Multi-Granularity Sparse Relationship Matrix Prediction Network for End-to-End Scene Graph Generation

Lei Wang, Zejian Yuan, and Badong Chen

Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
leiwangmail@stu.xjtu.edu.cn, {yuan.ze.jian, chenbd}@mail.xjtu.edu.cn

Abstract. Current end-to-end Scene Graph Generation (SGG) relies solely on visual representations to separately detect sparse relations and entities in an image. This leads to the issue where the predictions of entities do not contribute to the prediction of relations, necessitating post-processing to assign corresponding subjects and objects to the predicted relations. In this paper, we introduce a sparse relationship matrix that bridges entity detection and relation detection. Our approach not only eliminates the need for relation matching, but also leverages the semantics and positional information of predicted entities to enhance relation prediction. Specifically, a multi-granularity sparse relationship matrix prediction network is proposed, which utilizes three gated pooling modules focusing on filtering negative samples at different granularities, thereby obtaining a sparse relationship matrix containing entity pairs most likely to form relations. Finally, a set of sparse, most probable subject-object pairs can be constructed and used for relation decoding. Experimental results on multiple datasets demonstrate that our method achieves a new state-of-the-art overall performance. Our code is available at https://github.com/wanglei0618/Mg-RMPN.

Keywords: Scene Graph Generation · End-to-End · Sparse Relationship Matrix · Multi-Granularity

1 Introduction

Scene Graph Generation (SGG) is a fundamental visual comprehension task that captures semantic information by detecting relation triplets <subject entity, predicate, object entity> in an image. This structured representation can facilitate many downstream tasks, such as image captioning [4, 41], visual question answering [9, 28], image retrieval [11, 25] and image generation [10, 19]. The end-to-end SGG methods [6, 16, 30] can directly generate sparse relations from an image based on a fixed number of relation queries, avoiding dense prediction that contains a lot of background relations in two-stage SGG. This leads to faster inference speed, and because fewer background relations are predicted, it exhibits superior predictive performance for rare samples of tail classes.

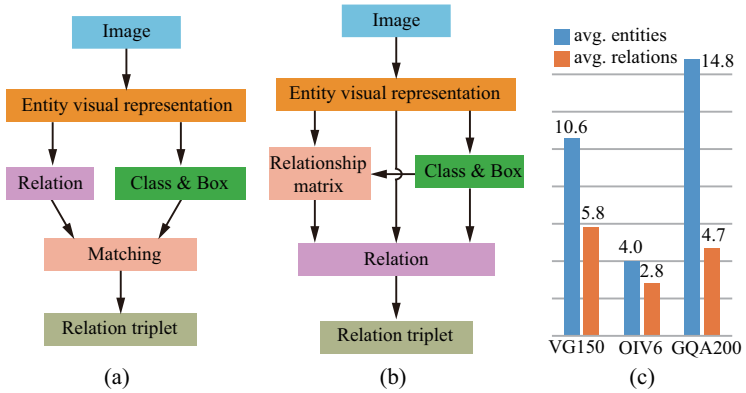

Fig. 1: (a) End-to-end SGG pipeline. (b) Our proposed end-to-end pipeline based on multi-granularity sparse relationship matrix prediction network. (c) The average number of entities and relations in the VG150, OIv6, and GQA200 datasets demonstrates the sparsity of relations in an image.

Current end-to-end approaches [6,16,22] still lag behind two-stage methods in performance, which is attributed to their reliance solely on visual representations to separately detect entities and relations, as shown in Figure 1 (a). This leads to two drawbacks: (1) It requires post-processing to match the corresponding subjects and objects for the relations, and the model's performance is constrained by the capabilities of matching. (2) Due to the inability to determine the subject and object in advance, entity predictions cannot help with relation prediction, and relation prediction can only rely on visual features.

In this paper, we introduce a sparse relationship matrix that bridges entity detection and relation detection. Our approach not only eliminates the need for relation matching, but also leverages the semantics and positional information of predicted entities to enhance relation prediction. Figure 1 (b) shows our pipeline based on the "pair then relation" [30] framework, which first constructs a relationship matrix based on entity predictions to represent the possibility of forming a relation between two entities and then selects the sparsest, most likely subject-object pairs for relation prediction.

Compared to all possible pairs formed by the entity proposals, the subject-object pairs with relations are extremely sparse. Figure 1 (c) illustrates the average entities and relations in the datasets. Taking VG150 [13] as an example, when generating $N_e = 100$ entities, there are a total of $(100(100-1)) = 9900$ possible combinations. But the average relations per image is only 5.8, which means that more than 99. 9% pairs are negative samples. Therefore, directly learning this extremely sparse relationship matrix is very challenging.

In this work, we propose a Multi-granularity sparse Relationship Matrix Prediction Network (Mg-RMPN) to partition the negative samples into different granularities, achieving layer-by-layer filtering of dense negative samples. Specifically, Mg-RMPN employs three Gated Pooling Modules (GPM) with identical

structures but independent parameters, each of which focuses on filtering negative samples at different granularities. The GPMs are used to generate relation matrices at different granularities and construct the corresponding ground truth matrices for supervised learning. The experimental results from various datasets reveal that our approach not only accurately identifies numerous head-class samples, but also excels in predicting sparse tail-class samples, achieving state-of-the-art overall performance.

The contributions are summarized as follows: (1) An end-to-end SGG framework based on the relationship matrix is proposed that bridges entity detection and relation detection. (2) An Mg-RMPN is proposed to achieve sparse relationship matrix prediction based on multi-granularity negative sample learning. (3) Experimental results have confirmed that our method achieves state-of-the-art overall performance in both end-to-end and two-stage methods.

