Causal Property based Anti-Conflict Modeling with Hybrid Data Augmentation for Unbiased Scene Graph Generation

Ruonan Zhang[1,2] and Gaoyun An[1,2*]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China
{21120318,gyan}@bjtu.edu.cn

Abstract. Scene Graph Generation(SGG) aims to detect visual triplets of pairwise objects based on object detection. There are three key factors being explored to determine a scene graph: visual information, local and global context, and prior knowledge. However, conventional methods balancing losses among these factors lead to conflict, causing ambiguity, inaccuracy, and inconsistency. In this work, to apply evidence theory to scene graph generation, a novel plug-and-play Causal Property based Anti-conflict Modeling (CPAM) module is proposed, which models key factors by Dempster-Shafer evidence theory, and integrates quantitative information effectively. Compared with the existing methods, the proposed CPAM makes the training process interpretable, and also manages to cover more fine-grained relationships after inconsistencies reduction. Furthermore, we propose a Hybrid Data Augmentation (HDA) method, which facilitates data transfer as well as conventional debiasing methods to enhance the dataset. By combining CPAM with HDA, significant improvement has been achieved over the previous state-of-the-art methods. And extensive ablation studies have also been conducted to demonstrate the effectiveness of our method.

Keywords: Scene graph generation · D-S evidence theory · Data augmentation.

1 Introduction

Scene Graph Generation(SGG) is an important task in computer vision, which can bridge low-level visual tasks such as object detection [26] and high level visual tasks such as visual question answering [44], 3D scene synthesis [25], image caption [38], etc. Today's SGG [3, 30, 32, 42] is considered as a deterministic task, recognizing relationships between pairwise objects. Recent work has made steady progress on SGG and provides powerful models to encode both visual and linguistic context of the scene. As illustrated in Fig. 1, a scene graph is composed of visual triplets in the form of $\langle subject - predicate - object\rangle$.

[*] Corresponding author

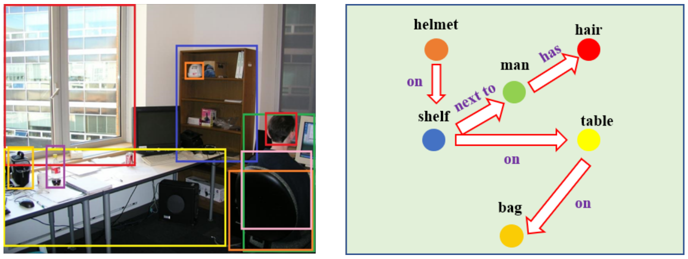


Fig. 1. An example of Visual Genome [14] dataset and its scene graph. Obviously, most of the predicates in this image are trivial, and the same is true for the dataset.

However, due to the long-tailed distribution of annotations, SGG is far from practical. In Visual Genome [14], there are 50 predicate classes, yet more than 100K samples are top 5 predicate classes [43]. BA-SGG [12] divided all predicates into two categories: informative and common. The frustrating fact is that annotators prefer common predicates, which are exactly the "head" predicates in the dataset. Therefore, we should not blame the model generates trivial and less informative predicates. To address this problem, we propose a Hybrid Data Augmentation (HDA) to deal with it.

Meanwhile, as Fig. 2 shown, counterfactual inference by causal graph [29] is also adopted to eliminate this highly-skewed long-tailed bias. A causal graph summarizes three key factors to determine a scene graph: visual appearance feature, context feature, and prior knowledge. Specifically, visual features are extracted from paired objects, context features are encoded by RNN or GNN for each object and prior knowledge denotes the predicate class distribution in the dataset after the categories of subject and object are identified. Then the predicate confidence scores are predicted by these features separately. Finally, these scores are fused to get a final result, usually by adding or gating.

Though models based on causal graphs outperform the conventional ones, we find an extra bias introduced by casual graphs. As Fig.3 shown, the output distributions of branches in casual graph may conflict each other. For the triplet of $\langle helmet - on - shelf \rangle$, if only context branch is considered, the final prediction is "on", while visual branch tends to wrongly predict "in". Meanwhile, background is also regarded as a highest probable prediction by prior knowledge. What's worse, after these three predicted distributions are fused together, the final result is misdirected to "in", although one branch performs correctly. To eliminate this extra bias, we propose a Causal Property Anti-conflict Modeling(CPAM) module, direct modeling these three branches via Dempster-Shafer(D-S) evidence theory [40], which is first applied in expert system [20] to handle uncertain information. In D-S evidence theory, evidences are mutually exclusive probability distributions which represent all possible answers to a question. By transferring this theory to causal graph based SGG task, the predicted confidence scores of these three branches can be explicitly modeled as evidences. Then the fused scores can be computed by Dempster combination rules, and thus this additional bias may be removed. By the way, CPAM is a plug-and-play module.

