Expanding Scene Graph Boundaries: Fully Open-vocabulary Scene Graph Generation via Visual-Concept Alignment and Retention

Zuyao Chen[1,2], Jinlin Wu[2,3], Zhen Lei[2,3,4], Zhaoxiang Zhang[2,3,4], and Chang Wen Chen[1] *

[1] The Hong Kong Polytechnic University
[2] Centre for Artificial Intelligence and Robotics, HKISI-CAS
[3] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences
zuyao.chen@connect.polyu.hk
{wujinlin2017, zhen.lei, zhaoxiang.zhang}@ia.ac.cn
changwen.chen@polyu.edu.hk

Abstract. Scene Graph Generation (SGG) offers a structured representation critical in many computer vision applications. Traditional SGG approaches, however, are limited by a closed-set assumption, restricting their ability to recognize only predefined object and relation categories. To overcome this, we categorize SGG scenarios into four distinct settings based on the node and edge: Closed-set SGG, Open Vocabulary (object) Detection-based SGG (OvD-SGG), Open Vocabulary Relation-based SGG (OvR-SGG), and Open Vocabulary Detection + Relation-based SGG (OvD+R-SGG). While object-centric open vocabulary SGG has been studied recently, the more challenging problem of relation-involved open-vocabulary SGG remains relatively unexplored. To fill this gap, we propose a unified framework named OvSGTR towards fully open vocabulary SGG from a holistic view. The proposed framework is an end-to-end transformer architecture, which learns a visual-concept alignment for both nodes and edges, enabling the model to recognize unseen categories. For the more challenging settings of relation-involved open vocabulary SGG, the proposed approach integrates relation-aware pre-training utilizing image-caption data and retains visual-concept alignment through knowledge distillation. Comprehensive experimental results on the Visual Genome benchmark demonstrate the effectiveness and superiority of the proposed framework. Our code is available at https://github.com/gpt4vision/OvSGTR/.

1 Introduction

Scene Graph Generation (SGG) aims to generate a descriptive graph that localize objects in an image and simultaneously perceive visual relationships among

* Corresponding author.

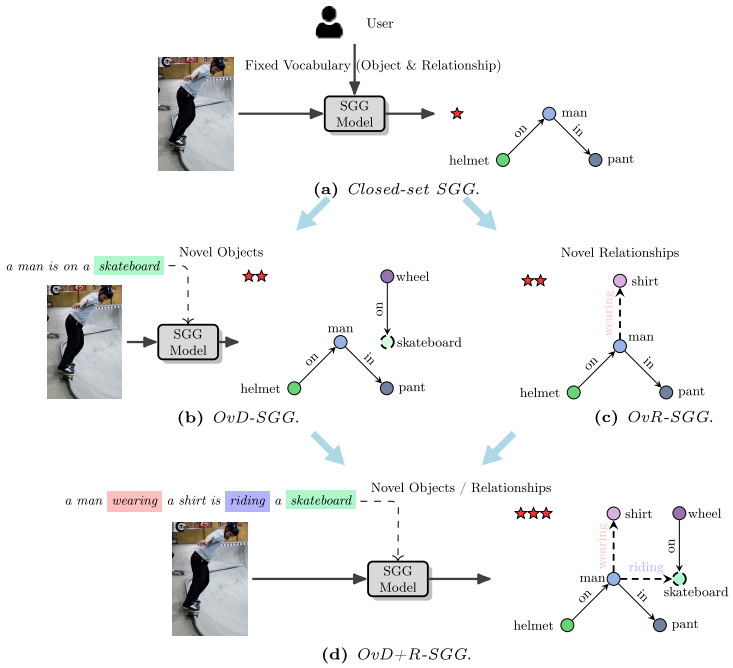

Fig. 1: Illustration of SGG Scenarios (best view in color). Dashed nodes or edges in (a) - (d) refer to unseen category instances, and stars refer to the difficulty of each setting. Previous works [2, 5, 18, 34, 35, 41, 46, 47] mainly focus on *Closed-set SGG* and few studies [10, 48] cover *OvD-SGG*. In this work, we give a more comprehensive study towards fully open vocabulary SGG.

object pairs. Such a structured representation has gained much attention, serving as a foundational component in many vision applications, including image captioning [1, 9, 26, 37, 43], visual question answering [12, 14, 27, 36], and image generation [11, 42].

