Stacked Hybrid-Attention and Group Collaborative Learning for Unbiased Scene Graph Generation

Xingning Dong[1], Tian Gan[1]†, Xuemeng Song[1], Jianlong Wu[1], Yuan Cheng[2]†, Liqiang Nie[1]
[1]Shandong University, [2]Ant Group
dongxingning1998@gmail.com, gantian@sdu.edu.cn, sxmustc@gmail.com
jlwu1992@sdu.edu.cn, chengyuan.c@antgroup.com, nieliqiang@gmail.com

Abstract

Scene Graph Generation, which generally follows a regular encoder-decoder pipeline, aims to first encode the visual contents within the given image and then parse them into a compact summary graph. Existing SGG approaches generally not only neglect the insufficient modality fusion between vision and language, but also fail to provide informative predicates due to the biased relationship predictions, leading SGG far from practical. Towards this end, we first present a novel Stacked Hybrid-Attention network, which facilitates the intra-modal refinement as well as the inter-modal interaction, to serve as the encoder. We then devise an innovative Group Collaborative Learning strategy to optimize the decoder. Particularly, based on the observation that the recognition capability of one classifier is limited towards an extremely unbalanced dataset, we first deploy a group of classifiers that are expert in distinguishing different subsets of classes, and then cooperatively optimize them from two aspects to promote the unbiased SGG. Experiments conducted on VG and GQA datasets demonstrate that, we not only establish a new state-of-the-art in the unbiased metric, but also nearly double the performance compared with two baselines. Our code is available at https://github.com/dongxingning/SHA-GCL-for-SGG.

1. Introduction

Scene Graph Generation (SGG) [41] targets at organizing all the objects and their pairwise relationships into a compact summary graph. As an intermediate visual understanding task, SGG could benefit various vision-and-language tasks, including cross-modal retrieval [6, 11, 28], image captioning [2, 10, 51], and visual question answering [12, 32, 48]. However, SGG is still far from satisfactory for practical applications due to the insufficient modality fusion and the biased relationship predictions.

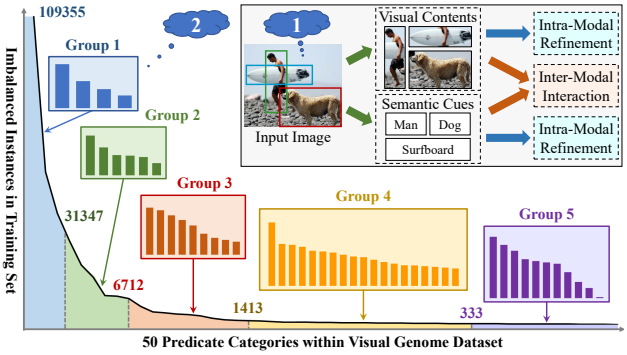


Figure 1. Two intentions to promote the unbiased SGG. (1) For the insufficient modality fusion, we aim to enhance both the intra-modal refinement and the inter-modal interaction (see the top-right corner of the figure). And (2) we split the extremely unbalanced dataset into a set of relatively balanced groups, based on which we configure the classification space for all the newly-added classifiers (see the rest part of the figure).

Though it is manifestly proved that incorporating semantic cues (language priors of object class names) into visual contents (object proposals) could significantly improve the generation capability [18,21], most of the recent approaches [17, 26, 30, 31, 42, 43, 46, 47] simply fuse these visual and semantic features by summing up directly or concatenation, which limits the model to further infer their interaction information. To address this under-explored insufficient modality fusion between visual contents and semantic cues, we aim to strengthen the encoder via jointly exploring the intra-modal refinement and the inter-modal interaction, as illustrated in Figure 1. To implement this intention, we first design the Self-Attention (SA) unit and the Cross-Attention (CA) unit to capture the intra-modal and inter-modal information, respectively. We then organize these two units into a Hybrid-Attention (HA) layer, and stack several HA layers to build the encoder. The proposed Stacked Hybrid-Attention (SHA) network could adequately explore the multi-modal interaction, thus improving the relationship prediction performance.

†Corresponding authors.

The other prominent issue faced by existing SGG methods is the biased relationship predictions due to the long-tailed data distribution. Since only a few head predicates (*e.g.*, *on*, *has*) possess massive and various instances, they would dominate the training procedure and lead the output scene graphs with few informative tail predicates (*e.g.*, *riding*, *watching*), which could hardly support a wide range of downstream tasks. Though various debiasing approaches [4, 29, 37] have been proposed, they are vulnerable to over-fitting the tail classes and sacrificing much on the head ones, leading to the other extreme. In a sense, we conjecture that this dilemma may root in the fact that a naive SGG model, regardless of the conventional or debiasing one, could only differentiate a limited range of predicates whose amount of training instances are relatively equal.

Intuitively, since a single classifier struggles in achieving a reasonable prediction trade-off, we can divide the biased predicate classes into several balanced subsets, then introduce more classifiers to conquer each of them, and ultimately leverage these classifiers to cooperatively address this challenge. To fulfill this "divide-conquer-cooperate" intuition, we propose the Group Collaborative Learning (GCL) strategy, where we 1) **first divide**: As a single classifier is adequate to differentiate the classes within a balanced dataset, we first divide all the predicates into a set of relatively balanced groups according to their amount of training instances, as illustrated in Figure 1. 2) **Then conquer**: We then borrow the idea from the class-incremental learning [14] to force all the classifiers to follow a continuously growing classification space, *i.e.*, each classifier would extend the previous classification space by incorporating a newly-added group of predicates. Besides, we devise the Median Re-Sampling strategy to provide each classifier with a relatively balanced training set. Based on this group-incremental configuration, these nested classifiers could fairly treat the predicates within their classification space, thus they would be more likely to learn the discriminating representations, especially towards the newly-added group. 3) **Ultimately cooperate**: We further leverage these classifiers to cooperatively enhance the unbiased relationship predictions from two aspects. First, we propose the Parallel Classifier Optimization (PCO) to jointly optimize all the classifiers. This can be seen as a "weak constraint", since we expect that gathering all the gradients could promote the recognition capability of each classifier. Second, we devise the Collaborative Knowledge Distillation (CKD) to ensure that the discriminating capability learned previously could be well translated to the subsequent classifiers. This can be seen as a "strong constraint", since we force each classifier to mimic the prediction behavior from its predecessors. By employing these two constraints, we effectively mitigate the overwhelming punishments to the tail classes as well as compensate for the under-fitting on the head ones.