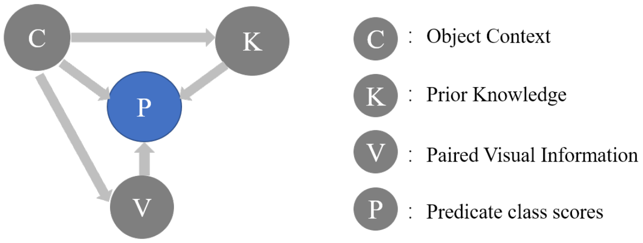


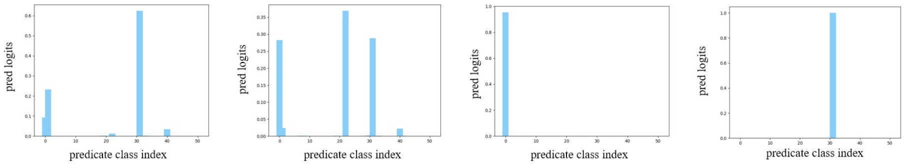
Fig. 2. The causal graph [29] in SGG. The final confidence scores of predicates come for three branches: context feature encoded by RNN or GNN from each proposal, visual appearance feature extracted by paired proposals, and prior knowledge distribution after the categories of subject and object are identified. Meanwhile, as input of the context branch, object visual features are also passed to the other two branches.

To sum up, the main contributions of our work contain: (1) We systematically reveal the long-tailed bias which limits SGG's overall performance and an extra bias introduced by multibranch prediction of a causal graph. (2) We propose a novel Causal Property Anti-conflict Modeling (CPAM) module, which applies Dempster-Shafer evidence theory and can serve as a plug-and-play module. (3) By combining data transfer and conventional debiasing method, we propose a Hybrid Data Augmentation (HDA) method to balance data distribution in Visual Genome. (4) Experimental results demonstrate the effectiveness of the proposed CPAM and HDA, which achieve state-of-the-art performance under existing evaluation metrics on Visual Genome and may improve the mean recall metric significantly.

2 Related Work

2.1 Scene Graph Generation

Scene Graph Generation has drawn widespread attention in computer vision community since Lu *et.al* [24] formalized SGG as a visual task to recognize relationships between objects. After the large-scale image semantic understanding dataset Visual Genome [14] is proposed, SGG has become a popular task in computer vision gradually. However, more and more researchers recognize the highly-skewed long-tailed distribution in the dataset which limits SGG performance seriously. To address this issue, TDE [29] introduced counterfactual causal analysis in the inference stage. BA-SGG [12] transferred common predicates to informative predicates by semantic adjustment. BGNN [16] proposed a bi-level data sampling strategy to sample instances of different entities and predicates. SHA-GCL [9] first grouped all the predicates and then employed a median-resampling method to make a balanced distribution. We, therefore, also propose an effective strategy HDA to reduce the impact of long-tailed bias and better results are shown by experiments.



(a) context branch (b) visual branch (c) frequency branch (d) CPAM

Fig. 3. This is an example of $\langle helmet - on - shelf \rangle$ in Fig.1. Predicate probability distributions arise from different causal properties in re-implemented MOTIFS [42] as shown in (a) to (c). Predictions generated from prior knowledge in (c) are intuitive guesses obtained from the probability distribution in the dataset, after the categories of subject and object are known. It is obvious that predictions of these branches cause ambiguity because maximum logits of (a), (b), and (c) are not all the same. After CPAM, the wrong prediction is corrected.

The existing SGG models can be roughly divided into two categories: two-stage methods [22, 35, 45] and one-stage methods. One-stage methods [17–19] predict object labels and relationships at last simultaneously, which loses the assistance of language semantics during training. The co-occurrence of object pairs and predicates in the dataset shows that knowing the categories of subjects and objects are quite helpful to make a correct prediction, such as the predicate between “person” and “horse” is probably “riding on”. Specifically, two-stage methods can provide extra semantic information of object labels, while the one-stage methods are not able to do this. Currently, two-stage methods are in the slight majority, and we also use this method to build our pipeline.

2.2 Evidence Theory

In 1967, Dempster proposed evidence theory and applied it to statistical problems [4, 5]. After that, Shafer published the first monograph of evidence theory in 1976. By introducing the concept of belief function, Shafer developed and improved evidence theory, namely Dempster Shafer (D-S) theory, marking the birth of evidence theory [27].

Currently, evidence theory has been rapidly developed because of its strong ability in uncertain inference and this is why we apply it to model the ambiguity brought by different predictions of casual graphs. Evidence theory is widely used in pattern recognition [6], information fusion [10], artificial intelligence [1], expert systems [36], etc. Researchers in the community of computer vision are trying to introduce evidence theory to model the uncertainty [8, 15, 31, 33]. More detail background knowledge can be found in [11, 13, 37, 40].

3 Approach

3.1 Overview of Approach

Conventional SGG models can be defined as $SG = (B, O, R)$, where B, O and R represent the bounding box, object and relational prediction model respectively. The whole process of SGG can be expressed in the following form conventionally. I is the given image.

$$P(SG|I) = P(B|I)P(O|B, I)P(R|O, B, I) \tag{1}$$

The Faster R-CNN [26] is adopted to return coordinate $b_i \in \mathbb{R}^4$ of each proposal, visual feature $v_i \in \mathbb{R}^{4096}$ and object classification confidence scores $c_i \in \mathbb{R}^O$ of each node. O is the number of object categories. The visual feature x_i of each node is formed by concatenating semantic embedding feature, spatial embedding feature and extracted feature by Faster R-CNN. Then the global context is encoded and object-level context and edge-level context are obtained. Furthermore, paired boxes feature of i-th and j-th proposals is extracted, denoting as $u_{ij} \in \mathbb{R}^{4096}$. As Fig. 4(c) shown, to better capture the connection of properties in a causal graph and avoid conflict, our proposed CPAM module is adopted, which employs uncertainty modeling instead of a discriminative manner. Conventional causal properties fusion method is replaced with CPAM. As Fig.4(a) shown, our proposed HDA method is also adopted to relieve long-tail distribution.