2 Related Works

2.1 Scene Graph Generation

Early work [5, 27, 32, 36] in SGG is based on the two-stage pipeline, first employing a Faster-RCNN [24] object detector to generate entity proposals, and the class of each entity is predicted. Then, entities are paired to form relations, and all possible pairs are classified. Its drawback lies that dense relation pairs include numerous background relations, which dilute the sparse tail-class samples, thus aggravating the imbalance issue. Therefore, some recent efforts [8, 23] have focused on addressing the imbalanced predictions, developing a series of rebalancing strategies such as logit adjustment [2, 26], loss reweighting [18, 33, 35], and data resampling [17, 31, 37]. Unlike these methods, we introduce a relationship matrix to generate sparse subject-object pairs for relation prediction. The sparse relations generated in this way have a higher signal-to-noise ratio, which can alleviate the imbalance problem. The results of the experiments validate that our method maintains outstanding performance for head classes and also demonstrates high performance for tail classes.

2.2 End-to-End Scene Graph Generation

End-to-end SGG, inspired by DETR [1], generates sparse relations directly from an image based on queries. For instance, [6] introduced the Relation Transformer to directly generate a set of relations from visual features. However, the generated relations did not match the corresponding subjects and objects, hence [16] proposed a graph assembling module to address the relation matching. [30] further introduced a "pair then relation" framework that predetermines subject-object pairs, circumventing the relation matching. However, these approaches suffer from the issue of relying solely on visual representations to detect relations and entities separately. Therefore, this paper introduces Mg-RMPN to predict a sparse relationship matrix to link entity detection and relation detection, not only eliminating the need for relation matching but also utilizing the predicted entity semantics and positional information to assist in relation prediction.

3 Method

3.1 Problem Definition

SGG aims to generate a scene graph $\mathcal{G} = \{\mathcal{E}, \mathcal{R}\}$ from an input image $\mathcal{I}$. The scene graph consists of a set of n entities $\mathcal{E} = \{e_i\}_{i=1}^{n}$ and a set of m relations $\mathcal{R} = \{r_k\}_{k=1}^{m}$ between entities. The set of entities $\mathcal{E}$ can be further decomposed into a set of bounding boxes $\mathcal{B} = \{\mathbf{b}_i\}_{i=1}^{n}$ and a set of class labels $\mathcal{C} = \{c_i\}_{i=1}^{n}$ The generation of a scene graph $\mathcal{G}$ can be formulated as the joint probability distribution:

$$\Pr(\mathcal{G}|\mathcal{I}) = \Pr(\mathcal{B}, \mathcal{C}|\mathcal{I}) \Pr(\mathcal{R}|\mathcal{I}, \mathcal{B}, \mathcal{C}), \tag{1}$$

where $\Pr(\mathcal{B}, \mathcal{C}|\mathcal{I})$ represents the entity representation obtained from the object detector, including the class labels and bounding boxes. $\Pr(\mathcal{R}|\mathcal{I}, \mathcal{B}, \mathcal{C})$ represents the relation prediction based on pairs of entities by the relation decoder. The process from $\Pr(\mathcal{B}, \mathcal{C}|\mathcal{I})$ to $\Pr(\mathcal{R}|\mathcal{I}, \mathcal{B}, \mathcal{C})$ requires our Mg-RMPN to predict the pairs of entities most likely to form a relation from all possible pairs of entities.

3.2 Overall Architecture

As shown in Figure 2 (a), our method comprises three modules: (1) entity detection, (2) relationship matrix prediction, and (3) relation prediction. The entity detection is responsible for generating a set of entities based on the object detector, including their visual representations, class labels, and bounding boxes. The relationship matrix prediction is responsible for producing a sparse relationship matrix that represents the relevance between two entities, and then obtaining the indices of the subject and object entities that are most likely to form a relation from all possible combinations. Finally, the relation of these sparse subject-object pairs is predicted through the relation decoder.

3.3 Object Detector

This paper employs a deformable DETR [42] as the object detector, which contains a transformer encoder-decoder architecture with a set of entity queries. Here, the entity queries interact with encoder features through multi-scale deformable attention. Given an image $\mathcal{I}$, the object detector produces N_e entity representations $\mathbf{Q}_e = \{\mathbf{q}_i^e\}_{i=1}^{N_e} \in \mathbb{R}^{N_e \times d}$ from a set of learnable entity queries, where d is the embedding dimensions. Based on $\mathbf{Q}_e$, the object detector then follows with a classification head and a regression head to predict the entity's class predictions $\mathbf{C} = \{\hat{\mathbf{c}}_i\}_{i=1}^{n} \in \mathbb{R}^{N_e \times C_e}$ and bounding boxes $\mathbf{B} = \{\hat{\mathbf{b}}_i\}_{i=1}^{n} \in \mathbb{R}^{N_e \times 4}$, respectively, where C_e is the number of entity classes.

3.4 Multi-Granularity Relationship Matrix Prediction Network

Given N_e entities, Multi-granularity sparse Relationship Matrix Prediction Network (Mg-RMPN) generates a $N_e \times N_e$ adjacency relationship matrix $\mathbf{M}_r$. Each