The contributions of our work are three-folds:

- We present a novel Stacked Hybrid Attention network to strengthen the encoder in SGG, which addresses the under-explored insufficient modality fusion problem.
- We design the Group Collaborative Learning strategy to optimize the decoder in SGG. Particularly, we deploy a group of classifiers and cooperatively optimize them from two aspects, thus effectively addressing the intractable biased relationship prediction problem.
- Experiments conducted on VG and GQA dataset indicate that, we not only establish a new state-of-the-art in the unbiased metric, but also nearly double the performance compared with two typical baselines when employing our model-agnostic GCL.

2. Related Work

Scene Graph Generation. SGG provides an efficient way for scene understanding by decoding the visual relationships into a summary graph. Early approaches [5, 18, 19, 21] were mainly dedicated to incorporating more features from various modalities, but they neglected the rich visual context, leading to sub-optimal performance. In order to tackle such deficiency, later approaches employed more powerful feature refinement modules to encode the rich contextual information, such as message passing strategy [17, 40], sequential LSTMs [31, 47], graph neural networks [3, 46], and self-attention networks [20, 26]. Though the performance is improved in the regular metrics, the relations they predicted are often trivial and less informative due to the biased training data, which could hardly support the downstream vision-and-language tasks. Therefore, various approaches [4, 29, 37] have been proposed to tackle the biased relationship predictions, including employing debiasing strategies like re-sampling [17] or re-weighting [42], disentangling unbiased representations from the biased [30], and utilizing the tree structure to filter the irrelevant predicates [43]. However, these approaches are vulnerable to over-fitting on the tail classes with much sacrifice on the head ones. Based on the observation that a single classifier could hardly differentiate all the classes within a biased dataset, and inspired by the "divide-conquer-cooperate" intuition, we propose the Group Collaborative Learning strategy to guide the training of the decoder. In this way, we not only significantly improve the prediction performance towards the tail classes, but also effectively preserve the discriminating capability learned by the head ones, thus achieving a reasonable prediction trade-off.

Cross-attention Models. Research towards improving multi-modal fusion [35, 36] and building cross-attention models [38, 50] have been attracting increasing interest in various vision-and-language tasks. For example, Yu *et al.* [44] proposed the deep Modular Co-Attention Network to

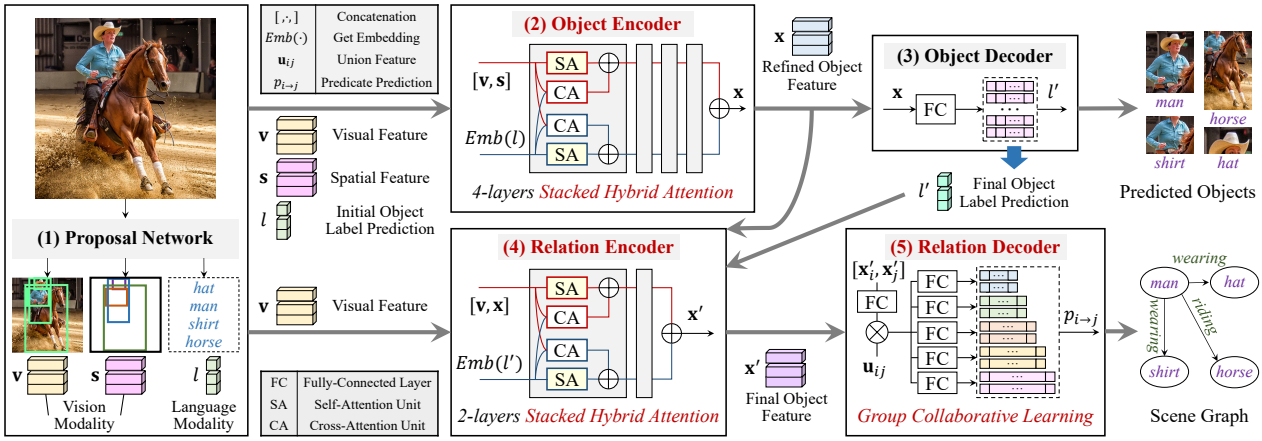


Figure 2. The framework of the common pipeline in SGG, which includes five key components. Notably, we improve three key components marked in red in the figure. Specifically, we propose the Stacked Hybrid-Attention network to enhance the object encoder and the relation encoder, and we also devise the Group Collaborative Learning strategy to guide the training of the relation decoder.

fully model the interaction between question words and image regions in VQA, and Lu *et al.* [22] proposed ViLBERT to extend BERT architecture for jointly pre-training images and texts. Nevertheless, few of the approaches in SGG dedicate to addressing the insufficient modality fusion between object proposals and their corresponding class names. Therefore, we propose the Stacked Hybrid-Attention (SHA) network to facilitate both the intra-modal refinement and the inter-modal interaction.

Knowledge Distillation. Knowledge distillation [9, 13, 23] aims to distill the knowledge from a larger deep network into a small one, which is widely employed in various tasks, including model compression [1,34], label smoothing [27, 45], and data augmentation [7, 8]. Note that the conventional knowledge distillation approaches generally follow a teacher-student pipeline. These two networks are optimized in different time steps as the teacher network is usually available beforehand. Different from this model-to-model paradigm, after adding several classifiers, we allow the previous classifiers to generate the outputs as soft labels to constrain the training of the subsequent, thus establishing a layer-to-layer "knowledge transfer".

3. Methodology

3.1. Problem Formulation

SGG aims to generate a summary graph $\mathcal{G}$ that highly generalizes the contents of a given image $\boldsymbol{I}$. Towards this end, we first detect all the objects within the image $\boldsymbol{I}$, denoted as $\mathcal{O} = \{o_i\}_{i=1}^{N}$. Then for each object pair (o_i, o_j), we predict its predicate $p_{i\to j}$. Ultimately, we organize all these predictions in the form of triplets to construct the scene graph, which can be formulated as $\mathcal{G} = \{(o_i, p_{i\to j}, o_j) | o_i, o_j \in \mathcal{O}, p_{i\to j} \in \mathcal{P}\}$, where $\mathcal{P}$ stands for the set of all the possible predicates.

