / 15.5 / 18.0
–CFA [25]	– /35.7 /38.2	– / 17.0 / 18.4	– / 13.2 / 15.5
–CaCao [49]	30.9 / 37.1 / 38.9	**20.4 / 23.3 / 24.4**	12.6 / 17.1 / 20.0
–Ours	**31.9 / 37.9 / 40.5**	18.9 / 21.9 / 22.8	**13.5/ 17.9/ 20.6**
VCTree (Base)	12.4 / 15.4 / 16.6	6.3 / 7.5 / 8.0	4.9 / 6.6 / 7.7
–NICE [26]	– / 30.7 / 33.0	– / 19.9 /21.3	– / 11.9 / 14.1
–IETrans [51]	31.7 / 37.0 / 39.7	18.2 /19.9 / 21.8	9.8 / 12.0 / 14.9
–CFA [25]	– / 34.5 / 37.2	– / 19.1 / 20.8	– / 13.1 / 15.5
–CaCao [49]	33.1 / 37.5 / 38.9	**23.8 / 27.5 / 28.7**	11.8 / 16.4 / 19.1
–Ours	**33.4 / 39.0 / 41.1**	23.0 / 26.6 / 27.6	**12.8 / 16.7 / 19.6**
Transf. (Base)	14.8 / 19.2 / 20.5	8.9 / 11.6 / 12.6	5.6 / 7.7 / 9.0
–IETrans [51]	29.1 / 35.0 / 38.0	17.9 / 20.8 / 22.3	11.7 / 15.0 / 18.1
–CFA [25]	– / 30.1 /33.7	– / 15.7 / 17.2	– / 12.3 / 14.6
–CaCao [49]	**36.2 / 41.7 / 43.7**	**21.1/ 24.0 / 25.0**	13.5 / **18.3 / 22.1**
–Ours	34.0 / 39.6 / 41.7	20.1 / 23.0 / 23.9	**14.3 / 18.3** / 20.8

Table 1. Comparing to the state-of-the-art methods evaluated on VG dataset. We use context-based SRD. **Bold text** indicates best result and underlined text indicates the second-best result.

Experimental Setup. To be consistent with prior works [14, 32, 43, 51], we use mean recall at k (mR@k) for the evaluation. Our method is assessed with VCTree [44], Motif [50], and Transformer [43] models, employing a pre-trained Faster-RCNN with ResNeXt-101-FPN for feature extraction in predicate classification, scene graph classification and scene graph detection. We adhere to standard training protocols from prior methods [43] and use BERT [39] embeddings for object and predicate similarity computation. For image context clustering, we choose the number of clusters based on Silhouette coefficient, i.e., 25 clusters for VG and 20 clusters for GQA-200.

To mitigate the long-tailed distribution issue, we also employ the reweighting loss strategy, the same as the approach outlined in [51]. The reweighting loss technique addresses this issue by assigning distinct weights to different predicate classes during the training process. The weight of relation r_k is determined as $w_k = \frac{\text{median of all relation frequencies}}{\text{frequency of } r_k}$. Subsequently, these computed weights are incorporated into our classification loss (*i.e.*, the class-weighted cross-entropy loss) when re-training the model with the enhanced dataset. This encourages the model to prioritize less frequent classes that potentially improve the overall performance. Moreover, during SRD, triplet frequencies are weighted by the inverse relation frequency of each predicate category (inspired by TFIDF –Term Frequency Inverse Document Frequency) to reduce the impact of frequency bias.