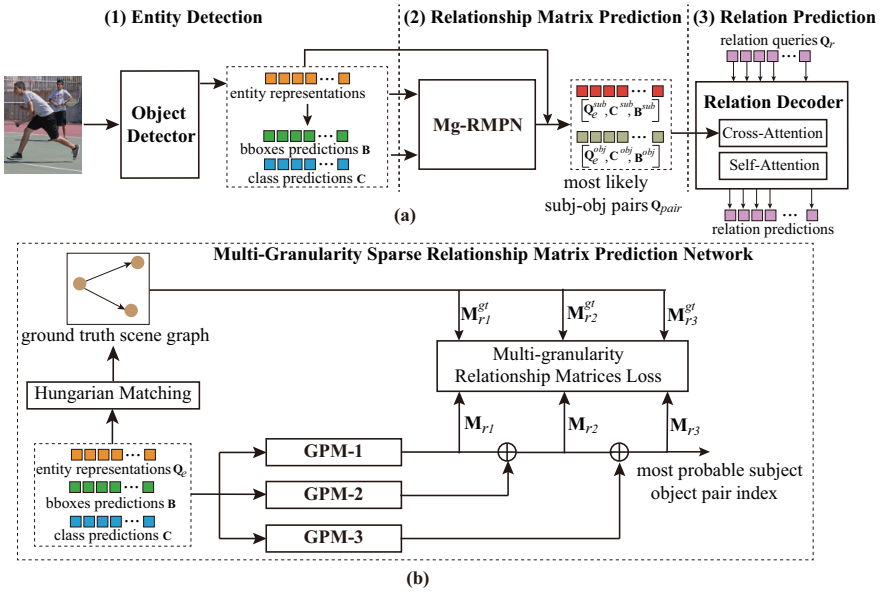

Fig. 2: (a) Overall pipeline of our end-to-end SGG method. (b)Multi-granularity sparse Relationship Matrix Prediction Network (Mg-RMPN). GPM is Gated Pooling Module.

node represents the probability of forming a relationship between two entities. Due to the sparse entities and relations in an image, $\mathbf{M}_r$ contains a large number of pairs without relations, that is, negative samples.

Therefore, Mg-RMPN defined three $\mathbf{M}_r$ based on multi-granularity negative samples, where the three $\mathbf{M}_r$ are learned through three structurally identical but parameter-independent Gated Pooling Modules (GPM), as shown in Figure 2 (b). With the outputs of the GPMs, Mg-RMPN is able to generate the final sparse relationship matrix, enabling the identification of most probable subject-object pairs to predict relations.

Gated Pooling Module Unlike [30], which only uses visual features, our GPM determines the correlation of two entities based on their visual representation, linguistic prior and location information. Given a set of visual representations $\mathbf{Q}_e$ of the entities in an image, GPM first uses two fully connected networks $\mathcal{F}_{sub}$ and $\mathcal{F}_{obj}$ to project it to the visual representations of the subject and object, $\mathbf{Q}_{sub}$ and $\mathbf{Q}_{obj}$, respectively, as follows:

$$\mathbf{Q}_{sub} = \mathcal{F}_{sub}(\mathbf{Q}_e) = \{\mathbf{q}_i^{sub}\}_{i=1}^{N_e}, \quad \mathbf{Q}_{obj} = \mathcal{F}_{obj}(\mathbf{Q}_e) = \{\mathbf{q}_j^{obj}\}_{j=1}^{N_e}. \tag{2}$$

Subsequently, the visual similarity between the subject i and the object j can be calculated as $v_{ij}^{sim} = \mathbf{q}_i^{sub} \cdot (\mathbf{q}_j^{obj})^{\mathrm{T}}$.

Given a set of class prediction distributions $\mathbf{C}$ for entities, GPM uses the prior linguistic function $\mathcal{F}_{prior}$ from [26] to predict the correlation between two

entities, $s_{ij}^{prior} = \mathcal{F}_{prior}(\hat{\mathbf{c}}_i, \hat{\mathbf{c}}_j)$. $\mathcal{F}_{prior}$ aims to capture the linguistic inductive bias that exists between two entities to enhance the prediction of the relational matrices. Similarly to [26], $\mathcal{F}_{prior}$ obtains a reasonable initialization based on the distribution of relations generated by entity pairs in the training dataset.

Based on the positions $\mathbf{B}$ of the entities, GPM uses the function $\mathcal{F}_{overlap}$ to obtain the relative position information between two entities $\mathbf{b}_{ij}^{overlap} = \mathcal{F}_{overlap}(\hat{\mathbf{b}}_i, \hat{\mathbf{b}}_j) \in \mathbb{R}^8$, where 8 elements indicate difference in x and y coordinates, width ratio, height ratio, IOU, center distance, area ratio, and aspect ratio. Additionally, we can also obtain the union bounding box $\hat{\mathbf{b}}_{ij}^{u} \in \mathbb{R}^9$ for the entity pairs, where 9 elements indicate the union bounding box coordinates (x_1, y_1, x_2, y_2), center $(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$, size $(x_2 - x_1, y_2 - y_1)$ and area $(x_2 - x_1)(y_2 - y_1)$, respectively. Then, we can obtain the relevance representation of the subject i and object j as follows:

$$\mathbf{e}_{ij} = [v_{ij}^{sim}, s_{ij}^{prior}, \mathbf{b}_{ij}^{overlap}, \hat{\mathbf{b}}_{ij}^{u}], \tag{3}$$

where $[\cdot, \cdot]$ denotes the concatenation operation. Subsequently, the feature representation of the entity pair can be expressed as

$$\mathbf{f}_{ij} = [\mathbf{q}_i^{sub}, \mathbf{q}_j^{obj}, f_s(s_{ij}^{prior}), pos(\mathbf{b}_{ij}^{overlap}), pos(\hat{\mathbf{b}}_{ij}^{u})], \tag{4}$$

Where f_s is a fully-connected layer for language prior encoding, and pos is a fully-connected layer for positional encoding. Then, GPM uses a gated pooling layer f_{gate} to obtain the probability of relation between subject i and object j as follows:

$$\hat{p}_{ij} = \text{Sigmoid}(\sum f_{gate}(\mathbf{f}_{ij}) \odot \mathbf{e}_{ij}), \tag{5}$$

where $\odot$ represents element-wise multiplication. Finally, we can obtain a $N_e \times N_e$ relationship matrix $\mathbf{M}_r = \text{GPM}(\mathbf{Q}_e, \mathbf{C}, \mathbf{B})$. Note that we set the diagonal elements of $\mathbf{M}_r$ to zero, removing the cases where an entity serves as both subject and object simultaneously.

Multi-Granularity Sparse Relationship Matrix Prediction Based on the output of three GPMs, we can construct three relationship matrices based on the granularity of negative samples and build the corresponding ground truth matrices (see Section 3.6) for supervised learning, enabling each GPM to focus on identifying negative samples of specific granularity.