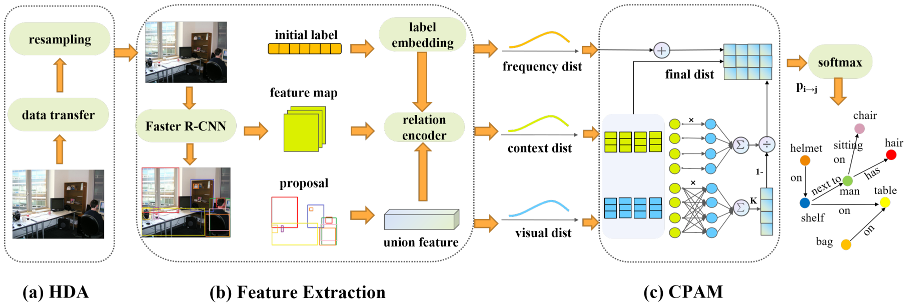


Fig. 4. The pipeline of our proposed method. In this work, we utilize Faster R-CNN as an object detector, and divide the method into three parts: (a) Enhancing the dataset to balance data distribution. (b) Extracting all the features used in the framework, including initial object relabel generated by Faster R-CNN, semantics embedding feature and encoded global context, etc. (c) Modeling data distributions of different branches based on the causal graph.

3.2 D-S Evidence Theory

D-S evidence theory models uncertainty by applying Dempster combination rules [40]. Some basic concepts of evidence theory are introduced below.

Defination 1. Frame of Discernment. The $\Theta = \{H_1, H_2, \ldots, H_N\}$ is defined as a frame of discernment which consists of a set of mutually exclusive non-empty events. Namely, $\forall i, j \in [1, N]$ and $i \neq j$, $H_i \cap H_j = \Phi$. Φ is the empty set and H_i denotes i-th event. In SGG task, events are the predicate class distribution predicted by the three branches of a casual graph.

Defination 2. Mass function. Basic Probability Assignment(BPA) is the output of the mass function. The mass function is a mapping function and denotes as follows. In our method, n is the number of predicate categories and $m(H_i)$ is the confidence score corresponding to each predicate class.

$$m(\Phi) = 0, \sum_{i=1}^{n} m(H_i) = 1 \tag{2}$$

Defination 3. Belief Function and Plausibility Function. $Bel(\cdot)$ and $Pl(\cdot)$ are basic concepts in evidence theory to measure the confidence of evidence. $Bel(A)$ denotes the sum of the BPAs of all subsets of proposition A, while $Pl(A)$ denotes the sum of BPA of all subsets intersecting proposition A.

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{3}$$

$$Pl(A) = \sum_{B \cap A \neq \Phi} m(B) \tag{4}$$

Defination 4. Dempster Combination Rules. Suppose there are two independent and completely reliable evidence, corresponding to BPAs $m1$ and $m2$ respectively. $\forall A \subseteq \Theta$, the Dempster Combination Rules (DCR) are defined by:

$$m(A) = \begin{cases} 0 & , A = \Phi \\ \frac{1}{1-K} \sum\limits_{B \cap C = A} m_1(B) m_2(C) & , A \neq \Phi \end{cases} \tag{5}$$

with

$$K = \sum_{B \cap C = \Phi} m_1(B) m_2(C) \tag{6}$$

where K is the conflict coefficient between m_1 and m_2. Note that some works denote $1 - K$ as the conflict coefficient, and the effect is the same. Predicted distribution of visual branch is modeled as B, while C represents the result from context branch. The intersection of B and C are confidence scores corresponding to the same predicate class.

3.3 Causal Property Anti-Conflict Modeling

To reduce the possibility of conflict between causal properties, a Casual Property Anti-Conflict Modeling module is proposed. To apply D-S evidence theory to SGG, we should figure out which are the frame of discernment and exclusive events. Take "person" as the subject and "motorbike" as the object for example, the events will be: "The predicate between them is riding", or "The predicate between them is sitting on". There are 50 classes of predicate in Visual Genome. For each pair of objects, the discriminant results of the predicate category are events in the frame of discernment. To capture the intrinsic uncertainty of a causal graph, we try to model the output distribution of each branch as evidence explicitly.

