

Structured Sparse R-CNN for Direct Scene Graph Generation

Yao Teng Limin Wang 

State Key Laboratory for Novel Software Technology, Nanjing University, China

tengyao19980325@gmail.com, lmwang@nju.edu.cn

Abstract

Scene graph generation (SGG) is to detect object pairs with their relations in an image. Existing SGG approaches often use multi-stage pipelines to decompose this task into object detection, relation graph construction, and dense or dense-to-sparse relation prediction. Instead, from a perspective on SGG as a direct set prediction, this paper presents a simple, sparse, and unified framework, termed as Structured Sparse R-CNN. The key to our method is a set of learnable triplet queries and a structured triplet detector which could be jointly optimized from the training set in an end-to-end manner. Specifically, the triplet queries encode the general prior for object pairs with their relations, and provide an initial guess of scene graphs for subsequent refinement. The triplet detector presents a cascaded architecture to progressively refine the detected scene graphs with the customized dynamic heads. In addition, to relieve the training difficulty of our method, we propose a relaxed and enhanced training strategy based on knowledge distillation from a Siamese Sparse R-CNN. We perform experiments on several datasets: Visual Genome and Open Images V4/V6, and the results demonstrate that our method achieves the state-of-the-art performance. In addition, we also perform in-depth ablation studies to provide insights on our structured modeling in triplet detector design and training strategies. The code and models are made available at <https://github.com/MCG-NJU/Structured-Sparse-RCNN>.

1. Introduction

Scene graph generation (SGG) [45] aims at detecting objects with their pairwise relations in an image. This structured representation could serve as an effective and compact representation for high-level visual understanding tasks such as image captioning [47, 48] and visual question answering [2, 11, 32]. Structure information between visual entities is the key to the success of many SGG methods.

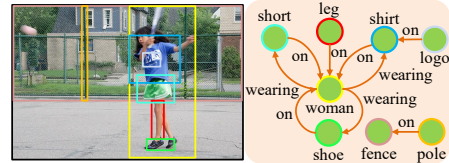



Figure 1. An example of scene graph generation. The scene graph is relatively sparser than the fully connected graph.

To capture this structure information, most existing methods typically follows a multi-stage pipeline to decompose this complex task into sub-tasks of object detection, fully-connected relation graph construction, dense relation classification [37, 49, 52], or dense-to-sparse relation classification [46], as shown in Fig. 2. These well-established methods often rely heavily on object detection performance and involve redundant computation for fully-connected relation graph construction.

In addition to structure information, we observe that sparsity is another important property on relation detection in natural images. For example, in Fig. 1, the ground-truth triplets of $\langle \text{leg}, \text{on}, \text{woman} \rangle$ and $\langle \text{logo}, \text{on}, \text{shirt} \rangle$ are more commonly expressed than the relation between *logo* and *leg*. Most existing dense or dense-to-sparse detection methods for SGG fails to well capture the general sparse and semantic priors. Accordingly, inspired by the recent sparse object detectors (e.g. DETR [3], Sparse R-CNN [34]), we present a new perspective on SGG by treating it as a direct sparse set prediction problem. However, unlike sparse object detection, sparse SGG is much more challenging due to its inherent difficulty in object pairing and relation prediction.

In this paper, we propose a direct sparse scene graph generation framework without explicit object detection and relation graph construction for inference, coined as Structured Sparse R-CNN. As shown in Fig. 2c, the key to our Structured Sparse R-CNN is a set of learnable *triplet queries* and a structured *triplet detector*. These learnable triplet queries, composed of two object boxes, two object content vectors and one relation content vector, are responsible for capturing the general prior for sparse detection and encoding the spatial and appearance information of objects and their relation. Based on the input of CNN fea-

 Corresponding author.

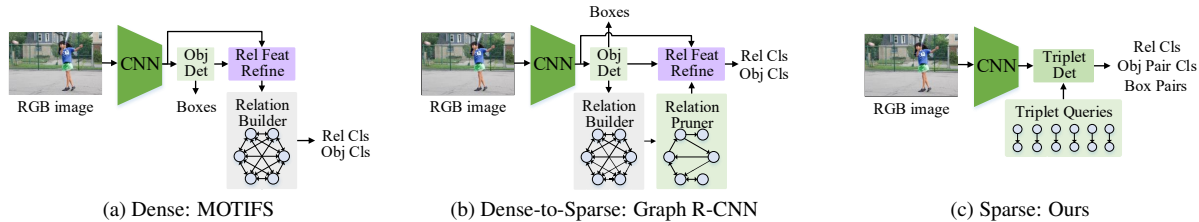


Figure 2. **Comparison of scene graph generation pipeline.** (a) The dense detectors enumerates all object pairs for relation inference, *e.g.* MOTIFS [49]. (b) The dense-to-sparse detectors utilize a pruning scheme to remove unreasonable pairs before the relation inference, *e.g.* Graph R-CNN [46]. (c) Our network directly generates sparse scene graphs with triplet queries.

ture maps and triplet queries, our triplet detector progressively detects the visual entities and recognizes their relations. The triplet detector contains two cascaded modules for object pair detection and their relation prediction, respectively. Specifically, we devise structured connections for each triplet query to capture the hierarchical context information. These structured connections first leverage the local interaction in object pairs (Pair Fusion) for better detection and then utilize the object information (E2R Fusion) for better relation prediction. The parameters of triplet queries are jointly optimized with network weights.