3.2. Overall Framework

As illustrated in Figure 2, our framework is based on the common pipeline followed by typical SGG approaches [31,42,43,47], which is a regular encoder-decoder structure.

Proposal Network is actually a pre-trained object detector. Given an image $\boldsymbol{I}$, it generates a set of object predictions $\mathcal{O} = \{o_i\}_{i=1}^{N}$. For each object o_i, it provides a visual feature $\mathbf{v}_i$, a spatial feature $\mathbf{s}_i$ of the bounding box coordinates, and an initial object label prediction l_i.

Object Encoder aims to obtain the refined object feature $\mathbf{x}_i$ for further predictions, which is calculated as:

$$\mathbf{x}_i = Enc^{obj}([\mathbf{v}_i, FC(\mathbf{s}_i)], Emb(l_i)), \tag{1}$$

where $Enc^{obj}(\cdot)$ represents the object encoder, which can be any feature refinement modules (*e.g.*, BiLSTMs [47] and GNNs [3]), $[,\cdot,]$ denotes the concatenation operation, $FC(\cdot)$ represents a fully-connected layer, and $Emb(\cdot)$ refers to a pre-trained language model to acquire the semantic feature of o_i based on its initial object label prediction l_i.

Object Decoder aims to obtain the final object label prediction l_i' based on the refined object feature $\mathbf{x}_i$, which is calculated as:

$$l_i' = \text{argmax}(\text{Softmax}(Dec^{obj}(\mathbf{x}_i))), \tag{2}$$

where $Dec^{obj}(\cdot)$ represents the object decoder, which is a single fully-connected layer.

Relation Encoder works on obtaining the final object feature $\mathbf{x}_i'$ for predicate predictions, which is calculated as:

$$\mathbf{x}_i' = Enc^{rel}([\mathbf{v}_i, \mathbf{x}_i], Emb(l_i')), \tag{3}$$

where $Enc^{rel}(\cdot)$ represents the relation encoder, which shares the same architecture with the object encoder.

Relation Decoder is responsible for predicting the predicate label $p_{i\to j}$ based on the final object features of subject

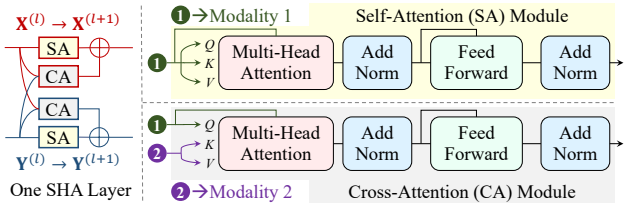


Figure 3. One single Stacked Hybrid-Attention (SHA) layer is composed of two types of attention units, *i.e.*, Self-Attention (SA) unit to facilitate the intra-modal refinement and Cross-Attention (CA) unit to promote the inter-modal interaction.

o_i and object o_j, which is calculated as:

$$p_{i \rightarrow j} = \text{argmax}(\text{Softmax}(Dec^{rel}(\mathbf{x}'_i, \mathbf{x}'_j, \mathbf{u}_{ij})), \quad (4)$$

where $Dec^{rel}(\cdot)$ represents the relation decoder. We also follow [47] to employ the union feature $\mathbf{u}_{ij}$ of the object pair (o_i, o_j) to enhance the predicate predictions.

It is worth noting that we improve three key components marked in red in Figure 2 to promote the unbiased SGG. Specifically, for the object encoder and the relation encoder, we propose the Stacked Hybrid-Attention (SHA) network to alleviate the insufficient modality fusion problem. Regarding the relation decoder, we devise the Group Collaborative Learning (GCL) strategy to address the intractable biased relationship prediction problem.

3.3. Encoder: Stacked Hybrid-Attention

Beyond understanding the visual contents (object proposals) of a given image, the semantic cues (refer to the class names in SGG) are also indispensable for robust relationship predictions. Unfortunately, most of the approaches in SGG simply fuse these two modal features by summing up directly or concatenation, which may be insufficient to mine the underlying inter-modal interaction, thus resulting in sub-optimal performance. To address this deficiency, we propose the Stacked Hybrid Attention (SHA) network, which is composed of several SHA layers. Each SHA layer contains two parallel Hybrid-Attention (HA) cells, and each HA cell is a composition of two types of attention units, *i.e.*, the Self-Attention (SA) unit to facilitate the intra-modal refinement, and the Cross-Attention (CA) unit to model the inter-modal interaction. As shown in Figure 3, both the SA unit and CA unit are built upon a multi-head attention module and a feed-forward module based on the attention mechanism [33]. The difference between SA and CA is whether the input features belong to the same modality.

Ultimately, we build our SHA network by cascading L SHA layers in sequential order. For the l-th SHA layer, the feature propagation process can be formulated as:

$$\begin{cases} \mathbf{X}^{(l)} = SA(\mathbf{X}^{(l-1)}) + CA(\mathbf{X}^{(l-1)}, \mathbf{Y}^{(l-1)}), \\ \mathbf{Y}^{(l)} = SA(\mathbf{Y}^{(l-1)}) + CA(\mathbf{Y}^{(l-1)}, \mathbf{X}^{(l-1)}), \end{cases} \quad (5)$$

where $SA(\cdot)$ and $CA(\cdot)$ denote the self-attention and cross-attention computation, respectively. For the first SHA layer, we set its input feature $\mathbf{X}^{(0)} = \mathbf{X}$ and $\mathbf{Y}^{(0)} = \mathbf{Y}$, where $\mathbf{X}$ and $\mathbf{Y}$ denote the original visual feature and semantic feature, respectively. After obtaining the final visual feature $\mathbf{X}^{(L)}$ and semantic feature $\mathbf{Y}^{(L)}$ generated by the last SHA layer, we sum them up to get the refined output, which contains rich multi-modal interaction information.