First, visual feature v_i and context feature c_i are projected into a subspace of the same dimension. $W_1, W_2 \in \mathbb{R}^{4096\times512}$ are linear transformation matrices. To satisfy EQU. (2), we normalize the projected features as:

$$v' = softmax(W_1^T v) \tag{7}$$

$$c' = softmax(W_2^T c) \tag{8}$$

To obtain conflict coefficient K, an auxiliary matrix A is introduced. All elements of A are 1 except that the diagonal elements are 0. N is the number of predicate categories. $I \in \mathbb{R}^{N\times N}$ is identity matrix. After getting the conflict coefficient K, the combined probability assignment is calculated as m, which is taken as the weighted coefficient of visual dist v' and context dist c' with the frequency dist of each pair of objects added. We call this approach binary causal attribute fusion. Also, a triple fusion is introduced among all of the three properties. These strategies are compared in Section 4.4. The classification score vector of relationships can be obtained as follows:

$$K = \sum_{i=1}^{N}\sum_{j=1}^{N} v'(i) \times c'(j) \times A(i,j) \tag{9}$$

$$m = \frac{1}{1-K}\sum_{i=1}^{N}\sum_{j=1}^{N} v'(i) \times c'(j) \times I(i,j) \tag{10}$$

$$p = m \times v'(i) + m \times c'(i) + frequency \tag{11}$$

The category of relationship between a pair of nodes is predicted by:

$$r = \arg\max_{r\in N}(p) \tag{12}$$

3.4 Hybrid Data Augmentation.

There are many semantically ambiguous triplets in the dataset. Some predicates share similar semantic spaces like "holding" and "carrying", while some predicates can reasonably describe relations at the same time, such as "on" and "standing on".

To deal with the highly-skewed long-tailed distribution, a Hybrid Data Augmentation method is proposed. Based on [43], internal and external data transfer are adopted to deal with semantic ambiguity and long-tailed problems. Internal data transfer tries to transform general predicates into informative predicates, and external data transfer takes advantage of negative samples. Specifically, negative samples are relabeled to generate a more diverse training set.

Furthermore, to get a more enhanced dataset, the conventional resampling strategy [2] is introduced. Predicate categories in the tail of data distribution were up-sampled according to sample fraction during training. $Count(\cdot)$ denotes the number of training samples in the dataset. We set $p = 3.0$ as default, and the sampling rate φ_i is calculated as below.

$$\varphi_i = \begin{cases} 1.0 \quad , if \quad \frac{\sum_{k=1}^{M} Count(p_k)}{Count(p_i)} < m \\ p \quad , if \quad \frac{\sum_{k=1}^{M} Count(p_k)}{Count(p_i)} \geq m \end{cases} \tag{13}$$

4 Experiment

In this section, a series of comprehensive experiments are conducted to validate the effectiveness of our method. The generalizability of our method is demonstrated by plugging into different baseline models. Below introduces implementation details while training. Then we show experimental analysis and do ablation studies on Visual Genome.

4.1 Evaluation Settings

Dataset. The popular Visual Genome [14] is used to train and evaluate our method, and followed the popular split VG-150 [32, 42] benchmark. VG-150 is composed of 108k images with the most frequent 150 object categories and 50 predicate categories. The whole dataset is divided into a training set and a testing set, which includes 70% images and 30% images separately. We also preserve 5k images for validation.

Tasks. In SGG, three widely-used subtasks are supposed to implement. (1) Predicate Classification (PredCls) classifies predicate categories with ground truth bounding boxes and labels. (2) Scene Graph Classification (SGCls) classifies object categories and predicate categories only with correct localization. (3) Scene Graph Detection (SGDet) requires a model to localize objects and recognize both objects and predicate classes. Namely, SGDet asks the model to detect scene graphs from scratch.

Metrics. Following the previous works [3, 29, 30], Mean Recall@K (mR@K) metric is used as our evaluation metrics. mR@K could evaluate the performance of a model more fairly, because it treats each predicate category as equal, and does not give more importance to "head" predicates due to the number of samples. mR@K first calculates Recall@K (R@K) of each predicate, and then averages them for all predicates. Furthermore, Zero-Shot Recall@K (zR@k) metric is introduced to better evaluate model performance, which was firstly proposed by [24]. zR@K aims to evaluate the generalization ability of the model when encountering triplets unseen in the training set. In other words, zR@K reports only the R@K of unseen triples.

4.2 Implementation Details

Object Detector. Following previous works [32, 42], a two-stage method is applied to build the overall model. For object detector, we employ a pre-trained Faster R-CNN and adopt ResNeXt-101-FPN as backbone. To reduce the computation cost, the parameters of the object detector are frozen while training. The object detector has 38.52 mAP on the training set and 28.14 mAP on the testing set.

Scene Graph Generation. In the training process, a SGD optimizer with an initial learning rate of 5e-3 is applied. We do not decay the learning rate while training but apply a warmup strategy to make the whole training process steady. The batch size of the three subtasks was set to be 8. Originally, the time complexity of all candidate pair boxes is O(n^2). For HDA, all sampling rates are set to 1.0 for SGDet. To limit the number of candidate pair boxes, candidate bounding boxes are sorted in descending order by confidence scores and only choose 256 candidate pairs, thus a lot of time and computation can be saved. Furthermore, the ground truth triplets are added while training in case some gt boxes are missing only for SGDet. And during preparing candidate pair of objects, overlapping boxes are not required, because there is no need for a spatial overlap between the subject and the object, like $\langle person - throw - ball\rangle$ or $\langle girl - looking\,at - kite\rangle$.