In practice, we find it is challenging to directly train our Structured Sparse R-CNN from scratch. The major challenge comes from the relatively sparse annotations of relations in the current datasets. The sparse relation annotations contain *few* related object labels, leading to incomplete supervision signal for our object pair detection. Furthermore, the negative samples are hard to define in the object pair level. To solve this issue, we propose to build a Siamese Sparse R-CNN to guide the training of our Structured Sparse R-CNN in a knowledge distillation framework [10]. This Sparse R-CNN only generates pseudo-labels [25, 31, 43] for training and is inactivated during testing. With the help of these pseudo-labels, we design a new relaxed matching criteria for set prediction loss and enable the training of Structured Sparse R-CNN to be more stable. Finally, to deal with imbalance distribution of object and relation categories in datasets, we propose an adaptive focusing parameter in our focal loss and utilize a post-hoc logit adjustment, to further boost the performance.

To verify the effectiveness of our framework, we perform experiments on several datasets: Visual Genome [15] and Open Images V4/V6 [17]. The experiment results demonstrate that our model is able to yield new state-of-the-art performance under setting of the same backbone on all datasets. In addition, we conduct in-depth ablation studies to verify the effectiveness of structure modeling in our design. In summary, our main contribution is threefold:

- We present a new sparse and unified framework for direct scene graph generation, without explicit object detection and preceding graph construction for inference. This new framework equipped with the struc-

ture connection proposed by us shares several advantages, namely simplicity without multi-stage design, effective context modeling, and high efficiency.

- We present a practical training strategy to overcome the training difficulty of Structured Sparse R-CNN. The knowledge distilled from a Siamese Sparse R-CNN can generate useful pseudo-labels to guide our training. We also propose an adaptive focusing parameter and utilize logit adjustment for imbalance distribution of objects and relations.
- Experiment results demonstrate that our simple framework is able to yield the state-of-the-art performance for scene graph generation on Visual Genome and Open Images V4/V6. We also perform detailed ablation studies to provide insights on our designs.

2. Related Work

Scene Graph Generation (SGG). In this part, we will discuss the existing works for SGG from three aspects: relation modeling, pipeline and long-tailed distribution. The explicit modeling for relations [18–20, 38, 41, 45, 46] is commonly considered. Xu *et al.* [45] built a bipartite graph composed of object proposals as object nodes and union proposals as relation nodes, and the message was passed between them to emphasize features. Yang *et al.* [46] utilized the pairwise object features to select few candidates of relation nodes in GCN [14] for classification. Since MOTIFS [49] was proposed, many works [23, 36, 37, 42, 49, 52] begin to aggregate context information for relation classification into object nodes, and the explicit relation features are only treated as attachments. As for the pipeline, almost all previous works revolve around the concept of multi-stage relation detection and ignore the feasibility of the one-stage paradigm. Recently, some works started to focus on the one-stage relation detection [5, 13, 24, 29, 35, 50, 54], but almost all of them still consider explicit post-hoc object detection to boost the performance, thereby making the overall framework a bit complicated. Some of them even do not make full use of the sparse and semantic priors. As for the long-tailed relation

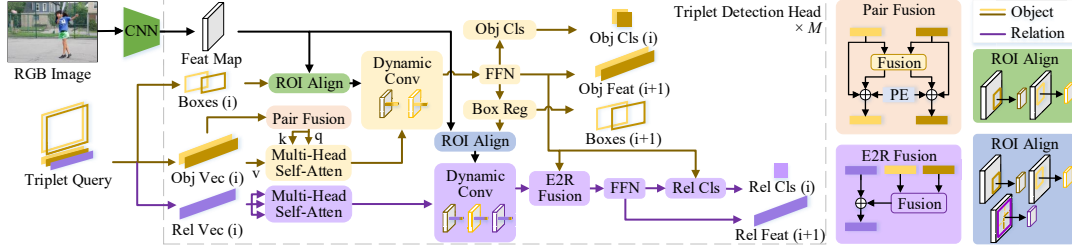


Figure 3. **Structured Sparse R-CNN.** Our method presents a simple, sparse and unified framework for direct scene graph generation without explicit object detection and relation graph construction in advance. Our framework is composed of CNN backbone, triplet queries, and triplet detector. The triplet queries encode the prior information on object boxes, object appearance, and relation appearance. The triplet detector consists of a series of detection heads. The detector takes CNN features and triplet queries as input, and progressively refine the relation detection results with two cascaded modules (marked in yellow and purple). The vectors of triplet queries are jointly optimized with the network weights with back-propagation. (i) in this figure denotes the index of current head. PE denotes positional encoding.

distribution, Tang *et al.* [36] performed a variant of logit calibration based on causal analysis. Li *et al.* [18] studied the re-sampling approach for SGG. In this paper, we directly generate a graph based on a sparse set of queries as its basic elements with efficiency and accuracy. We also propose a corresponding training strategy to get rid of explicit object detection when performing inference. In addition, we revisit the explicit modeling of relations and utilize logit adjustment [28] for the long-tailed datasets.

Sparse Object Detector. Recently, numerous works for sparse object detection were proposed. DETR [3] uses the Hungarian loss [3] and a transformer [40] architecture for object detection based on few queries as sparse anchors. Deformable DETR [53] boosts the performance by combining the deformable convolution [6] with the transformer and utilizing multi-scale features. Sparse R-CNN [34] is more lightweight than these methods and easier to serve as a baseline. In this paper, we extend Sparse R-CNN into our sparse triplet detector for generalized relation detection, and design corresponding structures as well as specific training strategy for sparse SGG.