3.4. Decoder: Group Collaborative Learning

As aforementioned, when facing an extremely unbalanced dataset, a naive SGG model could hardly achieve a satisfactory prediction performance on all the predicate classes. Towards this end, we aim to deploy several classifiers which are expert in distinguishing different subsets of predicates, and organize these classifiers to cooperatively address the biased relationship predictions. Based on this "divide-conquer-cooperate" intention, we propose the Group Collaborative Learning (GCL) strategy. As shown in Figure 4, GCL contains five key steps as follows:

Predicate Class Grouping aims to split the unbalanced dataset into several relatively balanced groups, and then configure the classification space for all the classifiers. Based on the observation that the recognition capability would suffer from the biased data distribution, we aim to provide each classifier with a relatively balanced training set, thus it could adequately learn the discriminating representations towards a subset of predicates. Therefore, We first sort the predicate classes by their amount of training instances in descending order, obtaining a sorted set $\mathcal{P}_{all} = \{p_i\}_{i=1}^{M}$. We then divide $\mathcal{P}_{all}$ into K mutually exclusive groups $\{\mathcal{P}_k\}_{k=1}^{K}$ according to the pre-defined threshold μ. The workflow is summarized in Algorithm 1, where $Count(p_i)$ denotes the amount of training instances towards the predicate p_i. Line 3 in Algorithm 1 ensures that, for each group $\mathcal{P}_k$, the maximal amount of training instances will be no more than μ times of the minimal amount, thus the predicates in $\mathcal{P}_k$ share a relatively equal amount.

Algorithm 1: Predicate Class Grouping.

Input: A sorted predicate set $\mathcal{P}_{all} = \{p_i\}_{i=1}^{M}$, μ
Output: K mutually exclusive groups $\{\mathcal{P}_k\}_{k=1}^{K}$

```
1 Set cur = 1, k = 1, and P_1 = {};
2 for i ← 1 to M do
3     if Count(p_cur) > μ * Count(p_i) then
4         cur = i;
5         k = k + 1;
6         Set P_k = {};
7     end
8     P_k = P_k ∪ {p_i}
9 end
```

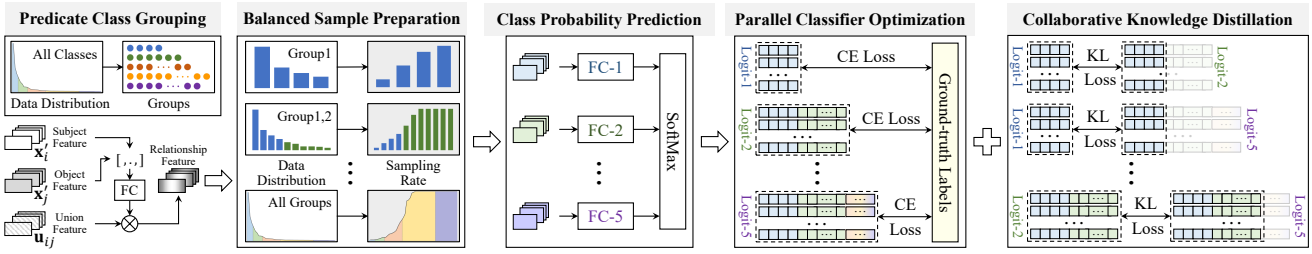


Figure 4. Illustration of the proposed Group Collaborative Learning (GCL) strategy, which includes five key steps. It is worth noting that we design two optimization mechanisms, namely Parallel Classifier Optimization (PCO) and Collaborative Knowledge Distillation (CKD), to jointly guide the training of the relation decoder.

We then borrow the idea from the class-incremental learning [14], and deploy a set of classifiers $\{\mathcal{C}_k\}_{k=1}^{K}$ which follow a continuously growing classification space. Except for the first classifier $\mathcal{C}_1$, other classifiers should recognize the predicate classes from both previous and current groups, *i.e.*, the classification space in $\mathcal{C}_k$ is $\mathcal{P}'_k = \mathcal{P}_1 \cup \mathcal{P}_2 \cup \cdots \cup \mathcal{P}_k$. Note that we only choose the last classifier $\mathcal{C}_K$ to obtain the final predicate predictions in the evaluation stage.

Balanced Sample Preparation aims to achieve several balanced training sets provided for further joint optimization by re-sampling the instances. For each classifier $\mathcal{C}_k$ that incorporates a newly-added group $\mathcal{P}_k$ to extend the previous classification space $\mathcal{P}'_{k-1}$ as $\mathcal{P}'_k = \mathcal{P}_k \cup \mathcal{P}'_{k-1}$, we expect it could adequately learn the discriminating representations towards the predicates, particularly within the newly-added group $\mathcal{P}_k$. Therefore, for the predicates in the group $\mathcal{P}_k$, we should retain all of its training instances to facilitate the convergence. And for the predicates in the previous classification space $\mathcal{P}'_{k-1}$, since they have more samples in the original dataset, we should under-sample their training instances to avoid biased predictions.

To implement this intention, we propose the Median Re-Sampling strategy to perform the re-sampling operation. For each classification space $\mathcal{P}'_k$, we first calculate the median amount $Med(\mathcal{P}'_k)$ over all the classes within $\mathcal{P}'_k$. For example, if $\mathcal{P}'_k$ is sorted in descending order and contains 9 predicate classes, the median amount $Med(\mathcal{P}'_k)$ is equal to $Count(p_5)$. Then for each predicate class p_i^k in $\mathcal{P}'_k$, we calculate the sampling rate ϕ_i^k as follows:

$$\phi_i^k = \begin{cases} \dfrac{Med(\mathcal{P}'_k)}{Count(p_i)}, & \textbf{if } Med(\mathcal{P}'_k) < Count(p_i), \\ 1.0 \quad , & \textbf{if } Med(\mathcal{P}'_k) \geq Count(p_i). \end{cases} \tag{6}$$

By employing the above strategy, each classifier would be expert in distinguishing the predicates, particularly in the newly-added group. For example, since we would under-sample the instances in Group 3 for training the 4^{th} and 5^{th} classifiers, the 3^{rd} classifier is more likely to achieve a better performance in distinguishing the predicates in Group 3, as we retain all the samples of this group to let the 3^{rd} classifier adequately learn the discriminating representations.