Despite significant advancements in SGG, prevailing approaches predominantly operate within a confined set-up, *i.e.*, they constrain object and relation categories to a predefined set. This setting hampers the broader applicability of SGG models in diverse real-world applications. Influenced by the achievements in open vocabulary object detection [7,15,40,45,50], recent works [10,48] attempt to extend the SGG task from closed-set to open vocabulary domain. However, they focus on an object-centric open vocabulary setting, which only considers the scene graph nodes. A holistic approach to open vocabulary SGG requires a comprehensive analysis of nodes and edges. This raises two crucial questions that serve as the driving force behind our research: ***Can the model predict unseen objects or relationships ? What if the model encounters both unseen objects and unseen relationships?***

Given these two questions, we recognize the need to re-evaluate the traditional settings of SGG and propose four distinct scenarios: *Closed-set SGG*, Open

Vocabulary (object) Detection-based SGG (*OvD-SGG*), which expands to detect objects beyond a closed set, Open Vocabulary Relation-based SGG (*OvR-SGG*), focusing on identifying a broader range of object relationships, and Open Vocabulary Detection+Relation-based SGG (*OvD+R-SGG*), which combines open vocabulary detection and relation analysis, as shown in Fig. 1. **1)** *Closed-set SGG*, extensively studied in previous works [2,5,18,34,35,41,46,47], involves predicting nodes (*i.e.*, objects) and edges (*i.e.*, relationships) from a predefined set. Generally, *Closed-set SGG* focuses on feature aggregation and unbiased learning for long-tail problems. **2)** *OvD-SGG*, which has recently gained attention [48], extends *Closed-set SGG* from the node perspective, aiming to recognize unseen object categories during inference. However, it still operates on a limited set of relationships. **3)** On the other hand, *OvR-SGG* introduces open vocabulary settings from the edge perspective, requiring the model to predict unseen relationships, *a more challenging task due to the absence of pre-trained relation-aware models and the dependence on less accurate scene graph annotations.* Specifically, *OvD-SGG* omits all unseen object categories during training, resulting in a graph with fewer nodes but correct edges. By contrast, *OvR-SGG* eliminates all unseen relation categories during training, yielding a graph with fewer edges. As a result, *the model for OvR-SGG is required to distinguish unseen relationships from "background"*. **4)** The most challenging scenario, *OvD+R-SGG*, involves both unseen objects and unseen relationships, resulting in sparse and less accurate graphs for learning. These distinct settings present different intrinsic characteristics and unique challenges.

With a clear understanding of the challenges posed by these settings, we introduce *OvSGTR* (*Open-vocabulary Scene Graph Transformers*), a novel framework designed to address the complexities of open vocabulary SGG. Our approach not only predicts unseen objects or relationships but also handles the challenging scenario where both object and relationship categories are unseen during the training phase. *OvSGTR* employs a visual-concept alignment strategy for nodes and edges, utilizing image-caption data for weakly-supervised relation-aware pre-training. The framework comprises three main components: a frozen image backbone for visual feature extraction, a frozen text encoder for textual feature extraction, and a transformer for decoding scene graphs. During the relation-aware pre-training, the captions are parsed into relation triplets, *i.e.*, (*subject*, *relation*, *object*), which provides a coarse and unlocalized scene graph for supervision. For the fine-tuning phase, relation triplets with location information (*i.e.*, bounding boxes) are sampled from manual annotations. These relation triplets are associated with visual features, and visual-concept similarities are computed for nodes and edges, respectively. Predictions regarding object and relation categories are subsequently derived from visual-concept similarities, which promotes the model's generalization ability on unseen object and relation categories.

Upon evaluating the settings for relation-involved open vocabulary SGG (*i.e.*, *OvR-SGG* and *OvD+R-SGG*), we empirically identified a significant issue of catastrophic forgetting pertaining to relation categories. Catastrophic forgetting

leads to a degradation in the model's ability to recall previously learned information from image-caption data when exposed to new SGG data with fine-grained annotations. To preserve the semantic space while minimizing compromises on the new dataset, we propose visual-concept retention with a knowledge distillation strategy to mitigate this concern. The knowledge distillation component utilizes a pre-trained model on image-caption data as a teacher to guide the learning of our student model, ensuring the retention of a rich semantic space of relations. Simultaneously, the visual-concept retention ensures that the model maintains its proficiency in recognizing new relations.

In short, the contributions of this work can be summarized as follows,

- We give a comprehensive and in-depth study on open vocabulary SGG from the perspective of nodes and edges, discerning four distinct settings including *Closed-set SGG*, *OvD-SGG*, *OvR-SGG*, and *OvD+R-SGG*. Our analysis delves into both quantitative and qualitative aspects, providing a holistic understanding of the challenges associated with each setting;
- The proposed framework is fully open vocabulary as both nodes and edges are extendable and flexible to unseen categories, which largely expand the application of SGG models in the real world;
- The integration of a visual-concept alignment with image-caption data significantly enriches relation-involved open vocabulary SGG, while our visual-concept retention strategy effectively counters catastrophic forgetting;
- Extensive experimental results on the VG150 benchmark demonstrate the effectiveness of the proposed framework, showcasing state-of-the-art performances across all settings.