3. Proposed Approach

Overview. Unlike the previous SGG methods composed of multiple stages, our Structured Sparse R-CNN presents a simple, direct and unified framework for relation detection. Our method takes image features and a set of triplet queries as inputs, and passes them into the stacked detection heads to progressively detect objects and predict their relations. The parameters of triplet queries can be jointly optimized with network weights in an end-to-end manner. We detail these components in the sequel.

3.1. Structured Sparse R-CNN

Backbone. The image is fed into a convolution neural network (CNN) [44] with Feature Pyramid Network (FPN) [21] for feature extraction, and then the feature maps

are fed into our triplet detector to detect objects and predict relations. More details can be found in Section 4.2.

Triplet query. To localize objects and recognize their categories and relations, our Structured Sparse R-CNN uses a set of learnable triplet queries to represent the general distribution prior of triplets. Specifically, each triplet query is composed of two proposal boxes representing the locations of objects, two object content vectors encoding the appearance of objects, one relation content vector capturing the structure information between objects. Each box is a 4-d parameter to represent the normalized box center, width, and height. The object and relation features are represented by 1024-d and 256-d parameters respectively, which encode the semantics of objects and relations.

These triplet queries are randomly initialized during training and jointly optimized with network weights via back-propagation algorithm. Once the training is finished, these learnt triplet queries serve as the general prior for SGG and are the same for all testing images. Basically, the learnt triplet queries could be viewed as the general statistics of potential objects location, appearance, and their relations, discovered in a data-driven manner from training set. They provide an initial guess for the triplet candidates, which is then refined progressively with the triplet detector.

Triplet detection head. Our Structured Sparse R-CNN is composed of a series of modular network building blocks, termed as Triplet Detection Head, to progressively refine the location and categories of objects as well as the prediction of relations. As shown in Fig. 3, each head of our triplet detector presents two modules to perform object pair detection and their relation prediction, respectively. These two modules are cascaded together with structured connections to accomplish the task of SGG.

Object pair detection. Given N triplet queries, triplet detector first uses object feature vectors to perform the global and local information interaction. The traditional multi-head self-attention mechanism is employed for aggregating

global context information into objects. To better describe the context features within object pairs, we propose a pair fusion module (PF) to relate the object feature vectors by using a multi-layer perception (MLP). The meaning of this structured connection lies in emphasizing each object feature via utilizing the unique properties of the internal interaction, *e.g.* in one triplet, its subject will be aware of which object to pair with. Moreover, the relations are unlikely to occur between the same objects. Therefore, this operation is designed to separate the objects and enhance their semantics. Its specific process is as follows:

$$\begin{aligned} X_p &= \text{ReLU}(\text{LN}(W_0^s X_s + W_0^o X_o)), \\ X'_s &= X_s + W_1^s X_p + P_s, \quad X'_o = X_o + W_1^o X_p + P_o, \end{aligned} \quad (1)$$

where X_s and X_o denote the subject and object content vectors, respectively. P_s and P_o are positional encoding for subjects and objects. $W_0^s, W_0^o, W_1^s, W_1^o$ and W_0 are learnable matrices. $\text{LN}(\cdot)$ and $\text{ReLU}(\cdot)$ represent the layer normalization [1] and ReLU activation [8]. X'_s and X'_o are used for generating the key and query vectors in self-attention.

Then the enhanced object feature vector is used to attend the RoI pooled feature of each object independently with a *dynamic* convolution [12], where the kernels for convolution are produced by object feature vectors. Subsequently, a feed-forward network (FFN) [40] with two MLPs (*i.e.*, *cls* and *reg* heads) is constructed for object box regression and category classification, respectively.

Relation recognition. After the object pair detection, our triplet detector perform visual relation prediction for each detected object pair. After performing a similar dynamic convolution on relation-level features from ROI Align [9] with relation vectors, we introduce a bottom-up connection to combine the object-level features with our relation feature vectors. This bottom-up structured connection is called as visual entities to relation fusion, denoted by E2R Fusion (E2R). These object-level features are expected to enhance the relation vectors by providing low-level object information via other MLPs:

$$\begin{aligned} H_r &= W_x \text{ReLU}(\text{LN}(W_r^s F_s)) + W_y \text{ReLU}(\text{LN}(W_r^o F_o)), \\ F'_r &= \text{LN}(F_r + H_r + W_r^p \text{ReLU}(W_p^s P_s + W_p^o P_o)), \end{aligned} \quad (2)$$

where F_s, F_o and F_r denote the features of the subjects, objects and relations, respectively. $W_r^s, W_r^o, W_x, W_y, W_p^s, W_p^o$ and W_r^p are linear transformation matrices. Finally, a FFN with relation classification head is used to conduct relation prediction with the enhanced relation vectors.

In addition, due to the object feature is helpful for relation prediction [52], we use object-level features to directly predict the relation categories as another branch. The final classification comes from the sum of the outputs of the master branch and this branch.