Since the N_e entity proposals are typically much greater than the number of entities contained in an image, resulting in a large number of background entities among all possible pairs of entities. Since background entities do not form relationships, we divide the negative samples in the relationship matrix into three levels of granularity: (1) high-confidence negative samples composed of background entities (background-background), (2) medium-confidence negative samples constructed from background and foreground entities (entity-background or background-entity), and (3) low-confidence negative samples composed of foreground entities (entity-entity).

Then we can define three relationship matrices based on the output of the GPMs as follows.

$$\begin{aligned}\mathbf{M}_{r1} &= \text{GPM-1}(\mathbf{Q}_e, \mathbf{C}, \mathbf{B}) \\ \mathbf{M}_{r2} &= \mathbf{M}_{r1} + \text{GPM-2}(\mathbf{Q}_e, \mathbf{C}, \mathbf{B}) \\ \mathbf{M}_{r3} &= \mathbf{M}_{r2} + \text{GPM-3}(\mathbf{Q}_e, \mathbf{C}, \mathbf{B})\end{aligned} \tag{6}$$

In Eq.6, each GPM can focus on learning the filtering of negative samples at different granularities, as follows: (1) GPM-1 focuses on learning from pairs between background and filters out a large number of high-confidence negative samples. (2) GPM-2 focuses on learning from pairs between background and entities and filters out medium-confidence negative samples. (3) GPM-3 focuses on learning from no relation pairs between entities and only needs to filter out entity pairs that do not form relationships from the remaining entity pairs.

Based on the final sparse relationship matrix $\mathbf{M}_{r3}$, we can select the top-N_r subject-object pairs most likely to form relations and obtain their indices in the entity visual representations $\mathbf{Q}_e$, the entity's class predictions $\mathbf{C}$, and the bounding boxes $\mathbf{B}$.

3.5 Relation Decoder

With the $\mathbf{Q}_e$, $\mathbf{C}$ and $\mathbf{B}$ of subjects and objects, we can obtain the subject-object pair representation $\mathbf{Q}_{pair}$, which integrates visual, semantic, and positional encodings as follows:

$$\mathbf{Q}_{pair} = [\mathbf{Q}_e^{sub}, emb(\mathbf{C}^{sub}), pos(\mathbf{B}^{sub}), \mathbf{Q}_e^{obj}, emb(\mathbf{C}^{obj}), pos(\mathbf{B}^{obj})]. \tag{7}$$

where emb is a pre-trained Glove language model to acquire the word embedding.

In this paper, we adopt the relation decoder in [30], which consists of transformer decoders in the style of DETR, to predict relations. In the relation decoder, we initialize a relation query $\mathbf{Q}_r \in \mathbb{R}^{N_r \times d}$ as the query input, and the subject-object pair $\mathbf{Q}_{pair}$ is projected as the key and value of cross-attention. Subsequently, the relation representation after the Relation Decoder (RD) can be expressed as $\tilde{\mathbf{Q}}_r = \text{RD}(\mathbf{Q}_r, \mathbf{Q}_{pair})$. Then, the relation prediction for the entity pairs can be expressed as $\hat{\mathbf{r}} = W_{cls}\tilde{\mathbf{Q}}_r$, where W_{cls} is the liner relation classifier.

3.6 Training

Our end-to-end SGG is divided into three subtasks: (1) an entity detection task based on the object detector, (2) a sparse relationship matrix prediction task based on the Mg-RMPN, and (3) a relation prediction task based on the relation decoder. During training, each subtask generates the corresponding supervisory information and losses, as follows.

Entity Detection Loss We uses the end-to-end deformable DETR [42] as object detector, which uses the set prediction loss proposed in DETR [1] by assigning the ground truth entities to the predictions. A cost function is applied to compute the matching cost between a prediction and a ground-truth entity. With the cost matrix, the entity prediction-ground truth assignment is computed with the Hungarian Matching [14]. Giver a set of N_e entity proposals $\{e_i\}_{i=1}^{N_e}$ from object detector, the set prediction loss for can be presented as:

$$\mathcal{L}_e = \Sigma_{i=1}^{N_e}[\mathcal{L}_{cls}^e + \mathbf{1}_{e_i \neq \phi}\mathcal{L}_{box}^e], \tag{8}$$

where $\mathcal{L}_{cls}^e$ denotes the cross-entropy loss for label classification and $e_i \neq \phi$ means that <background> is not assigned to the ith entity prediction. $\mathcal{L}_{box}^e$ consists of L_1 loss and generalized IoU loss for the box regression.

Multi-granularity Relationship Matrices Loss To predict the final sparse relationship matrix, we need to define the supervisory information and loss functions during the training phase.

Multi-granularity Supervision Matrices In the training phase, we assign labels to the entity proposals $\{e_i\}_{i=1}^{N_e}$ by Hungarian matching and obtain the ground truth for each node in $\mathbf{M}_r$ regarding the subject, object, and whether they form a relation. In response to the sparsity of $\mathbf{M}_r$, we propose Mg-RMPN to achieve hierarchical identification of negative samples at different granularities.

Consequently, based on the granularity of the negative samples that each GPM focuses on in Section 3.4, the corresponding ground truth can be obtained. (1) $\mathbf{M}_{r1}^{gt}$ filters out all pairs composed of background entities as follows:

$$\mathbf{M}_{r1}^{gt} = \{p_{ij} \mid p_{ij} = \begin{cases} 0 & \text{if } (e_i = \phi \text{ and } e_j = \phi) \text{ or } i = j \\ 1 & \text{otherwise} \end{cases}, 1 \leq i \leq N_e, 1 \leq j \leq N_e\}, \tag{9}$$

where $e_i = \phi$ ($e_j = \phi$) means that <background> is assigned to the ith (jth) entity prediction, $i = j$ means removing self-connected entity pairs. (2) $\mathbf{M}_{r2}^{gt}$ further filters out all pairs composed of backgrounds and entities as follows:

$$\mathbf{M}_{r2}^{gt} = \{p_{ij} \mid p_{ij} = \begin{cases} 0 & \text{if } (e_i = \phi \text{ or } e_j = \phi) \text{ or } i = j \\ 1 & \text{otherwise} \end{cases}, 1 \leq i \leq N_e, 1 \leq j \leq N_e\}. \tag{10}$$