2 Related Work

Scene Graph Generation (SGG) aims to generate an informative graph that localizes objects and describes the relationships between object pairs. Previous methods mainly focus on contextual information aggregation [35, 41, 46] , and unbias learning for long-tail problem [5, 18, 34]. Typically, a closed-set object detector like Faster-RCNN is used and cannot handle unseen objects or unseen relations, which limits the application of SGG models in the real world. Recent works [10, 48] attempt to extend closed-set SGG to object-centric open vocabulary SGG ; However, they still fail to generalize on unseen relations and the combination of unseen objects and unseen relations.

An alternative approach to boosting the SGG task lies in the utilization of weak supervision, particularly by harnessing image caption data, leading to the emergence of language-supervised SGG [19, 48, 49]. This method of language supervision provides a cheaper way for SGG learning than expensive and time-cost manual annotation. Although previous research [19, 48, 49] has shown the potential of this technique, it remains confined predominantly to closed-set relation recognition. By contrast, our framework is fully open vocabulary. It discards the synsets matching as used in [48, 49], enabling our model to learn rich semantic concepts for generalization on downstream tasks. Furthermore, we also build

a connection between language-supervised SGG and open vocabulary SGG, in which language-supervised SGG aims to reduce the alignment gap between visual and language semantic space.

In essence, our work can be perceived as a generalization of open vocabulary SGG, harmoniously integrated with closed-set SGG. To our understanding, ours is a pioneering effort in formulating a consolidated framework dedicated to realizing a fully open vocabulary SGG, encompassing both the nodes and edges of scene graphs.

Vision-Language Pretraining (VLP) has gained increasing attention recently for numerous vision-language tasks. Generally, the core problem of vision-language pretraining is learning an alignment for visual and language semantic space. For instance, CLIP [30] shows promising zero-shot image classification capabilities by utilizing contrastive learning on large-scale image-text datasets. Later, many methods [15, 23, 50] have been proposed for learning a fine-grained alignment for image region and language data, enabling the object detector to detect unseen objects by leveraging language information. The success of VLP on downstream tasks provides an exemplar for learning an alignment between visual features and relation concepts, which is fundamental to building a fully open vocabulary SGG framework.

Open-vocabulary Object Detection (OvD) expects to detect unseen classes in inference, which breaks the limitation of a fixed pre-defined object set (*e.g.*, 80 categories in COCO). To accomplish this goal, Ov-RCNN [45] transfers semantic knowledge learned from captions to the downstream object detection task. It is worth noting that supervision signals for unseen or novel classes are excluded during training detectors, while unseen classes can be included in the large vocabulary set of captions. Except for OvD, a series of methods and applications have been developed such as open-vocabulary segmentation [8], open-vocabulary video understanding [38], and open-vocabulary SGG [10,17,48]. A more in-depth analysis of open-vocabulary learning can refer to the literature [39] and [51].

3 Methodology

Given an image I, the objective of the SGG task is to produce a descriptive graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which node $v_i \in \mathcal{V}$ has location information (*i.e.*, bounding box) and object category information, and edge $e_{ij} \in \mathcal{E}$ measure the relationship between node v_i and node v_j. For open-vocabulary settings, the label set $\mathcal{C}$ (either for the node or the edge) is split into two disjoint sets : *base classes* $\mathcal{C}_B$ and *novel classes* $\mathcal{C}_N$ ($\mathcal{C}_B \cup \mathcal{C}_N = \mathcal{C}$, $\mathcal{C}_B \cap \mathcal{C}_N = \emptyset$).

3.1 Fully Open Vocabulary Architecture

As shown in Fig. 2, *OvSGTR* is a DETR-like architecture that comprises three primary components: a visual encoder for image feature extraction, a text encoder for text feature extraction, and a transformer for the dual purposes of