Class Probability Prediction aims to parse the sampled instances into the class probability logits for further loss computation and model optimization. For an object pair (o_i, o_j) chosen by the Median Re-Sampling strategy, after obtaining the subject feature $\mathbf{x}'_i$, the object feature $\mathbf{x}'_j$, and their union feature $\mathbf{u}_{ij}$, the class probability prediction $\mathbf{w}_{ij}^k$ generated by the classifier $\mathcal{C}_k$ is calculated as follows:

$$\mathbf{w}_{ij}^k = \text{Softmax}(FC([\mathbf{x}'_i, \mathbf{x}'_j]) \otimes \mathbf{u}_{ij}), \tag{7}$$

where $\otimes$ denotes the element-wise product.

Parallel Classifier Optimization aims to regularize the final classifier $\mathcal{C}_K$ by jointly optimizing all the classifiers. In the training stage, the parameters of all the K predicate classifiers would be optimized simultaneously, where the objective function can be defined as:

$$\mathcal{L}_{PCO} = \sum_{k=1}^{K} \frac{1}{|\mathcal{D}_k|} \sum_{(o_i,o_j)\in\mathcal{D}_k} \mathcal{L}_{CE}(y_{ij}, \mathbf{w}_{ij}^k), \tag{8}$$

where $\mathcal{D}_k$ denotes the set of the object pairs chosen by the Median Re-Sampling strategy, $|\cdot|$ denotes the length of the given set, y_{ij} denotes the ground-truth predicate label of the object pair (o_i, o_j), and $\mathcal{L}_{CE}(\cdot)$ is a regular Cross-Entropy cost function.

The Parallel Classifier Optimization can be seen as a "weak constraint" for Group Collaborative Learning, since we expect that gathering gradients from all the classifiers would facilitate the convergence of the final classifier $\mathcal{C}_K$.

Collaborative Knowledge Distillation aims to establish a knowledge transfer mechanism to promote the unbiased prediction capability of the final classifier $\mathcal{C}_K$. As aforementioned, each classifier specializes in distinguishing the predicates, particularly within the newly-added group. In order to preserve and translate this well-learned knowledge to compensate for the under-fitting on the head classes, we propose the Collaborative Knowledge Distillation (CKD), whose objective function is defined as:

$$\mathcal{L}_{CKD} = \frac{1}{|\mathcal{Q}|} \sum_{(m,n)\in\mathcal{Q}} \frac{1}{|\mathcal{D}_n|} \sum_{(o_i,o_j)\in\mathcal{D}_n} \mathcal{L}_{KL}(\mathbf{w}_{ij}^m, \widehat{\mathbf{w}}_{ij}^n), \tag{9}$$

where $\mathcal{Q}$ denotes the set of pairwise knowledge matching from the classifier $\mathcal{C}_m$ to the classifier $\mathcal{C}_n$ ($m < n$). We provide two alternatives, namely Adjacent and Top-Down strategy, to configure the set $\mathcal{Q}$ (these two strategies are illustrated in Figure 6 and Parameter Analysis). Note that the output $\mathbf{w}_{ij}^n$ generated by the classifier $\mathcal{C}_n$ incorporates new predicate classes which are not included in the previous classification space $\mathcal{P}'_m$, we utilize $\widehat{\mathbf{w}}_{ij}^n$ to indicate the sliced output by cutting off the incrementally-added classes which are not included in $\mathcal{P}'_m$, thus ensuring that $\widehat{\mathbf{w}}_{ij}^n$ shares the same dimension as $\mathbf{w}_{ij}^m$. $\mathcal{L}_{KL}(\cdot)$ is a regular Kullback-Leibler Divergence loss, which is defined as:

$$\mathcal{L}_{KL}(\mathbf{w}_m, \widehat{\mathbf{w}}_n) = -\sum_{l=1}^{L} \mathbf{w}_m^l \log \widehat{\mathbf{w}}_n^l. \tag{10}$$

By taking the previous predicate probability output $\mathbf{w}_{ij}^m$ from the classifier $\mathcal{C}_m$ as the soft label, CKD forces the current classifier $\mathcal{C}_n$ to mimic the prediction behaviour that $\mathcal{C}_m$ is expert in, thus can be treated as a "strong constraint".

Ultimately, the objective function of our proposed Group Collaborative Learning (GCL) is the combination of PCO and CKD, which is defined as:

$$\mathcal{L}_{GCL} = \mathcal{L}_{PCO} + \alpha \mathcal{L}_{CKD}, \tag{11}$$

where α is the pre-defined hyper-parameters to weigh the total loss $\mathcal{L}_{GCL}$. By employing these two types of constraint, we effectively mitigate the overwhelming punishments to the tail classes and compensate for the underfitting on the head ones, which benefits in establishing a reasonable trade-off during the predicate predictions.

4. Experiments

4.1. Experimental Settings

Dataset. We present the experimental results on two datasets: Visual Genome (VG) [16] and GQA [15]. VG is the most widely-used benchmark for SGG, which is composed of more than 108K images and 2.3M relation instances. Following the prior approaches [3, 4, 17, 20, 29–31, 40, 42, 43, 47, 49], we adopt the most widely-used VG150 split, which contains the most frequent 150 object classes and 50 predicate classes. GQA is another vision-and-language benchmark with more than 3.8M relation annotations. In order to achieve a representative split like VG150, we manually clean up a substantial fraction of annotations that have poor-quality or ambiguous meanings, and then select Top-200 object classes as well as Top-100 predicate classes by their frequency, thus establishing the GQA200 split. For both VG150 and GQA200, we use 70% of the images for training and the remaining 30% for testing. We also follow [47] to sample a 5K validation set from the training set for parameter tuning.

Tasks. To comprehensively evaluate the performance, we follow three conventional tasks: 1) Predicate Classification (**PredCls**) predicts the relationships of all the pairwise objects by employing the given ground-truth bounding boxes and classes; 2) Scene Graph Classification (**SGCls**) predicts the objects classes and their pairwise relationships by employing the given ground-truth object bounding boxes; and 3) Scene Graph Detection (**SGDet**) detects all the objects in an image, and predicts their bounding boxes, classes and pairwise relationships.