(3) $\mathbf{M}_{r3}^{gt}$ filters out all non-relationship pairs, and $\mathbf{M}_{r3}^{gt}$ is also the final ground truth relationship matrix of the scene graph, expressed as follows

$$\mathbf{M}_{r3}^{gt} = \{p_{ij} \mid p_{ij} = \begin{cases} 0 & \text{otherwise} \\ 1 & \text{if } r_{ij} = 1 \end{cases}, 1 \leq i \leq N_e, 1 \leq j \leq N_e\}, \tag{11}$$

where $r_{ij} = 1$ implies that there is a relation between subject i and object j. Here, the relationships of the negative sample subsets in $\mathbf{M}_{r1}^{gt}$, $\mathbf{M}_{r2}^{gt}$ and $\mathbf{M}_{r3}^{gt}$ are $\mathbf{M}_{r1|p_{ij}=0}^{gt} \subseteq \mathbf{M}_{r2|p_{ij}=0}^{gt} \subseteq \mathbf{M}_{r3|p_{ij}=0}^{gt}$.

Multi-granularity Relationship Matrices Loss Considering that the large number of negative samples in $\mathbf{M}_r$ have different confidences, we modify the binary cross-entropy function using focal loss [20], which applies weighted discrimination on well-classified samples, forcing the model to focus on wrongly classified samples. Then, the multi-granularity learning loss is

$$\mathcal{L}_{Mg} = \mathcal{L}^1_{RM}(\mathbf{M}_{r1}, \mathbf{M}^{gt}_{r1}) + \mathcal{L}^2_{RM}(\mathbf{M}_{r2}, \mathbf{M}^{gt}_{r2}) + \mathcal{L}^3_{RM}(\mathbf{M}_{r3}, \mathbf{M}^{gt}_{r3}), \tag{12}$$

where the relationship matrices loss $\mathcal{L}^k_{RM}(k = 1, 2, 3)$ is defined as

$$\begin{aligned} \mathcal{L}^k_{RM}(\mathbf{M}_r, \mathbf{M}^{gt}_r) = -\sum_i \sum_j \{&\alpha_k (1 - \hat{p}_{ij})^\gamma p_{ij} \log(\hat{p}_{ij}) \\ &+ (1 - \alpha_k) \hat{p}^\gamma_{ij} (1 - p_{ij}) \log(1 - \hat{p}_{ij})\} / N_{pos} \\ &+ \|\mathbf{M}_r\|_1 + \|\mathbf{M}_r\|_2, \end{aligned} \tag{13}$$

where the first part is the focal loss based on binary cross-entropy, and the second part is the L1 and L2 regularization of $\mathbf{M}_r$, $\alpha_k(k = 1, 2, 3)$ is hyperparameters that adjusts the weights of positive and negative samples, γ is a hyperparameter for focal loss, and N_{pos} is the number of positive samples in each mini-batch.

Relation Prediction Loss In this paper, the relation decoder also considers relation prediction as a set prediction based on a query. Therefore, during the training phase, we also use Hungarian Matching to assign labels for relation prediction. Due to the imbalanced relation classes in SGG, we use Seesaw loss [29] to dynamically adjust the gradients of samples of different classes. The relation loss is as follows:

$$\mathcal{L}_r = -\sum_{i=1}^{C_r} y_i \log(\hat{\sigma}_i), \quad \hat{\sigma}_i = \frac{e^{\hat{r}_i}}{\sum_{i \neq j}^{C_r} \mathcal{S}_{ij} e^{\hat{r}_j} + e^{\hat{r}_i}}, \tag{14}$$

where C_r is the number of relation classes, $y_i \in \{0, 1\}, 1 \leqslant i \leqslant C_r$ is the one-hot ground truth label, $\mathcal{S}_{ij}$ is a tunable balancing factor between different classes, detail can be found in [29].

In summary, the total loss of our method is

$$\mathcal{L} = \mathcal{L}_e + \lambda_1 \mathcal{L}_{Mg} + \lambda_2 \mathcal{L}_r, \tag{15}$$

Where λ_1 and λ_2 are hyperparameters used to respectively adjust multi-granularity learning loss $\mathcal{L}_{Mg}$ and relation prediction loss $\mathcal{L}_r$.

4 Experiments

4.1 Experimental Settings

Datasets: We evaluate our proposed method on the following datasets: (1) **Visual Genome (VG150)** [13] is the most widely used dataset in SGG, consisting of the most frequent 150 object classes and 50 predicate classes. (2) **Open Images V6 (OIv6)** [15] is a large-scale dataset proposed by Google. We follow

the data processing and evaluation protocols in [17, 40]. OIv6 consists of 601 object classes and 30 predicate classes. (2) **Generalized Question Answering (GQA200)** [9] is another vision and language benchmark with more than 3.8M relation annotations. We follow the data processing in [7], which consists of Top-200 object classes and Top-100 predicate classes.

Task & Evaluation Metrics: In this work, we focus on the Scene Graph Detection task, which detects all objects in an image and predicts their bounding boxes, labels, and relations. The initial evaluation metric Recall@K (R@K) was found to be dominated by head classes. Therefore, the mean Recall@K (mR@K) across all relations classes is proposed to evaluate unbiased SGG. However, focusing solely on mR@K and neglecting R@K can result in a tail bias. Therefore, we adopts the harmonic mean F@K of R@K and mR@K as an overall metric. For OIv6, the weighted evaluation metrics (wmAP_{rel}, wmAP_{phr}, $\text{score}_{\text{wtd}}$) are used for a more class-balanced evaluation.