4.3 Comparisons with State-of-the-art Methods

We evaluate three models on VG-150 dataset: MOTIFS [42], VCTree [30] and VTransE [45] to demonstrate the generalization ability of our proposed method. As Table 1 shown, our proposed method can boost all metrics in three subtasks (PredCls, SGCls, SGDet) and achieve state-of-the-art. The model architecture includes LSTM, TreeLSTM ,and translation embedding. Furthermore, the training process can be divided into supervised learning and reinforcement learning. For the MOTIFS model, our method achieves 25.0% higher on mR@100 for PredCls. Compared with other plug-and-play methods, our method outperforms all of them in nearly all metrics. For VTransE model, in particular, which has the

Table 1. Performance (%) of our method and other state-of-the-art models on VG-150. The re-implemented models under codebase of [29] are denoted by †.

	Predicate Classification			Scene Graph Classification			Scene Graph Detection		
Models	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
FC-SGG [23]	4.9	6.3	7.1	2.9	3.7	4.1	2.7	3.6	4.2
KERN [3]	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
GBNet [41]	-	22.1	24.0	-	12.7	13.4	-	7.1	8.5
BA-SGG [12]	26.7	31.9	34.2	15.7	18.5	19.4	11.4	14.8	17.1
PCPL [34]	-	35.2	37.8	-	18.6	19.6	-	9.5	11.7
BGNN [16]	-	30.4	32.9	-	14.3	16.5	-	10.7	12.6
GPS-Net [22]	-	19.2	21.4	-	11.7	12.5	-	7.4	9.5
MOTIFS† [42]	11.5	14.6	15.8	6.5	8.0	8.5	4.1	5.5	6.8
-TDE [29]	18.5	25.5	29.1	9.8	13.1	14.9	5.8	8.2	9.8
-CogTree [39]	20.9	26.4	29.0	12.1	14.9	16.1	7.9	10.4	11.8
-CPAM+HDA(ours)	29.3	37.3	40.8	17.3	20.8	22.2	9.7	12.2	13.7
VCTree† [30]	11.7	14.9	16.1	6.2	7.5	7.9	4.2	5.7	6.9
-TDE [29]	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
-CogTree [39]	22.0	27.6	29.7	15.4	18.8	19.9	7.8	10.4	12.1
-CPAM+HDA(ours)	25.9	33.5	38.2	17.7	22.9	26.0	9.7	11.7	13.8
VTransE† [45]	11.6	14.7	15.8	6.7	8.2	8.7	3.7	5.0	6.0
-TDE [29]	18.9	25.3	28.4	9.8	13.1	14.7	6.0	8.5	10.2
-CPAM+HDA(ours)	25.9	34.6	38.7	15.1	18.9	21.4	10.3	14.6	17.1

weakest performance for SGDet in baseline, our method has made tremendous improvement with 65.0% on mR@100.

Meanwhile, compared with other strong baselines, our method still achieve competitive performance when compared with state-of-the-art KERN [3], GB-Net [41], BA-SGG [12], PCPL [34], BGNN [16] and GPS-Net [22]. Considering SGDet, our method is slightly lower than BA-SGG, while performs best in PredCls and SGCls.

On the other hand, model performance on zR@K is also evaluated as Table 2 shown. The focal loss [21], reweight [34] and resample [7, 16] are chosen as conventional plug-and-play debiasing methods, while EBM [28] as a debiasing method using deep learning model. Compared with these debiasing methods, our method outperforms them. Specifically, for MOTIFS, our method performs almost five times better than the baseline, changing the embarrassing result that almost no unseen triplets are recognized on zR@20.

4.4 Ablation Studies

Model Components. CPAM is proposed to solve the conflict problem among causal attributes, and HDA to deal with the imbalance distribution in VG-150. To prove the effectiveness of the above components, ablation studies are performed to get a clear sense of how these different components affect final performance. Experiments on VCTree [30] are performed, and the results are reported in Table 3. w/o-HDA represents models without HDA, and only CPAM works. We observe that all the metrics are improved, which means the existence of causal property conflict while training and our proposed CPAM can effectively

Table 2. Performance (%) of Zero-Shot Recall (zR@K) of baseline model and model using our method on VG-150.

	Scene Graph Detection		
Models	zR@20	zR@50	zR@100
MOTIFS [42]	0.0	0.1	0.2
-Focal [21]	-	0.1	0.3
-Resample	-	0.1	0.3
-Reweight	-	0.0	0.0
-EMB [28]	0.2	0.3	-
-CPAM+HDA(ours)	0.4	0.6	1.0
VCTree [30]	0.2	0.5	0.7
-EMB [28]	0.3	0.6	-
-CPAM+HDA(ours)	0.6	1.1	1.5

eliminate it. Meanwhile, CPAM improves the performances of all metrics. w/o-CPAM represents models evaluating only with HDA. We also find the performance outperforms the baseline with a large margin, which means the imbalance distribution (i.e. long-tailed bias) impairs model performance extremely.

To sum up, the necessity of all components are demonstrated. Only when these two components are applied at the same time, the model may achieve the best performance.