Evaluation Metrics. Following [4, 17, 20, 29, 30, 42, 43], we use mean Recall@K (mR@K) [3, 31], which computes the average Recall@K (R@K) for each predicate class, to evaluate the unbiased SGG. As R@K is easily dominated by the head classes due to the extremely unbiased dataset, mR@K could give a fair performance appraisal for both head and tail classes, which is widely used as an unbiased evaluation metric.

Implementation Details. We adopt a pre-trained Faster R-CNN [25] with ResNeXt-101-FPN [39] provided by [30] as the object detector. We employ Glove [24] to obtain the semantic embedding. The object encoder and the relation encoder contain four and two SHA layers, respectively. We set the division threshold $\mu = 4$, and employ the Top-Down strategy (each classifier is forced to learn the prediction behavior from all its predecessors, see Figure 6 for more details) to construct the pairwise knowledge matching set $\mathcal{Q}$. The hyper-parameter α which balances the optimization objective is set to be 1.0. We optimize the proposed network by the Adam optimizer with a momentum of 0.9. For all three tasks, the total training stage lasts for 60,000 steps with a batch size of 8. The initial learning rate is 0.001, and we adopt the same warm-up and decayed strategy as [30]. One RTX2080 Ti is used to conduct all the experiments.

4.2. Compared Methods

We want to declare that our proposed method is not only powerful in generating unbiased scene graphs, but also applicable for a variety of SGG approaches. For the former, we compare it with state-of-the-art approaches, including re-produced IMP+ [40], KERN [3], GPS-Net [20], PCPL [42], re-produced VTransE+ [49] and BGNN [17]. For the latter, we adopt two typical baselines, namely Motifs [47] and VCTree [31], to give a fair comparison with other model-agnostic approaches, such as Reweighting [4], TDE [30], CogTree [43], DLFE [4] and EBM [29].

Table 1 and Table 2 present the performance of different approaches conducted on VG150 and GQA200, respectively. We have several observations as follows: 1) Our proposed SHA+GCL significantly outperforms all the baselines on all three tasks. To the best of our knowledge, our work is the first to breakthrough the 40% precision in both mR@50 and mR@100 on PredCls, and we also achieve the

Model	PredCls			SGCls			SGDet		
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
IMP+†	-	9.8	10.5	-	5.8	6.0	-	3.8	4.8
KERN†	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
GPS-Net†	17.4	21.3	22.8	10.0	11.8	12.6	6.9	8.7	9.8
PCPL†	-	35.2	37.8	-	18.6	19.6	-	9.5	11.7
VTransE+	13.6	17.1	18.6	6.6	8.2	8.7	5.1	6.8	8.0
SG-CogTree	22.9	28.4	31.0	13.0	15.7	16.7	7.9	11.1	12.7
BGNN	-	30.4	32.9	-	14.3	16.5	-	10.7	12.6
Motifs	11.7	14.8	16.1	6.7	8.3	8.8	5.0	6.8	7.9
Motifs + Reweight$_d$	14.3	17.3	18.6	9.5	11.2	11.7	6.7	9.2	10.9
Motifs + TDE$_d$	18.5	25.5	29.1	9.8	13.1	14.9	5.8	8.2	9.8
Motifs + CogTree$_d$	20.9	26.4	29.0	12.1	14.9	16.1	7.9	10.4	11.8
Motifs + DLFE$_d$	22.1	26.9	28.8	12.8	15.2	15.9	8.6	11.7	13.8
Motifs + EBM$_d$	14.2	18.0	28.8	8.2	10.2	11.0	5.7	7.7	9.3
Motifs + GCL	**30.5**	**36.1**	**38.2**	**18.0**	**20.8**	**21.8**	**12.9**	**16.8**	**19.3**
VCTree	13.1	16.7	18.1	9.6	11.8	12.5	5.4	7.4	8.7
VCTree + Reweight$_d$	16.3	19.4	20.4	10.6	12.5	13.1	6.6	8.7	10.1
VCTree + TDE$_d$	18.4	25.4	28.7	8.9	12.2	14.0	6.9	9.3	11.1
VCTree + CogTree$_d$	22.0	27.6	29.7	15.4	18.8	19.9	7.8	10.4	12.1
VCTree + DLFE$_d$	20.8	25.3	27.1	15.8	18.9	20.0	8.6	11.8	13.8
VCTree + EBM$_d$	14.2	18.2	19.7	10.4	12.5	13.5	5.7	7.7	9.1
VCTree + GCL	**31.4**	**37.1**	**39.1**	**19.5**	**22.5**	**23.5**	**11.9**	**15.2**	**17.5**
SHA	14.4	18.8	20.5	8.7	10.9	11.6	5.7	7.8	9.1
SHA + GCL (ours)	**35.6**	**41.6**	**44.0**	**19.6**	**23.0**	**24.3**	**14.2**	**17.9**	**20.9**

Table 1. Performance comparison of different methods on PredCls, SGCls, and SGDet tasks of VG150 with respect to mR@20/50/100 (%). The superscript † denotes that the method employs Faster R-CNN with VGG-16 as the object detector, while the subscript d denotes that the method is model-agnostic and targets to address the biased relationship predictions in SGG.

Model	PredCls	SGCls	SGDet
	mR 50/100	mR 50/100	mR 50/100
VTransE	14.0 / 15.0	8.1 / 8.7	5.8 / 6.6
VTransE + GCL	**30.4 / 32.3**	**16.6 / 17.4**	**14.7 / 16.4**
Motifs	16.4 / 17.1	8.2 / 8.6	6.4 / 7.7
Motifs + GCL	**36.7 / 38.1**	**17.3 / 18.1**	**16.8 / 18.8**
VCTree	16.6 / 17.4	7.9 / 8.3	6.5 / 7.4
VCTree + GCL	**35.4 / 36.7**	**17.3 / 18.0**	**15.6 / 17.8**
SHA	19.5 / 21.1	8.5 / 9.0	6.6 / 7.8
SHA + GCL	**41.0 / 42.7**	**20.6 / 21.3**	**17.8 / 20.1**

Table 2. Performance comparison of different methods on three tasks of GQA200 with respect to mR@50/100 (%).

best performance on SGCls and SGDet. 2) Motifs+GCL and VCTree+GCL nearly double the performance in mean Recall on all three tasks compared with Motifs and VCTree. It demonstrates that the proposed GCL is model-agnostic and can largely enhance the unbiased relationship predictions. 3) Compared with Motifs+GCL and VCTree+GCL, we witness an obvious performance gain in SHA+GCL. It indicates that the proposed SHA module could facilitate both the intra-modal refinement and the inter-modal interaction, thus leading to more accurate predictions. In conclusion, SHA+GCL effectively addresses two aforementioned concerns in SGG, *i.e.*, insufficient modality fusion and biased relationship predictions.