Discussion. Our Structured Sparse R-CNN is an extension of the original Sparse R-CNN to the structure predic-

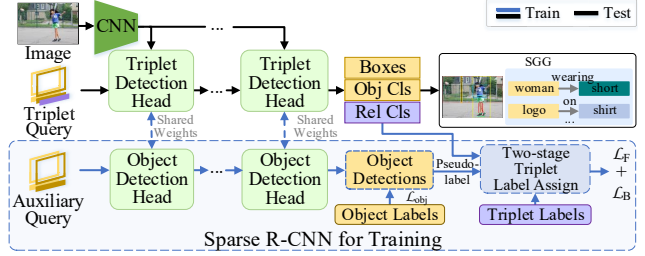


Figure 4. **Learning with Siamese Sparse R-CNN.** We present a relaxed and enhanced training strategy for our Structured Sparse R-CNN based on the knowledge distillation from a Siamese Sparse R-CNN which is composed of object detection heads. This extra Sparse R-CNN generate pseudo-labels for our triplet label assignment and also augment our triplet queries at each layer, benefiting the training of Structured Sparse R-CNN.

tion task. To mitigate the difficulty of relation detection over object detection, our Structured Sparse R-CNN introduces customized structure modeling in our triplet detector. The key difference with the original Sparse R-CNN is the consideration of structure information. First, we introduce the pairwise object context information for better object detection. Second, we model the hierarchical context information between two objects and their relation. As shown in experiments, this structure modeling is of great importance in our Structured Sparse R-CNN design to accomplish SGG.

3.2. Learning with Siamese Sparse R-CNN

Unlike Sparse R-CNN [34], we observe that it is challenging to directly train our Structured Sparse R-CNN only with the ground-truth triplets. These triplet annotations cover *too few object samples*. However, for our triplet detector, training its object pair detection component requires a large number of object samples. Therefore, we consider generating some virtual *object pairs* as pseudo-labels for training so as to increase the recall of object pair detection. For this objective, we present a relaxed and enhanced training strategy based on knowledge distillation [10] from an extra Sparse R-CNN which can yield a set of such pseudo-labels. To allocate these pseudo-labels to predicted object pairs in training, we design a two-stage triplet label assignment with specific classification and regression loss. Even if these pseudo-labels are not in annotations, they could be used to train the object pair detection component under the new label assignment and loss.

Siamese Sparse R-CNN. As shown in Fig. 4, we propose to build an extra Sparse R-CNN *only activated in the training phase*. This network shares the same weight with our Structured Sparse R-CNN and thus is called as Siamese Sparse R-CNN. It is separately employed for object detection, and has the *auxiliary queries* independent of the triplet ones. It is jointly trained with our triplet detector, and has its own *object label assignment* just like the object detec-

tors [3, 34]. The detected objects are grouped into pairs and act as pseudo-labels for training Structured Sparse R-CNN.

Two-stage triplet label assignment. In training, we first directly use Hungarian matching [3] to assign ground-truth relations with their objects to a set of triplet candidates. Then, for the remaining triplet candidates not matching the ground-truth triplets, instead of padding their objects with background label, we use another Hungarian matching to assign these *object pairs* to a subset of pseudo-labels provided by Siamese Sparse R-CNN. With such a matching, these triplets are forced to approximate object pairs that most resemble them. Finally, with the two-stage label assignment, we compute the loss for triplet detection as the sum of \mathcal{L}_F for the triplets matched in the first stage and \mathcal{L}_B for the ones matched in the second stage.

In the first stage, a bipartite matching is conducted between ground-truth triplets and all predicted triplets [16]. The following is the matching cost between a prediction and a ground-truth triplet, as well as a part of the final loss:

$$\begin{aligned} \mathcal{L}_F = & \lambda_{cls_r} \mathcal{L}_{cls_r}^g + \sum_{i \in \{s, o\}} \lambda_{cls_i} \mathcal{L}_{cls_i}^g \\ & + \lambda_{L_{1_i}} \mathcal{L}_{L_{1_i}}^g + \lambda_{giou_i} \mathcal{L}_{giou_i}^g, \end{aligned} \quad (3)$$

where $\mathcal{L}_{cls_i}^g$ and $\mathcal{L}_{cls_r}^g$ are focal loss [22] between ground-truth and predicted labels of objects and relations, respectively. *s/o* refers to the subject/object in one object pair. $\mathcal{L}_{L_{1_i}}^g$ and $\mathcal{L}_{giou_i}^g$ are L1 loss and generalized IoU loss [30] between the bounding boxes of objects and the corresponding ground-truth boxes, respectively. λ_{cls_r} , λ_{cls_i} , $\lambda_{L_{1_i}}$ and λ_{giou_i} are the coefficients of each component.

In the second stage, for the set of pseudo-labels, we remove some of its pairs that detect the ground-truth, and denote the remaining pairs as the pseudo-label set U . In principle, we could directly use U to train our triplet detector, but some works show that the hard-label format benefits training [39]. Therefore, considering the objects in U are also assigned with labels during the previous object label assignment, we keep the boxes of the objects *not* matching ground-truth objects *unchanged* and replace all the classification scores as well as other predicted boxes with the assigned labels. Then, we perform another bipartite matching between U and the object pairs from remaining triplet predictions. Due to the existence of predicted boxes, this stage of label assignment is like distillation. The matching cost between a predicted pair and a pair in U is as follows:

$$\mathcal{L}_B^m = \sum_{i \in \{s, o\}} \eta_{L_{1_i}} \mathcal{L}_{L_{1_i}}^u + \eta_{giou_i} \mathcal{L}_{giou_i}^u + \mathbf{1}_i^u \eta_{cls_i} \mathcal{L}_{cls_i}^u, \quad (4)$$

where $\mathcal{L}_{cls_i}^u$, $\mathcal{L}_{giou_i}^u$ and $\mathcal{L}_{L_{1_i}}^u$ is loss between predicted objects in triplets and the objects in U , just like those in Eq. (3). η_{cls_i} , $\eta_{L_{1_i}}$ and η_{giou_i} are coefficients. $\mathbf{1}_i^u$ is 1 if the object from U hits the ground-truth, otherwise is 0.