Fusion types for CPAM. As aforementioned, the two fusion types for CPAM are proposed: binary fusion and triple fusion. Binary fusion fuses visual distribution and context distribution, with the frequency distribution added. Triple fusion fuses all of the three causal attribute distributions simultaneously. As shown in Table 4, the performance of these two fusion types is compared on several baseline models.

Intuitively, the more causal attributes fuse, the better performance a model could behave. However, we observe an opposite result: triple fusion is generally weaker than binary fusion. Therefore, the reasons are shown as follows: (1) As prior knowledge, the frequency distribution is a fixed attribute, while visual and context distribution is uncertain. If certain attributes and uncertain attributes are forced to combine, it will violate the premise of D-S evidence theory, introducing extra bias at the same time. (2) CPAM is applied at the end of a model to calculate the joint distribution of the three attributes directly, and this will lead to a slight gradient vanishing through backward propagation. The fact is that the fusion of three pieces of evidence is more computationally intensive than the fusion of two, so the problem of gradient vanishing will be enlarged.

In a word, the performance of binary fusion with triple fusion is compared and we conclude that binary fusion is better. Also, the experimental results are analyzed and the reasons for them are summarized.

Table 3. Ablation study of model components.

	PredCls		SGCls		SGDet	
	mR@50	mR@100	mR@50	mR@100	mR@50	mR@100
baseline	14.9	16.1	7.5	7.9	5.7	6.9
w/o-HDA	17.4	18.9	11.1	11.9	7.3	8.7
w/o-CPAM	30.3	33.9	16.5	18.1	11.5	14.0
CPAM+HDA	33.5	38.2	22.9	26.0	11.7	14.1

Table 4. Ablation study of CPAM using different fusion types.

		PredCls		SGCls		SGDet	
		mR@50	mR@100	mR@50	mR@100	mR@50	mR@100
MOTIFS [42]	triple	15.1	16.4	8.0	9.0	6.7	8.1
	binary	15.6	16.9	9.0	9.6	6.9	8.2
VTransE [45]	triple	15.6	16.8	8.0	8.5	6.0	7.1
	binary	16.7	18.0	8.9	9.5	6.4	7.6
VCTree [30]	triple	15.7	17.8	9.1	9.7	5.6	6.5
	binary	17.4	18.9	11.1	11.9	7.3	8.7

4.5 Qualitative Studies

Several testing examples of VTransE on the PredCls subtask are visualized . As Fig. 5 shown, our method generates more fine-grained relationships such as "people-near-train" v.s. "people-looking at-train" in the first row and "tree-near-building" v.s. "tree-in front of-building" in the second row. Take predicate *on*, for example, it is divided into *painted on*, *parked on*, *walking on*, etc, and the latter predicates are more semantically informative instead of trivial and common. Head predicates such as *has*, *of*, *in*, and *near* rarely appear in predictions and the performance of tail predicates is promoted.

Furthermore, R@100 of each predicate category is calculated for baseline and our method as Fig. 6 shown. It can proved our inference proposed before that long-tailed distribution is alleviated. Specifically, we observe the drop of head predicate and substantial improvement of the tail.