4.1. Comparison to State-of-the-Art Methods

We compare with prior scene graph dataset enhancement methods. In Table 1, our model, being model-agnostic,

Method	P.Cls		S.Cls		S.Det	
	@50	@100	@50	@100	@50	@100
Motif (Baseline)	16.8	17.9	8.2	8.6	6.4	7.7
–GCL [10]	36.7	38.1	17.3	18.1	16.8	18.8
–CaCao [49] (ICCV23)	**37.5**	**40.5**	**19.6**	**21.9**	**17.8**	**19.6**
–EICR [33] (ICCV23)	36.3	38.0	17.2	18.2	16.0	18.0
–Ours	35.7	37.0	18.1	19.0	16.2	18.1
VCTree (Baseline)	16.6	17.4	7.9	8.3	6.5	7.4
–GCL [10]	35.4	36.7	17.3	18.0	15.6	17.8
–EICR [33] (ICCV23)	**35.9**	**37.4**	17.8	18.6	14.7	16.3
–Ours	**35.9**	37.0	**18.9**	**19.4**	**16.9**	**18.9**
Transformer (Baseline)	17.5	18.7	8.5	9.0	6.6	7.8
–GCL [10]	35.6	36.7	17.8	18.3	16.6	18.1
–CaCao [49] (ICCV23)	34.8	36.9	19.3	20.1	**18.8**	19.1
–Ours	**37.6**	**38.8**	**20.2**	**20.7**	18.7	**20.6**

Table 2. Comparing to the state-of-the-art methods evaluated on GQA dataset [15]. We use context-based SRD. **Bold text** indicates best result and underlined text indicates the second-best result.

outperforms established dataset enhancement methods like IETrans [51] and NICE [26], showcasing good performance. This success is attributed to our method's ability to leverage all relation and object categories (13,053 object classes and 5,232 predicates) in the Visual Genome dataset, along with a statistical and semantic-centric bias reduction mechanism. Using BERT for semantic similarities, CLIP for semantic clusters, and statistical distillation for priors, our approach effectively addresses long-tailed issues in predicate classification, particularly improving rare predicate classes (see Section 4.2). The method remains competitive with foundational model-based data enhancement methods such as CaCao [49] and excels with Motif and VCTree, demonstrating significant enhancements in predicate classification and scene graph detection. The integration of our distilled statistical prior proves beneficial across different methods, emphasizing the practical utility and seamless integration of our contribution into existing frameworks for competitive outcomes. In Table 2, we show the performance of our method on a more challenging GQA-200 dataset. Compared to the VG dataset, the number of object classes is reduced from 13,053 to 1684 and therefore our distillation process only transfers statistics from fewer total possible relations. Nevertheless, even in the GQA-200 dataset, our method performs comparably with other methods (10 out of 18 times our method obtains the best results) in the literature demonstrating the robustness of the main idea of this paper even under a more challenging setup.

4.2. Ablations and Analysis

In this section, we ablate our model using Transformer and VCTree models. **SRD's components.** As presented in Section 3, our model comprises key components: the statistical relation distillation module (SRD), and context-based SRD. We assess their impact in Section 4.1. The

	Transformer	VCTree
Method	mR@20 / 50 / 100	mR@20 / 50 / 100
SL	24.2 / 30.5 / 33.2	23.7 / 33.2 / 37.8
SRD	32.9 / 39.0 / 41.0	32.9 / 38.6 / 40.5
CSRD	**34.0 / 39.6 / 41.7**	**33.4 / 39.0 / 41.1**

Table 3. Evaluating the impact of different components of our method using predicate classification task using VG dataset. SL and CSRD stand for Self-labelling and context-based SRD.

	Transformer	VCTree
	mR@20 / 50 / 100	mR@20 / 50 / 100
Selected only	32.7 / 38.5 / 40.6	31.3 / 37.5 / 40.4
All	**34.0 / 39.6 / 41.7**	**33.4 / 39.0 / 41.1**

Table 4. Evaluating SRD when using different sets of object and predicate categories for relation distillation.

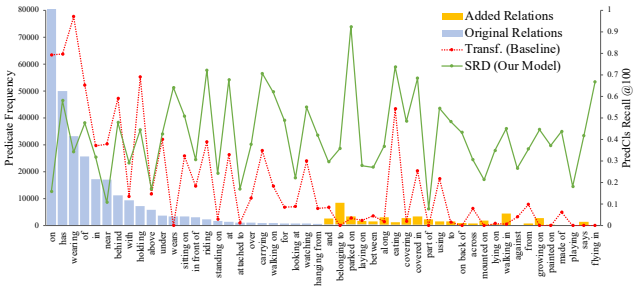


Figure 4. Comparison of original and augmented data shows increased number of rare predicate relations and per class recall@100 on PredCls.

first baseline, termed *Self-labelling*, solely relies on the pretrained SGG model $\phi()$ for conditional probability, neglecting the prior from SRD (Equation (1)). We use this model to label subject-object pairs that do not have predicate annotations similar to ours. From the results, we see that SRD consistently outperforms the self-labeling model across Transformer and VCTree. The inclusion of context improves the results in all cases. The biggest improvement comes from the inclusion of SRD.