Implementation Details: We utilized the pre-trained deformable DETR [42] as the object detector. We set the loss hyperparameters λ_1 and λ_2 to 0.5 and 3 respectively. For the multi-granularity relationship matrices loss, we set $\alpha_k (k = 1, 2, 3)$ to 0.75, 0.9, and 0.99, respectively, to control the weight of positive samples, with the focal loss parameter γ set to 2. The number of entities predicted by the detector N_e is 100, the number of most probable subject-object pairs selected N_r is 100, and the size of embedding dimemsions d is 256. Our model is implemented on 8 NVIDIA 3090 GPUs with learning rate $\times 10^{-4}$ and batch size 16 for 24 epochs. More implementation details are shown in the Supplementary Material.

4.2 Comparison with State of the Arts

VG150: Table 1 shows the comparison of our method on VG150, with methods divided into two-stage (above) and end-to-end (below). Due to the imbalance issue in SGG, the two-stage methods (Motifs, VCTree, GPS-Net, RelDN, PE-Net) perform better in R@K but show very poor performance in mR@K. Thus, many rebalancing techniques (VCTree+TDE, BGNN, SHA+GCL) have been developed. Although they improve mR@K, they inevitably impair R@K. For example, SHA+GCL, despite achieving the best mR@K, its recall performance significantly deteriorates, with R@50/100 being only 14.9/18.2.

Although the end-to-end methods are inferior to the two-stage methods on R@K, it demonstrates better potential on mR@K due to its advantage in sparse relation prediction, which does not dilute the scarce tail-class samples. Furthermore, compared to the rebalancing techniques of the two-stage methods, the end-to-end approach does not overly harm R@K and can achieve superior overall performance F@K, such as SGTR and Pair-Net.

Our Mg-RMPN leads comprehensively in end-to-end methods. Compared with two-stage methods, Mg-RMPN achieves competitive performance in R@K and significantly excels in mR@K, achieving the best overall performance F@K. For example, Mg-RMPN achieved 19.3/22.8 on F@50/100, surpassing the end-to-end and two-stage optimal methods Pair-Net and PE-Net, respectively.

Table 1: Performance comparison at scene graph detection task on VG150. $*$ denotes the results from [16].

Method	R@20	R@50	R@100	mR@20	mR@50	mR@100	F@20	F@50	F@100
Motifs [36]	**25.5**	**32.8**	**37.2**	5.0	6.8	7.9	8.4	11.3	13.0
VCTree [27]	24.5	31.9	36.2	5.4	7.4	8.7	8.8	12.0	14.0
VCTree+TDE [26]	14.0	19.4	23.2	6.9	9.3	11.1	9.2	12.6	15.0
GPS-Net [21]	22.3	28.9	33.2	6.9	8.7	9.8	**10.5**	13.4	15.1
BGNN [17]	-	31.0	35.8	-	10.7	12.6	-	15.9	18.6
RelDN [39]	-	31.4	35.9	-	6.0	7.3	-	10.1	12.1
SHA+GCL [7]	-	14.9	18.2	**14.2**	**17.9**	**20.9**	-	16.3	19.5
PE-Net [40]	-	30.7	35.2	-	12.4	14.5	-	**17.7**	**20.5**
FCSGG [22]	16.1	21.3	25.1	2.7	3.6	4.2	4.6	6.2	7.2
RelTR [6]	21.2	27.5	-	6.8	10.8	-	10.3	15.5	-
AS-Net* [3]	-	18.7	21.1	-	6.1	7.2	-	9.2	10.7
HOTR* [12]	-	23.5	27.7	-	9.4	12.0	-	13.4	16.7
SGTR [16]	-	24.6	28.4	-	12.0	15.2	-	16.1	19.8
Pair-Net [30]	18.8	24.9	29.3	8.9	12.4	15.4	12.1	16.6	20.2
Mg-RMPN(DETR)	21.0	27.1	31.3	9.9	13.5	16.2	13.5	18.0	21.3
Mg-RMPN(Ours)	**22.5**	**29.1**	**33.5**	**10.3**	**14.4**	**17.3**	**14.1**	**19.3**	**22.8**

We also compared with Pair-Net that is also based on relationship matrix learning. The experimental results show that our method is comprehensively superior to it, which demonstrates the effectiveness of our proposed Mg-RMPN. Furthermore, we also present the results of Mg-RMPN based on the DETR object detector to eliminate the influence of the detector. The results show that our Mg-RMPN(DETR) still achieves the best overall performance among both one-stage and two-stage methods.

Open Image V6: To validate the generalizability of Mg-RMPN, we performed experiments on Open Images V6 in Table 2 and compared with two-stage (above) and end-to-end (below) methods. Compared to classic SGG benchmarks, our method achieved the optimal results among all methods with 45.5, 77.8, 57.4 respectively in mR@50, R@50, F@50. Compared to open image benchmarks, we obtained competitive 35.5 and 36.4 in wmAP_{rel} and wmAP_{phr}, and achieved the best result 43.6 in the overall metric $\text{score}_{\text{wtd}}$.

GQA200: We also conducted experiments on the more challenging GQA200, as shown in Table 3. Compared to two-stage methods (VTransE, Motifs, VCTree, SHA+GCL), our method's overall performance F@K is only slightly inferior to SHA+GCL on F@100, which sacrificed a significant amount of R@K to enhance mR@K by using rebalancing techniques. Compared to the end-to-end method Pair-Net, which is also based on relationship matrices, our method is comprehensively superior.

Table 2: Performance comparison on OIv6. $*$ denotes the results from [16,17]. †denote results reproduced with the authors' code.