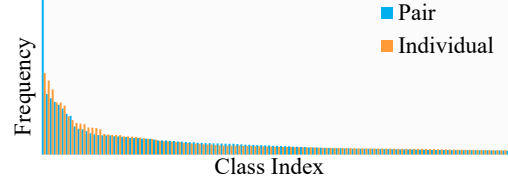


Figure 5. The class distributions of objects as individuals and pairs. The frequency from the two distributions is sorted in descending order separately.

After the bipartite matching of the second stage, we then pad the relation predictions in the remaining triplets with background label. The loss for these triplets is as follows:

$$\begin{aligned} \mathcal{L}_B = & \lambda_{cls_r} \mathcal{L}_{cls_r}^- + \sum_{i \in \{s, o\}} \eta_{cls_i} \mathcal{L}_{cls_i}^u \\ & + \mathbf{1}_i^u (\eta_{L_{1_i}} \mathcal{L}_{L_{1_i}}^u + \eta_{giou_i} \mathcal{L}_{giou_i}^u), \end{aligned} \quad (5)$$

where $\mathcal{L}_{cls_r}^-$ is focal loss between relation prediction and background label, the other terms are same with the above.

3.3. Imbalance Class Distribution

Adaptive focusing parameter. The format of triplets may deteriorate the imbalanced class distribution of entities. As shown in Fig. 5, the most frequented class of entities as the elements of pairs in distribution is far heavier than as individuals, which is attributed to the duplicates of entities induced by the format of triplets. Thus, we consider reducing the weights for majority classes in the loss for object classification. Inspired by [23], we re-balance the biased model by tailoring the focusing parameter (denoted as γ) in focal loss [22] for each category:

$$\gamma(c) = \min\{2, 3 - (1 - f_c)^\mu (-\log(f_c))^{\frac{1}{\mu}}\}, \quad (6)$$

where c denotes the object category. f_c denotes the frequency of each object category occurring in triplets. μ is a hyper-parameter.

Logit adjustment (LA). As for the imbalance class distribution of relations, we utilize logit adjustment [28]. We directly calculates the frequency for each relation category, and the final classification score is obtained from the logit minus the frequency multiplied by a tuning parameter τ .

4. Experiments

We conduct experiments on Visual Genome [15] and Open Images [17]. We describe evaluation settings, implementation details, ablation studies and comparisons to the state-of-the-art methods.

4.1. Datasets and Evaluation Settings

Visual Genome (VG). VG [15] is the most widely used dataset for SGG. We followed the widely adopted VG

Rel	E2R	PF	R@20	R@100	zR@20	zR@100	mR@20	mR@100	zR@100 (LA)	mR@100 (LA)
			23.16	34.62	0.79	2.44	5.09	8.62	3.03	19.04
		✓	25.33	36.57	0.94	2.41	5.67	9.13	2.96	20.25
✓		✓	25.52	36.58	1.38	3.41	6.03	9.83	3.61	19.96
✓	✓		24.32	35.34	1.24	3.32	5.77	9.37	3.83	19.55
✓	✓	✓	25.82	36.93	1.51	3.74	6.08	10.04	4.04	21.39

Table 1. Study on the structured modules. Rel: using relation feature vectors, PF: pairwise fusion, E2R: visual entities to relation fusion.

TLA	NMS	R@20	R@100	zR@20	zR@100	mR@20	mR@100	zR@100 (LA)	mR@100 (LA)	speed
full BG		24.62	35.00	1.21	3.11	5.75	9.20	3.52	18.57	0.19
full BG	✓	24.85	35.14	1.24	3.27	5.83	9.26	3.69	18.78	0.29
no BG		23.21	35.99	1.09	2.82	5.25	9.63	3.35	20.35	0.19
no BG	✓	25.45	<u>38.29</u>	1.33	3.79	5.93	<u>10.63</u>	3.98	<u>22.23</u>	0.29
p-label		25.82	36.93	1.51	3.74	6.08	10.04	4.04	21.39	0.19
p-label	✓	<u>26.49</u>	37.42	<u>1.62</u>	<u>4.09</u>	<u>6.27</u>	10.24	<u>4.19</u>	21.65	0.29

Table 2. Study on the triplet label assignment. TLA: triplet label assignment, no BG: training without background object label, full BG: assigning background label to all non-foreground entities, p-label: pseudo-label. Notably, in this table, the bolded and underlined values indicate the best results without and with NMS, respectively.

Adapt- γ	R@20	R@100	mR@20	mR@100	mR@100 (LA)
	25.58	36.57	5.93	9.76	20.90
✓	25.82	36.93	6.08	10.04	21.39

Table 3. Study on adaptive focusing parameter.

split [4, 37, 45, 49] including the most frequent 150 object categories and 50 relation categories. Since our paradigm generates triplet candidates based on queries, the modes based on ground-truth objects (e.g. PredCls and SGCls [27]) are not suitable here. We adopt the mode of SGDet, which considers both object detection and relation prediction. The traditional metrics on VG is Recalls [27]. Due to the imbalanced class distribution of relations in VG, Recalls are dominated by frequent categories. Thus, following [36], we also utilize mean Recalls (mR) [37] and zero-shot Recalls (zR) [27] for evaluation.