4.3. Ablation Study

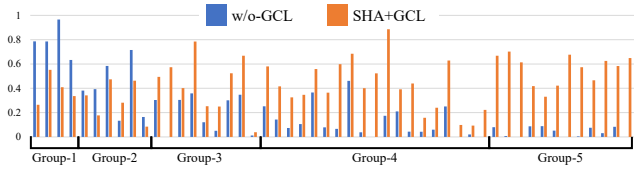
As aforementioned, we propose the Stacked Hybrid Attention (SHA) network to improve the object encoder and the relation encoder, and propose the Group Collaborative Learning (GCL) strategy, which employs the Parallel Classifier Optimization (PCO) as the "weak constraint" and Collaborative Knowledge Distillation (CKD) as the "strong constraint", to guide the training of the decoder. In order to prove the effectiveness of the above components, we test various ablation models on VG150 as follows:

- w/o-GCL: To evaluate the effectiveness of GCL, we let the relation decoder be a one-layer classifier, where a regular Cross-Entropy loss is performed.
- w/o PCO&CKD: To evaluate the effectiveness of PCO in GCL, we remove the PCO loss and CKD loss, and only employ the Median Re-Sampling strategy and a regular Cross-Entropy loss in the optimization step.
- w/o CKD: To evaluate the effectiveness of CKD in GCL, we remove the CKD loss but retain all the classifiers to compute the PCO loss.
- w/o CA or w/o SA: To evaluate the effectiveness of SHA, we remove either the Cross-Attention (CA) unit or the Self-Attention (SA) unit in every SHA layer.

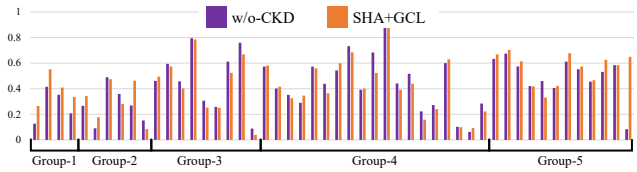
Table 3 presents the results of all the ablation models. We have several observations as follows: 1) Compared with w/o-GCL, SHA+GCL nearly doubles the per-

Model	PredCls	SGCls	SGDet
	mR 50/100	mR 50/100	mR 50/100
w/o - GCL	18.8 / 20.5	10.9 / 11.6	7.8 / 9.1
w/o - PCO&CKD	35.2 / 37.4	20.1 / 21.2	14.6 / 16.9
w/o - CKD	39.3 / 41.7	22.0 / 23.2	16.5 / 19.0
w/o - CA	39.8 / 42.5	22.6 / 23.6	16.8 / 19.3
w/o - SA	39.2 / 41.5	22.6 / 23.7	17.5 / 20.1
SHA + GCL	**41.6 / 44.0**	**23.0 / 24.3**	**17.9 / 20.9**

Table 3. Ablation study of the proposed method on VG150.



(a) R@100 of all the predicate classes of w/o-GCL and SHA+GCL on VG150.



(b) R@100 of all the predicate classes of w/o-CKD and SHA+GCL on VG150.

Figure 5. R@100 of 50 predicate classes on PredCls on VG150.

formance. Moreover, in Figure 5a, we compare w/o-GCL and SHA+GCL with respect to R@100 of all the predicate classes. As can be observed, SHA+GCL obviously improves the performance on most of the predicate classes, only with an acceptable decay on the head classes in Group 1 and Group 2, showing a powerful capability in generating unbiased scene graphs. 2) Compared with w/o-PCO&CKD, w/o-CKD evidently improves the prediction performance, demonstrating that the "weak constraint", namely gathering gradients from all the classifiers, would facilitate the convergence of the final classifier. 3) Compared with w/o-CKD, we witness an obvious performance gain in SHA+GCL. Moreover, we compare w/o-CKD and SHA+GCL on the detailed precision towards every predicate class on VG150. As shown in Figure 5b, CKD effectively prevents the model from sacrificing much on the head classes, as well as achieves a comparable performance towards the tail predictions. It demonstrates that the "strong constraint", namely a knowledge transfer paradigm, could effectively compensate for the under-fitting on the head classes by preserving the discriminating capability learned previously, and thus benefits in achieving a reasonable trade-off. 4) From the last three rows in Table 3, we witness an obvious performance decay when removing either the CA unit or the SA unit. It verifies that combining both attentions would effectively alleviate the insufficient modality fusion, thus leading to more accurate predictions.

Model		PredCls	SGCls	SGDet
μ	Strategy	mR 50/100	mR 50/100	mR 50/100
3	Adjacent	40.0 / 42.4	22.5 / 23.4	16.8 / 19.2
4	Adjacent	41.0 / 43.5	23.0 / 23.9	17.3 / 19.7
5	Adjacent	39.4 / 41.7	21.8 / 23.0	16.7 / 19.1
3	Top-down	40.9 / 43.2	22.9 / 23.8	17.0 / 19.9
4	Top-down	**41.6 / 44.0**	**23.0 / 24.3**	**17.9 / 20.9**
5	Top-down	39.7 / 42.0	23.1 / 23.8	16.9 / 19.6

Table 4. Parameter analysis towards the threshold μ and the pairwise knowledge matching strategies of GCL on VG150.