Method	mR@50	R@50	F@50	wmAP_{rel}	wmAP_{phr}	$\text{score}_{\text{wtd}}$
Motifs* [36]	32.7	71.6	44.9	29.9	31.6	38.9
VCTree* [27]	33.9	74.1	46.5	**34.2**	33.1	40.2
RelDN* [39]	34.0	73.1	46.4	32.2	33.4	40.8
G-RCNN* [34]	34.0	74.5	46.7	33.2	**34.2**	41.8
GPS-Net* [21]	35.3	74.8	48.0	32.9	34.0	41.7
BGNN* [17]	**40.5**	**75.0**	**52.6**	33.5	**34.2**	**42.1**
RelTR [6]	-	71.7	-	**37.2**	37.5	43.0
AS-Ne* [3]	35.2	55.3	43.0	25.9	27.5	32.4
HOTR* [12]	40.1	52.7	45.5	19.4	21.5	26.7
SGTR* [16]	42.6	59.9	49.8	37.0	**38.7**	42.3
Pair-Net† [30]	44.5	77.4	56.5	31.8	32.4	40.3
Mg-RMPN(Ours)	**45.5**	**77.8**	**57.4**	35.5	36.4	**43.6**

Table 3: Performance comparison at scene graph detection task on GQA200. $*$ denotes the results from [7]. †denote results reproduced with the authors' code.

Method	R@50	R@100	mR@50	mR@100	F@50	F@100
VTransE* [38]	27.2	30.7	5.8	6.6	9.6	10.9
Motifs* [36]	**28.9**	**33.1**	6.4	7.7	10.5	12.5
VCTree* [27]	28.3	31.9	6.5	7.4	10.6	12.0
SHA+GCL* [7]	14.8	17.9	**17.8**	**20.1**	**16.2**	**18.9**
Pair-Net† [30]	20.2	23.4	10.6	12.6	13.9	16.4
Mg-RMPN (Ours)	**23.2**	**25.7**	**12.8**	**14.5**	**16.5**	**18.5**

4.3 Visualization of Multi-Granularity Relationship Matrices

Figure 3 presents a visualization of the multi-granularity relationship matrices learned by our Mg-RMPN, where (d), (e), (f) are the sparse relationship matrices learned after being filtered by GPM-1, GPM-2, GPM-3, respectively, and (a), (b), (c) are the corresponding ground truth relationship matrices defined in Section 3.6. The visualization results show that the dense negative samples in the relationship matrix are filtered layer by layer according to different granularities. The consistency of the generated multi-granularity relationship matrices with their ground truth indicates that the three GPMs of Mg-RMPN indeed focus on filtering negative samples of different granularities.

Figures 3 (h) and (g), respectively, present a zoomed-in view of the final relationship matrix and its ground truth. We can observe that the predicted relationship matrix successfully captures the true relationships, for example: "man wearing shirt", "man holding racket", "man wearing short" and "pole on fence". Additionally, the relationship matrix also predicts some relationships that do not in the ground truth, as illustrated by the dashed lines in the rows

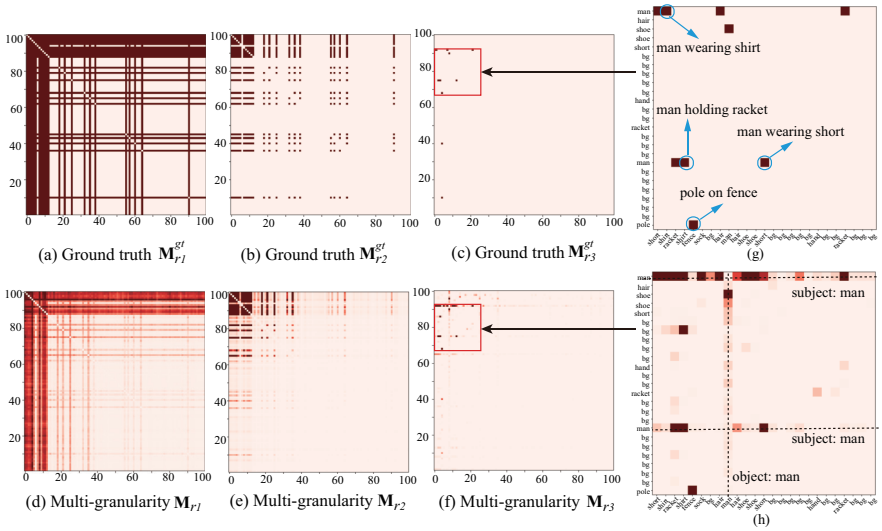

Fig. 3: Visualization of Multi-Granularity Relationship Matrices

Table 4: Ablation studies of Mg-RMPN components on VG150.

GPM-1	GPM-2	GPM-3	R@20/50/100	mR@20/50/100	F@20/50/100
✓	✗	✗	6.5/10.7/15.2	2.9/4.8/6.6	4.0/6.6/9.2
✗	✓	✗	12.5/17.5/21.6	5.2/7.4/9.3	7.3/10.4/13.0
✓	✓	✗	12.8/18.1/22.3	5.8/8.5/10.7	8.0/11.6/14.5
✗	✗	✓	21.2/27.6/32.1	9.2/13.1/15.8	12.8/17.8/21.2
✓	✗	✓	21.7/28.2/32.6	9.8/13.8/16.5	13.5/18.5/21.9
✗	✓	✓	21.8/28.2/32.5	9.7/13.7/16.7	13.4/18.4/22.1
✓	✓	✓	**22.5/29.1/33.5**	**10.3/14.4/17.3**	**14.1/19.3/22.8**

and columns, which represent the relationships with the corresponding entity "man" as the subject and object, respectively. Due to the abundant presence of "man" in SGG, the relationship matrix tends to favor predicting relations that include "man", which aligns with the actual language induction bias. Therefore, the results of visualization indicate that our Mg-RMPN can accurately predict the sparse relationship matrix.

4.4 Ablation Studies

Component Analysis of Mg-RMPN. To assess the effectiveness of each GPM in Mg-RMPN, we performed ablation experiments in Table 4. In Mg-RMPN, each GPM focuses on negative sample filtering of different granularities, and the three GPMs collaborate to complete the learning of the sparse relationship matrix. GPM-1 and GPM-2 are used to help GPM-3 filter negative

Table 5: Ablation studies of different input information for Mg-RMPN on VG150.