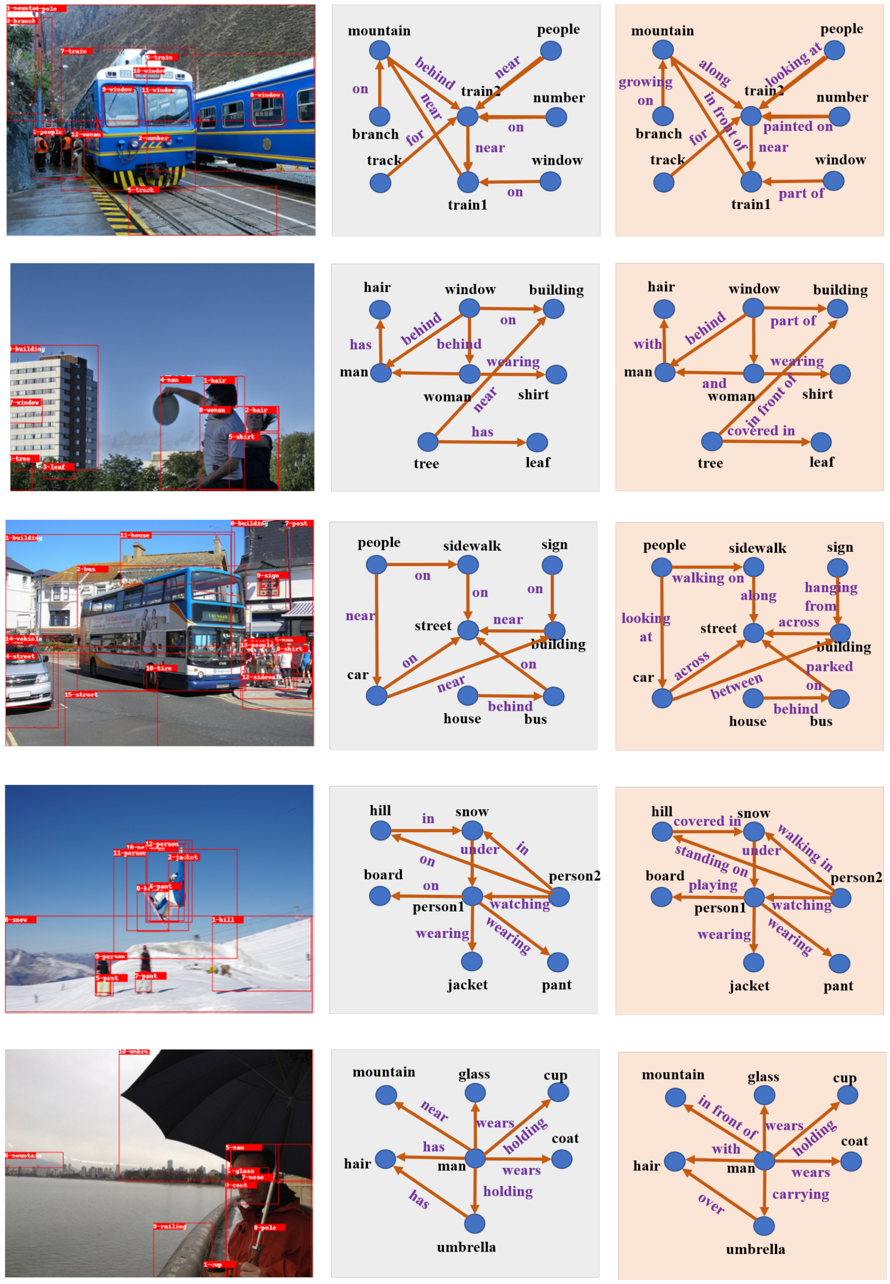


Fig. 5. Qualitative results of VTransE [45] (gray) and VTransE adopting our proposed method (orange).

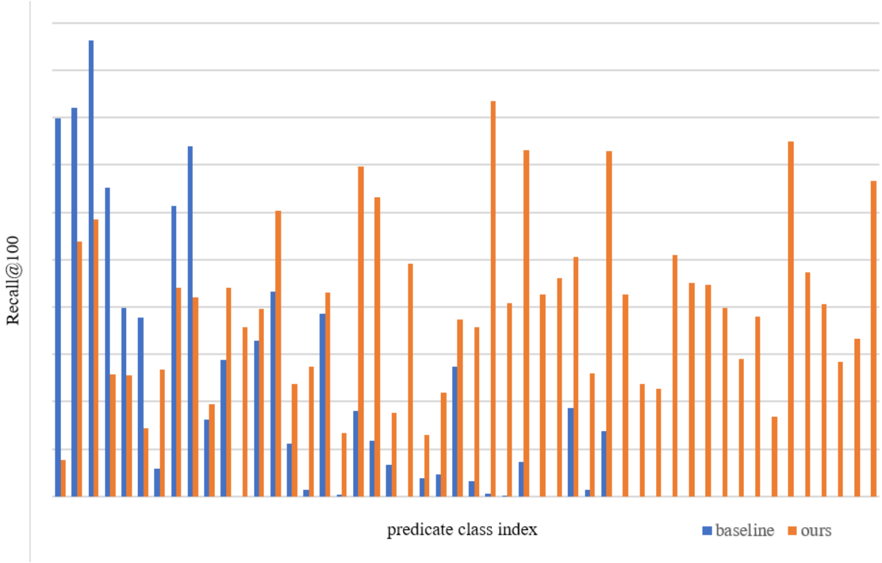


Fig. 6. R@100 of all the predicate classes of baseline and our method on VG-150.

5 Conclusion

In this work, the problems are analyzed in SGG, which could be attributed to long-tailed bias, and an extra bias caused by conflicts of different causal branches. To address the problems mentioned above, a novel plug-and-play CPAM module is proposed, applying the Dempster-Shafer evidence theory to eliminate conflict among causal attributions. The predicted distribution of each branch is regarded as evidence and Dempster combination rules are employed to model the uncertainty. Meanwhile, a HDA method is introduced to get an enhanced dataset by integrating IETrans and conventional debiasing strategies. Through extensive comparative experiments, our method achieved state-of-the-art performance under the existing evaluation metric mR@K of the three subtasks (PredCls, SGCls, SGDet), either model-agnostic baselines or specific baselines.

Furthermore, comprehensive ablation studies are conducted to demonstrate the universal effectiveness of CPAM and HDA, and the effectiveness of both model components are validated. For CPAM, we do more detailed experiments. Two fusion types of CPAM are proposed and the performance of binary fusion type and triple fusion type is compared. Then we show scene graphs generated by baseline and our method, and the results demonstrate our method can perform more fine-grained predictions and relieve long-tailed distribution effectively.

Acknowledgements This work was supported by the National Key R&D Program of China (2019YFB2204200) and the National Natural Science Foundation of China (62006015 and 62072028).