Open Images (OI). OI [17] is another large-scale dataset containing annotations for SGG. Currently, two benchmarks for SGG are built on the two versions of this dataset, namely OI V4 and OI V6, respectively. On each benchmark, we carried out experiments and utilized the same backbone as used in [18], and followed their data processing and evaluation metrics. The training sets and testing sets of OI V4 contain 54k images and 3k images, respectively. It contains 57 object categories and 9 relation categories. OI V6 includes 126k images for training, 2k and 5k images for validation and testing, respectively. It contains 601 object categories and 30 relation categories. For both OI V4 and OI V6, the results are evaluated with the metric of mean Recall@50, Recall@50, weighted mean AP of triplets (wmAP_{rel}), and weighted mean AP of phrase (wmAP_{phr}). The wmAP_{rel} evaluates the AP of the predicted triplet in

which both the subject and object boxes have an IoU of at least 0.5 with ground-truth, while the wmAP_{phr} uses the union area of the subject and object boxes for IoU calculation. The final evaluation score is calculated by $\text{score} = 0.2 \times \text{Recall@50} + 0.4 \times \text{wmAP}_{rel} + 0.4 \times \text{wmAP}_{phr}$.

4.2. Implementation Details

For fair comparison on OI V4, OI V6 and VG, we utilize the same ResNeXt-101-FPN [21, 44] as the backbone for training Structured Sparse R-CNN. Our network is optimized by AdamW [26], and its initial learning rate and batchsize are set to 6.4×10^{-5} and 8, respectively. The number of total iterations is 80k, and the learning rate is decayed by the factor of 10 on the 47kth and 64kth iterations. Following [34], both the triplet detector and the object detector both have 6 detection heads. Since the classification scores in a triplet share the same weight when inference, the parameters in our loss are set as follows: $\lambda_{cls_r} = \lambda_{cls_i} = \frac{4}{3}$, $\lambda_{L1_i} = 5$, $\lambda_{giou_i} = 2$, $\eta_{cls_i} = \frac{1}{3}$, $\eta_{L1_i} = \frac{5}{4}$ and $\eta_{giou_i} = \frac{1}{2}$. As for focal loss, we set α and the fixed γ to 0.25 and 2, respectively. We set μ to 4 for the adaptive focusing parameter. The number of triplet queries is set to 300 and can be extended into 800. The number of auxiliary queries is set to 100. The τ in logit adjustment is set to 0.3. Notably, like Sparse R-CNN [34], NMS can be removed.

4.3. Ablation Study

We perform ablation studies on VG and report the performance on various Recalls. Furthermore, we report the performance of our models equipped with logit adjustment.

Study on the structure modeling. We begin our ablation study by exploring the importance of structure modeling in our design. Specifically, we first propose a purely

Model	R@20	R@50	R@100	zR@20	zR@50	zR@100	mR@20	mR@50	mR@100	speed
IMP [45]	18.1	25.9	31.2	0.2	0.4	0.8	2.8	4.2	5.3	0.43
G-RCNN [46]	-	29.7	32.8	-	-	-	-	5.8	6.6	-
VTransE [51]	24.5	31.3	35.5	-	1.9	2.6	5.1	6.8	8.0	0.40
RelDN [52]	-	31.4	35.9	-	-	-	-	6.0	7.3	-
GPS-Net [23]	-	31.1	35.9	-	-	-	-	7.0	8.6	-
MOTIFS [49]	25.1	32.1	36.9	-	0.1	0.2	4.1	5.5	6.8	0.45
MOTIFS(Focal) [49]	24.7	31.7	36.7	-	0.1	0.3	3.9	5.3	6.6	-
MOTIFS(EBM) [33]	24.3	31.7	36.3	0.1	0.2	-	5.7	7.7	9.3	-
VCtree [37]	24.5	31.9	36.2	0.1	0.3	0.7	5.4	7.4	8.7	0.67
VCtree(EBM) [33]	24.2	31.4	35.9	0.2	0.4	-	5.7	7.7	9.1	-
Transformer [40]	25.6	33.0	37.4	0.0	0.1	0.3	6.0	8.1	9.6	0.38
VTransE _{TDE} [51]	13.5	18.7	22.6	-	2.0	2.7	6.3	8.6	10.5	-
MOTIFS _{TDE} [49]	12.4	16.9	20.3	-	2.3	2.9	5.8	8.2	9.8	-
VCtree _{TDE} [37]	14.0	19.4	23.2	-	2.6	3.2	6.9	9.3	11.1	-
VCtree(EBM) _{TDE} [33]	14.7	20.6	24.7	1.6	2.7	-	7.1	9.7	11.6	-
Transformer [†] _{TDE} [40]	11.2	15.6	19.0	1.4	2.0	2.5	6.9	9.2	10.9	-
BGNN [18]	-	31.0	35.8	-	-	-	-	10.7	12.6	-
Ours	25.8	32.7	36.9	1.5	2.7	3.7	6.1	8.4	10.0	0.19
Ours*	26.1	33.5	38.4	1.5	2.7	4.0	6.2	8.6	10.3	0.32
Ours _{TDE}	14.5	18.3	21.0	1.8	2.7	3.6	10.8	15.0	18.5	0.29
Ours* _{TDE}	15.0	19.7	22.9	1.6	2.7	3.8	9.8	14.6	18.0	0.54
Ours _{LA}	18.4	23.3	26.5	1.9	2.9	4.0	13.5	17.9	21.4	0.19
Ours* _{LA}	18.2	23.7	27.3	2.0	3.1	4.5	13.7	18.6	22.5	0.32

Table 4. Comparisons with the state-of-the-art methods at SGDet on Visual Genome (VG). * refers to the 800 queries. LA: logit adjustment [28]. The reimplemented model is denoted by the superscript †. The two blocks indicate the models with or without debiasing techniques, respectively.