Visual	Semantic	Position	R@20/50/100	mR@20/50/100	F@20/50/100
✓	✗	✗	21.2/27.6/32.1	9.8/13.6/16.4	13.4/18.2/21.7
✓	✗	✓	22.2/28.7/33.1	10.1/14.1/16.9	13.9/18.9/22.4
✓	✓	✗	21.2/27.6/32.2	9.8/13.8/16.8	13.4/18.4/22.1
✓	✓	✓	**22.5/29.1/33.5**	**10.3/14.4/17.3**	**14.1/19.3/22.8**

samples. Using GPM-1 and GPM-2 individually cannot make full use of the ground truth, hence their results are very poor. The results obtained using only GPM-3 are also not good enough. Further increasing the assistance of GPMs for negative sample filtering can significantly improve the performance of the model. Mg-RMPN fully utilized three GPMs to achieve optimal results, confirming that learning negative samples based on different granularities can effectively improve the model's learning ability for sparse relationship matrices.

Different input information for Mg-RMPN This paper bridges entity detection and relation detection through a relationship matrix, making full use of the semantic and positional information of entities and solving the problem that end-to-end SGG relies solely on visual information to predict relations. Table 5 presents ablation studies of Mg-RMPN based on different input information. Compared to models that only utilize visual information, separately incorporating the semantic and positional information of entities can enhance the performance of the model. Furthermore, the improvement in introducing semantic and positional information simultaneously is significantly greater than in introducing a single modality of information, indicating that semantic and positional information are two complementary information. Hence, the semantic and positional information incorporated in our approach plays a crucial role in predicting relations. More hyperparameter analysis can be found in Supplementary Material.

5 Conclusion

In this paper, our proposed end-to-end method bridges entity detection and relation detection through a sparse relationship matrix, which not only eliminates the need for post-processing of relation matching but also leverages the semantics and positional information of predicted entities to enhance relation prediction. To predict the sparse relationship matrix, we propose a multi-granularity sparse relationship matrix prediction network, which utilizes three gated pooling modules focusing on filtering negative samples at different granularities. Finally, a set of sparse, most likely subject-object pairs can be constructed and used for relation decoding. The experimental results demonstrate that our method achieves an optimal overall performance in both the end-to-end and two-stage methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (U21A20485, 61976170, 62088102) and the National Key R&D Program of China (2023YFB4704900).

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
2. Chen, C., Zhan, Y., Yu, B., Liu, L., Luo, Y., Du, B.: Resistance training using prior bias: Toward unbiased scene graph generation. Proceedings of the AAAI Conference on Artificial Intelligence **36**(1), 212–220 (Jun 2022)
3. Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C.: Reformulating hoi detection as adaptive set prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9004–9013 (2021)
4. Chen, S., Jin, Q., Wang, P., Wu, Q.: 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. pp. 9962–9971 (2020)
5. 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)
6. Cong, Y., Yang, M.Y., Rosenhahn, B.: Reltr: Relation transformer for scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
7. 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)
8. Gao, L., Lyu, X., Guo, Y., Hu, Y., Li, Y.F., Xu, L., Shen, H.T., Song, J.: Informative scene graph generation via debiasing. arXiv preprint arXiv:2308.05286 (2023)
9. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
10. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1219–1228 (2018)
11. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3668–3678 (2015)
12. Kim, B., Lee, J., Kang, J., Kim, E.S., Kim, H.J.: Hotr: End-to-end human-object interaction detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 74–83 (2021)
13. 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)

14. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
16. Li, R., Zhang, S., He, X.: Sgtr: End-to-end scene graph generation with transformer. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19486–19496 (2022)
17. 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)
18. Li, W., Zhang, H., Bai, Q., Zhao, G., Jiang, N., Yuan, X.: Ppdl: Predicate probability distribution based loss for unbiased scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19447–19456 (2022)
19. Li, Y., Ma, T., Bai, Y., Duan, N., Wei, S., Wang, X.: Pastegan: A semi-parametric method to generate image from scene graph. Advances in Neural Information Processing Systems **32** (2019)
20. 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)
21. 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)
22. 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)
23. Lyu, X., Gao, L., Xie, J., Zeng, P., Tian, Y., Shao, J., Shen, H.T.: Generalized unbiased scene graph generation. arXiv preprint arXiv:2308.04802 (2023)
24. 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)
25. Schroeder, B., Tripathi, S.: Structured query-based image retrieval using scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 178–179 (2020)
26. 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)
27. 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)
28. Teney, D., Liu, L., van Den Hengel, A.: Graph-structured representations for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2017)
29. Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9695–9704 (2021)
30. Wang, J., Wen, Z., Li, X., Guo, Z., Yang, J., Liu, Z.: Pair then relation: Pair-net for panoptic scene graph generation. arXiv preprint arXiv:2307.08699 (2023)

31. Wang, L., Yuan, Z., Chen, B.: Learning to generate an unbiased scene graph by using attribute-guided predicate features. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2581–2589 (2023)
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. 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)
34. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–685 (2018)
35. Yu, J., Chai, Y., Wang, Y., Hu, Y., Wu, Q.: Cogtree: Cognition tree loss for unbiased scene graph generation. In: Zhou, Z.H. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 1274–1280. International Joint Conferences on Artificial Intelligence Organization (8 2021)
36. 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)
37. Zhang, A., Yao, Y., Chen, Q., Ji, W., et al, L.: Fine-grained scene graph generation with data transfer (2022)
38. 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)
39. Zhang, J., Shih, K.J., Elgammal, A., Tao, A., Catanzaro, B.: Graphical contrastive losses for scene graph parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11535–11543 (2019)
40. Zheng, C., Lyu, X., Gao, L., Dai, B., Song, J.: Prototype-based embedding network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22783–22792 (2023)
41. Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: European Conference on Computer Vision. pp. 211–229. Springer (2020)
42. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)