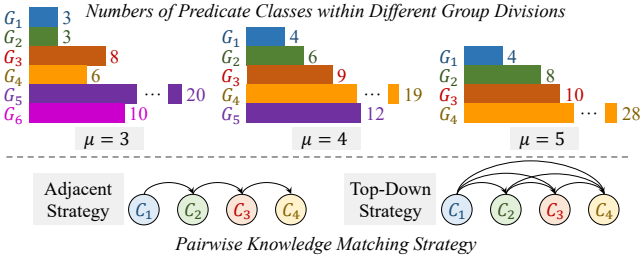


Figure 6. Illustration of three configurations of the balanced group divisions according to the threshold μ (top), and two alternatives of the pairwise knowledge matching strategy (down).

4.4. Parameter Analysis

As aforementioned, the threshold μ and the organization strategy would influence the performance of GCL. As Figure 6 illustrates, for the former, we set μ =3, 4, and 5, and obtain 6, 5, and 4 group divisions, respectively. For the latter, we provide two alternatives, namely Adjacent and Top-Down strategy, whose difference is whether each classifier could learn the knowledge from its nearest predecessor (Adjacent) or from all the predecessors (Top-Down).

Table 4 presents the performance comparisons, where $\mu = 4$ and the Top-Down strategy is the best combination.

5. Conclusion

In this work, we declare two concerns that restrict the practical applications of SGG, namely insufficient modality fusion and biased relationship predictions. To address such deficiency, we propose the Stacked Hybrid-Attention network and the Group Collaborative Learning strategy. In this way, we establish a new state-of-the-art in the unbiased metric and provide a model-agnostic debiasing method. In the future, we plan to explore more robust group dividing methods and devise more knowledge distillation strategies.

Acknowledgments. This work is supported by the National Natural Science Foundation of China, No.: 62176137, No.:U1936203, and No.: 62006140; the Shandong Provincial Natural Science and Foundation, No.: ZR2020QF106; Beijing Academy of Artificial Intelligence(BAAI); Ant Group.

References

[1] Haoli Bai, Jiaxiang Wu, Irwin King, and Michael Lyu. Few shot network compression via cross distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3203–3210, 2020. 3

[2] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9962–9971, 2020. 1

[3] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019. 2, 3, 6

[4] Meng-Jiun Chiou, Henghui Ding, Hanshu Yan, Changhu Wang, Roger Zimmermann, and Jiashi Feng. Recovering the unbiased scene graphs from the biased ones. *arXiv preprint arXiv:2107.02112*, 2021. 2, 6

[5] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3076–3086, 2017. 2

[6] Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D Hager, Federico Tombari, and Christian Rupprecht. Semantic image manipulation using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5213–5222, 2020. 1

[7] Lingyun Feng, Minghui Qiu, Yaliang Li, Hai-Tao Zheng, and Ying Shen. Learning to augment for data-scarce domain bert knowledge distillation. *arXiv preprint arXiv:2101.08106*, 2021. 3

[8] Mitchell A Gordon and Kevin Duh. Explaining sequence-level knowledge distillation as data-augmentation for neural machine translation. *arXiv preprint arXiv:1912.03334*, 2019. 3

[9] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021. 3

[10] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332, 2019. 1

[11] Yutian Guo, Jingjing Chen, Hao Zhang, and Yu-Gang Jiang. Visual relations augmented cross-modal retrieval. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 9–15, 2020. 1

[12] Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*, 2020. 1

[13] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3

[14] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14045–14054, 2020. 2, 5

[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

[16] 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. *arXiv preprint arXiv:1602.07332*, 2016. 6

[17] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. 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*, pages 11109–11119, 2021. 1, 2, 6

[18] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. Visual relationship detection with deep structural ranking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2

[19] Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. Natural language guided visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 444–453, 2019. 2

[20] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3746–3753, 2020. 2, 6

[21] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision*, pages 852–869, 2016. 1, 2

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 3

[23] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020. 3

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014. 6

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6):1137–1149, 2017. 6

[26] Sahand Sharifzadeh, Sina Moayed Baharlou, and Volker Tresp. Classification by attention: Scene graph classification

with prior knowledge. *arXiv preprint arXiv:2011.10084*, 2020. 1, 2

[27] Zhiqiang Shen, Zechun Liu, Dejia Xu, Zitian Chen, Kwang-Ting Cheng, and Marios Savvides. Is label smoothing truly incompatible with knowledge distillation: An empirical study. *arXiv preprint arXiv:2104.00676*, 2021. 3

[28] Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia*, 2021. 1

[29] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13936–13945, 2021. 2, 6

[30] 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. 1, 2, 6

[31] 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. 1, 2, 3, 6

[32] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2017. 1

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 4

[34] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip. Private model compression via knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1190–1197, 2019. 3

[35] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing*, 29:1–14, 2019. 2

[36] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1437–1445, 2019. 2

[37] Bin Wen, Jie Luo, Xianglong Liu, and Lei Huang. Unbiased scene graph generation via rich and fair semantic extraction. *arXiv preprint arXiv:2002.00176*, 2020. 2

[38] Olivia Wiles, Sebastien Ehrhardt, and Andrew Zisserman. Co-attention for conditioned image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15920–15929, 2021. 2

[39] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 6

[40] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017. 2, 6

[41] Pengfei Xu, Xiaojun Chang, Ling Guo, Po-Yao Huang, Xiaojiang Chen, and Alexander G Hauptmann. A survey of scene graph: Generation and application. *IEEE Trans. Neural Netw. Learn. Syst*, 2020. 1

[42] Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. Pcpl: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 265–273, 2020. 1, 2, 3, 6

[43] Jing Yu, Yuan Chai, Yue Hu, and Qi Wu. Cogtree: Cognition tree loss for unbiased scene graph generation. *arXiv preprint arXiv:2009.07526*, 2020. 1, 2, 3, 6

[44] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 2

[45] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3903–3911, 2020. 3

[46] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *Proceedings of the European Conference on Computer Vision*, pages 606–623, 2020. 1, 2

[47] 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, 3, 4, 6

[48] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. *arXiv preprint arXiv:1907.12133*, 2019. 1

[49] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5532–5540, 2017. 6

[50] Lecheng Zheng, Yu Cheng, Hongxia Yang, Nan Cao, and Jingrui He. Deep co-attention network for multi-view subspace learning. In *Proceedings of the Web Conference 2021*, pages 1528–1539, 2021. 2

[51] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *Proceedings of the European Conference on Computer Vision*, pages 211–229, 2020. 1