References

1. Barnett, J.A.: Computational methods for a mathematical theory of evidence. In: Classic Works of the Dempster-Shafer Theory of Belief Functions, pp. 197–216. Springer (2008)
2. Burnaev, E., Erofeev, P., Papanov, A.: Influence of resampling on accuracy of imbalanced classification. In: Eighth international conference on machine vision (ICMV 2015). vol. 9875, pp. 423–427. SPIE (2015)
3. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6171 (2019)
4. Dempster, A.P.: A generalization of bayesian inference. Journal of the Royal Statistical Society: Series B (Methodological) **30**(2), 205–232 (1968)
5. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. In: Classic works of the Dempster-Shafer theory of belief functions, pp. 57–72. Springer (2008)
6. Denœux, T., Zouhal, L.M.: Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy sets and systems **122**(3), 409–424 (2001)
7. Desai, A., Wu, T.Y., Tripathi, S., Vasconcelos, N.: Learning of visual relations: The devil is in the tails. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15404–15413 (2021)
8. Dong, M., Peng, J., Ding, S., Wang, Z.: Vision and emg information fusion based on ds evidence theory for gesture recognition. In: Proceedings of 2021 Chinese Intelligent Automation Conference. pp. 492–501. Springer (2022)
9. Dong, X., Gan, T., Song, X., Wu, J., Cheng, Y., Nie, L.: 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. pp. 19427–19436 (2022)
10. Florea, M.C., Jousselme, A.L., Bossé, É., Grenier, D.: Robust combination rules for evidence theory. Information Fusion **10**(2), 183–197 (2009)
11. Gordon, J., Shortliffe, E.H.: A method for managing evidential reasoning in a hierarchical hypothesis space. Artificial intelligence **26**(3), 323–357 (1985)
12. 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 the IEEE/CVF International Conference on Computer Vision. pp. 16383–16392 (2021)
13. Jiang, W.: A correlation coefficient for belief functions. International Journal of Approximate Reasoning **103**, 94–106 (2018)
14. 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. International journal of computer vision **123**(1), 32–73 (2017)
15. Li, B., Han, Z., Li, H., Fu, H., Zhang, C.: Trustworthy long-tailed classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6970–6979 (2022)
16. Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11109–11119 (2021)
17. Li, Y., Ouyang, W., Wang, X., Tang, X.: Vip-cnn: Visual phrase guided convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1347–1356 (2017)

18. Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., Wang, X.: Factorizable net: an efficient subgraph-based framework for scene graph generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 335–351 (2018)
19. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE international conference on computer vision. pp. 1261–1270 (2017)
20. Liao, S.H.: Expert system methodologies and applications—a decade review from 1995 to 2004. Expert systems with applications **28**(1), 93–103 (2005)
21. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
22. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3746–3753 (2020)
23. Liu, H., Yan, N., Mortazavi, M., Bhanu, B.: Fully convolutional scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11546–11556 (2021)
24. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision. pp. 852–869. Springer (2016)
25. Qi, S., Zhu, Y., Huang, S., Jiang, C., Zhu, S.C.: Human-centric indoor scene synthesis using stochastic grammar. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5899–5908 (2018)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
27. Shafer, G.: A mathematical theory of evidence. In: A mathematical theory of evidence. Princeton university press (2021)
28. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.: Energy-based learning for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13936–13945 (2021)
29. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3716–3725 (2020)
30. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6619–6628 (2019)
31. Wang, X., Wang, T.: Research on face recognition algorithm based on ds evidence theory and local domain pattern. In: 2021 International Conference on Intelligent Computing, Automation and Applications (ICAA). pp. 261–266. IEEE (2021)
32. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5410–5419 (2017)
33. Xu, Z., Zhang, B., Fu, H., Yue, X., Lv, Y.: Multi-branch recurrent attention convolutional neural network with evidence theory for fine-grained image classification. In: International Conference on Belief Functions. pp. 177–184. Springer (2021)
34. 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 the 28th ACM International Conference on Multimedia. pp. 265–273 (2020)

35. Yang, G., Zhang, J., Zhang, Y., Wu, B., Yang, Y.: Probabilistic modeling of semantic ambiguity for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12527–12536 (2021)
36. Yang, J.B., Liu, J., Wang, J., Sii, H.S., Wang, H.W.: Belief rule-base inference methodology using the evidential reasoning approach-rimer. IEEE Transactions on systems, Man, and Cybernetics-part A: Systems and Humans **36**(2), 266–285 (2006)
37. Yang, J.B., Xu, D.L.: Evidential reasoning rule for evidence combination. Artificial Intelligence **205**, 1–29 (2013)
38. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10685–10694 (2019)
39. Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q.: Cogtree: Cognition tree loss for unbiased scene graph generation. arXiv preprint arXiv:2009.07526 (2020)
40. Zadeh, L.A.: On the validity of Dempster's rule of combination of evidence. Infinite Study (1979)
41. Zareian, A., Karaman, S., Chang, S.F.: Bridging knowledge graphs to generate scene graphs. In: European Conference on Computer Vision. pp. 606–623. Springer (2020)
42. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5831–5840 (2018)
43. Zhang, A., Yao, Y., Chen, Q., Ji, W., Liu, Z., Sun, M., Chua, T.S.: Fine-grained scene graph generation with data transfer. arXiv preprint arXiv:2203.11654 (2022)
44. Zhang, C., Chao, W.L., Xuan, D.: An empirical study on leveraging scene graphs for visual question answering. arXiv preprint arXiv:1907.12133 (2019)
45. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5532–5540 (2017)