SRD on Different Set of Object and Predicate Categories. Here we ablate the impact of using additional triplets from the full VG dataset. We evaluate the performance of the models when we apply SRD on the selected objects (150 categories) and predicates (50 categories) only. We compare the results in Table 4. The results show that when applying SRD on all the VG triplets, the performance of both Transformer and VCTree are better compared to when using only relational triplets of selected categories only. This demonstrates the benefit of using SRD as it allows utilizing all annotations.

The Impact of SRD on Rare and Frequent Predicates. Using our method, we added 42029, 43759, and 41763

Method	Head (16)	Body (17)	Tail (17)
Motif (Base) [8]	42.3	9.8	0.6
–DT2-ACBS [8]	35.1 (-7.2)	45.2 (+35.4)	38.6 (+38.0)
Motif (Base)	40.8	10.2	3.0
–Ours	37.4 (-3.4)	**47.1** (+36.9)	36.7 (+33.7)
VCTree (Base)	40.8	10.2	3.0
–Ours	38.9 (-1.9)	46.0 (+35.8)	38.5 (+35.5)
Transf. (Base)	**42.8**	13.7	4.6
–Ours	38.8 (-4.0)	46.7 (+33.0)	**39.3** (+34.7)

Table 5. Mean Recall @100 on predicate classification (PredCls) for head, body, and tail predicate classes.

new relations to the training set of the VG dataset when using Motif, VCTree, and Transformer, respectively. Among 57,723 images in the training set, we added new relations to about 22,000 images. For GQA-200, we added 23496, 26927, and 24111 new relations to the training set when using Motif, VCTree, and Transformer, respectively. Among 57,623 images in the training set, about 16,000 images were added new relations. We now analyze the impact of our method on both the infrequent and frequent predicates and with the showcase of Transformer on both datasets. The other two methods present similar trends. Figure 4 shows the number of relations containing each predicate for the original data ("Original Relations") and the number of newly added relations ("Added Relations") for the enhanced dataset. It also compares per-class recall at 100 between the Transformer baseline and our method for VG. After SRD, our method can add more instances of rare predicates, increasing their frequency for both VG and GQA-200 (see supplementary). Notably, the frequencies of rare predicates such as "belonging to" in VG and "growing on" in GQA-200, are amplified by an order of magnitude, from 560 to 8,490 and 175 to 2,998, respectively. This demonstrate the effectiveness of SRD in addressing the long-tail issue in the SGG task. The addition of infrequent predicates significantly enhances the corresponding recall, as illustrated by the solid green line in Figure 4.

While a decline in recall was noted for frequent predicates like "on" and "wearing," augmenting the dataset with new samples for the 25 least frequent predicates improved performance across 42 out of 50 predicates. Following [8], we compute the mean recall for head, body, and tail classes (Table 5). Our results show a similar trend with reduced head-class performance (-3.4%, -1.9%, -4% for Motif, VCTree, Transformer) but significant gains for body (+36.9%, +35.8%, +33.0%) and tail classes (+33.7%, +35.5%, +34.7%). Notably, our method surpasses [8] across all class groups. Examples of relations added by our method for VG using Transformer as the base model are shown in Figure 5. Many important missing relations are