Sparse R-CNN baseline without explicitly introducing the relation feature vector. This baseline simply treats relation detection as object pair detection without structure modeling and its performance is lower than other variants of Structured Sparse R-CNN, as shown in Tab. 1. Then, we investigate the effectiveness of structure modeling module (PF and E2R) in our method by detailed ablations. In Tab. 1, the results demonstrate that these structured connections are helpful for the relation detection.

Study on the triplet label assignment. We perform the ablation study on the effectiveness of two-stage triplet label assignment and report the results in Tab. 2. For fair comparison, we report results all with co-training of Siamese Sparse R-CNN, and the only difference is label assignment strategy. First, we report the results of one-stage triplet label assignment, similar to the training of sparse object detectors, where the object candidates in unmatched triplets are all assigned with the background category (denoted by full BG). Then, we remove the background label assignment for those objects in unmatched triplets (denoted by no BG). From the comparison between these two settings, we find that the performance with background supervision at R@20 is better than without it. When the NMS post-processing is used, the performance of the full BG model is poor. We speculate the background supervision is key to duplicate

removal. However, for the objects in triplets with background relations, the full BG model will suppress them indiscriminately though some of them have localized ground-truth objects. Finally, we compare the previous strategies with our proposed pseudo-label assignment. Equipped with our pseudo-labels, the overall performance without NMS is better than that of the previous two training strategies, and NMS can still achieve good performance. Among these metrics, its performance on R@20 is consistently the highest. These results demonstrate the effectiveness of our proposed knowledge distillation framework on training our network. In addition, equipped with NMS, we observe that the absence of background object supervision leads to better performance than utilizing pseudo-labels. We speculate the noise in pseudo-labels influence the training, thereby resulting in more wrong predictions with low confidence, as well as the lower R@100 and mR@100 performance.

Study on the adaptive focusing parameter. We conduct comparative study on the adaptive focusing parameter. In Tab. 3, the improvement at R@100 and mR@100 shows its effectiveness.

4.4. Comparisons with the State of the Art

Visual Genome (VG). We compare our model to the results of the state-of-the-arts methods on VG, shown

Model	mR@50	R@50	wmAP _{rel}	wmAP _{phr}	score _{wtd}
RelDN [52]	70.40	75.66	36.13	39.91	45.21
GPS-Net [23]	69.50	74.65	35.02	39.40	44.70
BGNN [18]	72.11	75.46	37.76	41.70	46.87
Ours	72.62	74.92	43.47	48.17	51.64
Ours _{LA}	79.23	74.75	43.57	48.25	51.68

Table 5. Comparisons with the state-of-the-art methods on Open Images (OI) V4. Following [18], R@50 here is micro-Recall@50 [7], calculated directly on total ground-truth triplets.

Model	mR@50	R@50	wmAP _{rel}	wmAP _{phr}	score _{wtd}
MOTIFS [49]	32.68	71.63	29.91	31.59	38.93
RelDN [52]	33.98	73.08	32.16	33.39	40.84
VCTree [37]	33.91	74.08	34.16	33.11	40.21
G-RCNN [46]	34.04	74.51	33.15	34.21	41.84
GPS-Net [23]	35.26	74.81	32.85	33.98	41.69
BGNN [18]	40.45	74.98	33.51	34.15	42.06
Ours	42.84	76.66	41.47	43.64	49.38
Ours _{LA}	50.73	75.70	41.14	43.24	48.89

Table 6. Comparisons with the state-of-the-art methods on OI V6. Following [18], R@50 here is micro-Recall@50 [7].

in Tab. 4. Following the tradition, we first provide the results of each method on Recalls. However, VG has a long-tailed distribution of relation categories, and the traditional Recalls are dominated by the frequent categories such as "on". Accordingly, the mean and zero-shot Recalls are also utilized to evaluate the existing methods [36].

The results in Tab. 4 show that our Structured Sparse R-CNN achieves the state-of-the-arts performance on multiple metrics. Specifically, our model shows new state-of-the-arts performance in terms of zero-shot Recalls. As for mean Recalls, our basic model obtains the performance of 8.2% on average. Furthermore, our model with 800 queries achieves the best performance on Recalls with an average of 32.7%. We speculate that the reason why more queries lead to higher performance lies in the wide range of triplet combinations. With respect to the processing speed, we conduct experiments on the same server and our model achieves the fastest speed, 0.19 second per image, in the same experimental setting compared to other methods.

Following [36], we also report the results of various methods equipped with techniques against long-tailed relation category distribution in Tab. 4. Our model with TDE [36] or LA pushes the performance on zero-shot and mean Recalls in the task of SGDet to a new level. We think it is because the long-tailed relation class distribution limits the performance of models on mean Recalls. With the same debiasing techniques such as TDE, the effectiveness of our design on context feature utilization is revealed.