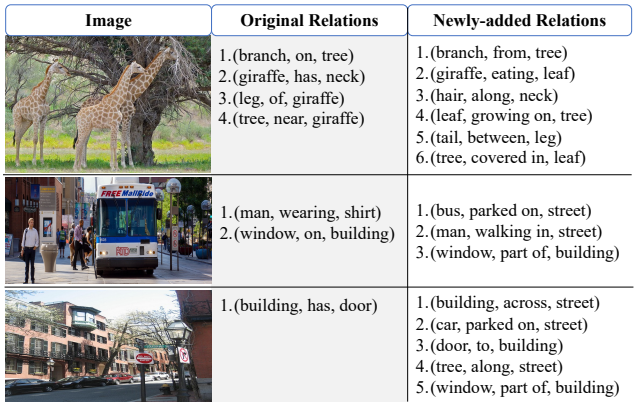


Figure 5. Example of newly-added relations using our method with Transformer on Visual Genome dataset.

added to the images. More examples for VG and GQA-200 can be found in the supplementary.

User Study to Evaluate Newly Added Relations. We qualitatively evaluated newly added relations by sampling 300 triples generated by the transformer model, creating 300 questions. Fourteen participants assessed these relations, marked in both image and text formats, answering 60 unique questions each, resulting in 880 responses (2-3 responses per question). Results show 59.3% (178) were unanimously valid, 15% (45) were unanimously invalid, and the rest had mixed votes. Excluding 16 ambiguous cases, the accuracy based on majority votes is 76.1%, validating the added relations and enhancing model performance for infrequent predicates in SGG tasks.

5. Conclusions

We introduced SRD to enhance the statistical prior of SGG datasets by distilling information from similar relational triplets. The enhanced prior probability is used along with a pre-trained SGG model to annotate missing triplets to augment the training set. The enhanced dataset is used to retrain SGG models, demonstrating significant improvements. Experiments with Motif, VCTree, and Transformer models show state-of-the-art performance in predicate classification and scene graph detection on VG and across all three tasks with multiple metrics on GQA-200. Notably, our method substantially improves the performance of tail predicate classes. The distilled relational statistics are interpretable, providing insights into the behavior and rationale behind statistical distillation.

Acknowledgment. This research/project is supported by the National Research Foundation, Singapore, under its NRF Fellowship (Award NRF-NRFF14-2022-0001) and ASTAR CRF award to B.F.

References

[1] Bashirul Azam Biswas and Qiang Ji. Probabilistic debiasing of scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10429–10438, June 2023. 3

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[3] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1–26, 2021. 1

[4] Vincent S Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2580–2590, 2019. 3

[5] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1581–1590, 2021. 1

[6] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. Spatial-temporal transformer for dynamic scene graph generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16372–16382, 2021. 2

[7] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[8] Alakh Desai, Tz-Ying Wu, Subarna Tripathi, and Nuno Vasconcelos. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15404–15413, 2021. 2, 8

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[10] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19427–19436, 2022. 6, 7

[11] Tianshi Gao, Michael Stark, and Daphne Koller. What makes a good detector?–structured priors for learning from few examples. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 354–367. Springer, 2012. 3

[12] Yansong Gao, Rahul Ramesh, and Pratik Chaudhari. Deep reference priors: What is the best way to pretrain a model? In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7036–7051. PMLR, 17–23 Jul 2022. 3

[13] 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)*, pages 1840–1844. IEEE, 2019. 2

[14] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022. 1, 3, 6

[15] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 6, 7

[16] Bowen Jiang, Zhijun Zhuang, and Camillo Jose Taylor. Enhancing scene graph generation with hierarchical relationships and commonsense knowledge. *arXiv preprint arXiv:2311.12889*, 2023. 2

[17] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 2

[18] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 2

[19] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. Devil's on the edges: Selective quad attention for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18664–18674, June 2023. 1

[20] Jongha Kim, Jihwan Park, Jinyoung Park, Jinyoung Kim, Sehyung Kim, and Hyunwoo J Kim. Groupwise query specialization and quality-aware multi-assignment for transformer-based visual relationship detection. In *CVPR*, 2024. 3

[21] Kibum Kim, Kanghoon Yoon, Jaehyeong Jeon, Yeonjun In, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Llm4sgg: Large language models for weakly supervised scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28306–28316, June 2024. 3

[22] Boris Knyazev, Harm de Vries, Cătălina Cangea, Graham W Taylor, Aaron Courville, and Eugene Belilovsky. Generative compositional augmentations for scene graph prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15827–15837, 2021. 2

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense

image annotations. *International journal of computer vision*, 123:32–73, 2017. 1, 6

[24] Jiankai Li, Yunhong Wang, Xiefan Guo, Ruijie Yang, and Weixin Li. Leveraging predicate and triplet learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28369–28379, June 2024. 3

[25] Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21685–21695, 2023. 1, 3, 6

[26] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. The devil is in the labels: Noisy label correction for robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18869–18878, 2022. 1, 2, 6, 7

[27] Wei Li, Haiwei Zhang, Qijie Bai, Guoqing Zhao, Ning Jiang, and Xiaojie Yuan. Ppdl: Predicate probability distribution based loss for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2022. 3

[28] Yikang Li, Tao Ma, Yeqi Bai, Nan Duan, Sining Wei, and Xiaogang Wang. Pastegan: A semi-parametric method to generate image from scene graph. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[29] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1347–1356, 2017. 2

[30] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018. 2

[31] Bingqian Lin, Yi Zhu, and Xiaodan Liang. Atom correlation based graph propagation for scene graph generation. *Pattern Recognition*, 122:108300, 2022. 3

[32] Xinyu Lyu, Lianli Gao, Yuyu Guo, Zhou Zhao, Hao Huang, Heng Tao Shen, and Jingkuan Song. Fine-grained predicates learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19467–19475, 2022. 1, 3, 6

[33] Yukuan Min, Aming Wu, and Cheng Deng. Environment-invariant curriculum relation learning for fine-grained scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13296–13307, October 2023. 1, 3, 6, 7

[34] Maëlic Neau, Paulo E Santos, Anne-Gwenn Bosser, and Cédric Buche. Fine-grained is too coarse: A novel data-centric approach for efficient scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11–20, 2023. 2

[35] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 3

[36] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3957–3966, 2019. 2

[37] Tianwen Qian, Jingjing Chen, Shaoxiang Chen, Bo Wu, and Yu-Gang Jiang. Scene graph refinement network for visual question answering. *IEEE Transactions on Multimedia*, 2022. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 6

[39] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019. 6

[40] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987. 6

[41] Chinthani Sugandhika, Chen Li, Deepu Rajan, and Basura Fernando. Situational scene graph for structured human-centric situation understanding. *arXiv preprint arXiv:2410.22829*, 2024. 3

[42] Libin Sun, Sunghyun Cho, Jue Wang, and James Hays. Good image priors for non-blind deconvolution: generic vs. specific. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pages 231–246. Springer, 2014. 3

[43] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 2, 6

[44] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019. 2, 6

[45] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017. 2

[46] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *European conference on computer vision*, pages 222–239. Springer, 2020. 2

[47] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 2

[48] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–338, 2018. 2

[49] Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, and Yueting Zhuang. Visually-prompted language model for fine-grained scene graph generation in an open world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21560–21571, October 2023. 3, 6, 7

[50] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 1, 2, 6

[51] Ao Zhang, Yuan Yao, Qianyu Chen, Wei Ji, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. Fine-grained scene graph generation with data transfer. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 409–424. Springer, 2022. 1, 2, 3, 6, 7

[52] Ce Zhang, Simon Stepputtis, Joseph Campbell, Katia Sycara, and Yaqi Xie. Hiker-sgg: Hierarchical knowledge enhanced robust scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28233–28243, June 2024. 3

[53] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017. 2

[54] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2915–2924, June 2023. 1, 3

[55] Shanshan Zhao, Lixiang Li, and Haipeng Peng. Aligned visual semantic scene graph for image captioning. *Displays*, 74:102210, 2022. 2

[56] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 211–229. Springer, 2020. 2

[57] Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Mingtao Feng, Xia Zhao, Qiguang Miao, Syed Afaq Ali Shah, et al. Scene graph generation: A comprehensive survey. *arXiv preprint arXiv:2201.00443*, 2022. 1