Open Images (OI). We demonstrate the effectiveness of our method on Open Images and the results are in Tab. 5

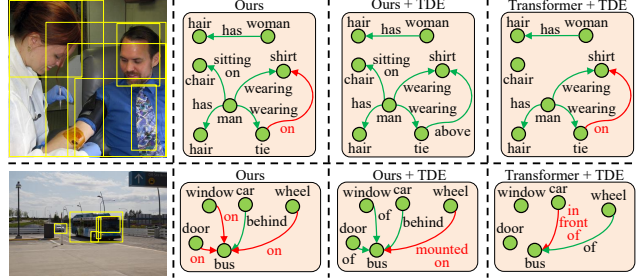


Figure 6. Results of Recall@100 from our model and another method. Due to space limitation, only the directed edges matching the ground-truth pairs are presented. The misclassified relations are marked in red.

and Tab. 6. Consistent with the higher performance at mean Recalls in VG, our method performs better under metrics for each class. When evaluated on R@50, our method still outperforms previous methods. Moreover, our model with logit adjustment performs slightly worse than the basic one under weighted metrics such as wmAP_{rel}, wmAP_{phr} and score_{wtd} in Tab. 6, with the drop on R@50.

Qualitative analysis. We visualize the detection results of SGG on VG in Fig. 6. In general, comparing the results between the last two column, we see that our model detects more correct relation prediction than previous state-of-the-art method, which demonstrates the effectiveness of our method.

5. Conclusion

In this paper, we have presented a new perspective on scene graph generation (SGG) as a direct set prediction problem, and proposed a simple, sparse, and unified framework for SGG, termed as Structured Sparse R-CNN. The key to our method is a set of learnable triplet queries and a structured triplet detector, which could be jointly optimized in an end-to-end manner. In addition, we present a relaxed and enhanced training strategy based on the knowledge distillation from a Siamese Sparse R-CNN. We also propose to use adaptive focusing parameter and logit adjustment for imbalance data distribution. We perform experiments on several datasets: Visual Genome and Open Images V4/V6, and the results demonstrate that our method achieves the state-of-the-art performance.

Acknowledgements. This work is supported by National Natural Science Foundation of China (No.62076119, No.61921006), Program for Innovative Talents and Entrepreneur in Jiangsu Province, and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

- [1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. 4
- [2] Hedi Ben-younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. BLOCK: bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *AAAI*, pages 8102–8109, 2019. 1
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 1, 3, 5
- [4] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *ICCV*, pages 4612–4622, 2019. 6
- [5] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating HOI detection as adaptive set prediction. In *CVPR*, pages 9004–9013, 2021. 2
- [6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 3
- [7] Nikolaos Gkanatsios, Vassilis Pitsikalis, Petros Koutras, and Petros Maragos. Attention-translation-relation network for scalable scene graph generation. In *ICCV Workshops*, pages 1754–1764, 2019. 8
- [8] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *AISTATS*, volume 15 of *JMLR Proceedings*, pages 315–323, 2011. 4
- [9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 4
- [10] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 2, 4
- [11] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 1
- [12] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NIPS*, pages 667–675, 2016. 4
- [13] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. HOTR: end-to-end human-object interaction detection with transformers. In *CVPR*, pages 74–83, 2021. 2
- [14] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2
- [15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 2, 5
- [16] Harold W. Kuhn. The hungarian method for the assignment problem. In *50 Years of Integer Programming*, pages 29–47. Springer, 2010. 5
- [17] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982, 2018. 2, 5, 6
- [18] 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. 2, 3, 6, 7, 8
- [19] Yikang Li, Wanli Ouyang, and Xiaogang Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. *CoRR*, abs/1702.07191, 2017. 2
- [20] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: An efficient subgraph-based framework for scene graph generation. In *ECCV*, pages 346–363, 2018. 2
- [21] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, 2017. 3, 6
- [22] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017. 5
- [23] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, pages 3743–3752, 2020. 2, 5, 7, 8
- [24] Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11546–11556, 2021. 2
- [25] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *ICLR*, 2021. 2
- [26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6
- [27] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016. 6
- [28] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 3, 5, 7
- [29] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *NIPS*, pages 2171–2180, 2017. 2
- [30] Hamid Rezaatoughi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019. 5
- [31] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh Singh Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *ICLR*, 2021. 2

- [32] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *CVPR*, pages 8376–8384, 2019. 1
- [33] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gérard G. Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *CVPR*, pages 13936–13945, 2021. 7
- [34] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse R-CNN: end-to-end object detection with learnable proposals. In *CVPR*, pages 14454–14463, 2021. 1, 3, 4, 5, 6
- [35] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, pages 10410–10419, 2021. 2
- [36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3713–3722, 2020. 2, 3, 6, 8
- [37] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019. 1, 2, 6, 7, 8
- [38] Yao Teng, Limin Wang, Zhifeng Li, and Gangshan Wu. Target adaptive context aggregation for video scene graph generation. In *ICCV*, pages 13668–13677, 2021. 2
- [39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357, 2021. 5
- [40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017. 3, 4, 7
- [41] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation. In *CVPR*, pages 8188–8197, 2019. 2
- [42] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *ECCV*, pages 222–239, 2020. 2
- [43] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan L. Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, pages 10857–10866, 2021. 2
- [44] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 3, 6
- [45] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 3097–3106, 2017. 1, 2, 6, 7
- [46] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for scene graph generation. In *ECCV*, pages 690–706, 2018. 1, 2, 7, 8
- [47] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694, 2019. 1
- [48] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, pages 711–727, 2018. 1
- [49] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018. 1, 2, 6, 7, 8
- [50] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage HOI detection. *Advances in Neural Information Processing Systems*, 34, 2021. 2
- [51] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, pages 3107–3115, 2017. 7
- [52] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, pages 11535–11543, 2019. 1, 2, 4, 7, 8
- [53] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3
- [54] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with HOI transformer. In *CVPR*, pages 11825–11834, 2021. 2