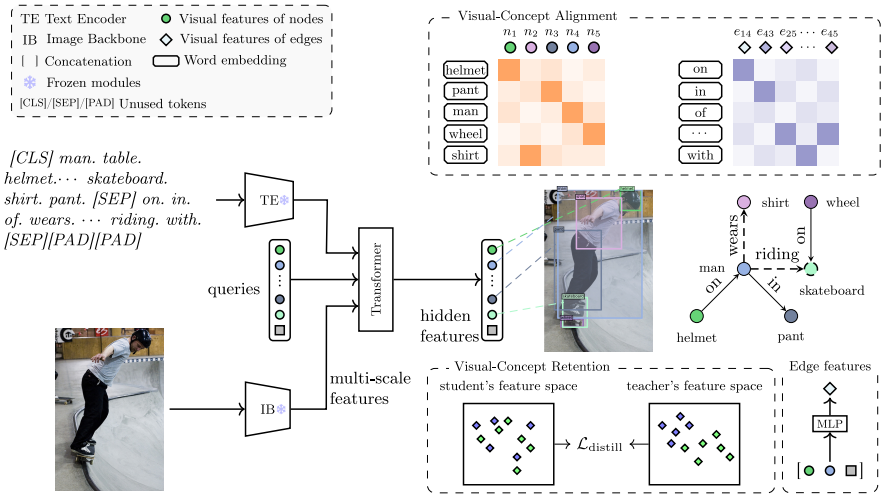

Fig. 2: Overview of our proposed OvSGTR . The proposed OvSGTR is equipped with a frozen image backbone to extract visual features, a frozen text encoder to extract text features, and a transformer for decoding scene graphs. Visual features for nodes are the output hidden features of the transformer; Visual features for edges are obtained via a light-weight relation head (*i.e.*, with only two-layer MLP). Visual-concept alignment associates visual features of nodes/edges with corresponding text features. Visual-concept retention aims to transfer the teacher's capability of recognizing unseen categories to the student.

object detection and relationship recognition. When provided with paired image-text data, *OvSGTR* is adept at generating corresponding scene graphs. To ease the optimization burden, the weights of both the image backbone and the text encoder are frozen during training.

Feature Extraction. Given an image-text pair, the model will extract multi-scale visual features with an image backbone like Swin Transformer [24] and extract text features via a text encoder like BERT [6]. Visual and text features will be fused and enhanced via cross-attention in the deformable encoder module of the transformer.

Prompt Construction. The text prompt is constructed by concatenating all possible (or sampled) noun phrases and relation categories, *e.g.*, *[CLS] girl. umbrella. table. bathing suit.⋯ zebra. [SEP] on. in. wears. ⋯ walking. [SEP][PAD][PAD]*, which is as similar as GLIP [15] or Grounding DINO [23] concatenating all noun phrases. For a large vocabulary set during training, we randomly sample negative words from the vocabulary set and constrain the number of positive and negative words to be M (*e.g.*, M=80).

Node Representation. Given K object queries, the model follows standard DETR to output K hidden features $\{\boldsymbol{v}_i\}_{i=1}^{K}$, which follow a bbox. head to decode the location information (*i.e.*, 4-d vectors), and a cls. head responsible for category classification. The bbox. head is a three-layer fully connected

layers. The cls. head is parameter-free, which computes the similarity between hidden features and text features. These hidden features are served as the visual representation for predicted nodes.

Edge Representation. Contrary to a complex and heavy message-passing mechanism for obtaining relation features, we design a lightweight relation head that concatenates node features for the subject and object, and relation query features. To learn a relation-aware representation, we use a random initialized embedding for querying relations. This relation-aware embedding will interact with image and text features by cross-attention in the decoder stage. Building on this design, given any possible subject-object pair (s_i, o_j), its edge representation can be obtained with $\boldsymbol{e}_{s_i \to o_j} = f_\theta([\boldsymbol{v}_{s_i}, \boldsymbol{v}_{o_j}, \boldsymbol{r}])$, where $\boldsymbol{v}_{s_i}, \boldsymbol{v}_{o_j}$ are node representation for the subject and object respectively, $\boldsymbol{r}$ refers to the relation query features, $[\cdot]$ refers to concatenation operation, and f_θ denotes a two-layer multi-perceptrons.

Loss Function. Following previous DETR-like methods [23, 52] , we use L1 loss and GIoU loss [32] for bounding box regression. For object or relation classification, we use Focal Loss [20] as the contrastive loss between prediction and language tokens.

To decode object and relation categories in a fully open vocabulary way, the fixed classifier (one fully connected layer) is replaced with a visual-concept alignment, which will be introduced in Sec. 3.2.

3.2 Learning Visual-Concept Alignment

Visual-concept alignment associates visual features for nodes or edges with corresponding text features. For node-level alignment, take an image as example, the model will output K predicted nodes $\{\tilde{v}_i\}_{i=1}^{K}$. These predicted nodes must be matched and aligned with N ground-truth nodes $\{v_i\}_{i=1}^{N}$. The matching is formulated as a bipartite graph matching, similar to the approach in standard DETR. This can be expressed as $\max_{\boldsymbol{M}} \sum_{i=1}^{N} \sum_{j=1}^{K} \mathrm{sim}(v_i, \tilde{v}_j) \cdot \boldsymbol{M}_{ij}$. Here, $\mathrm{sim}(\cdot, \cdot)$ measures the similarity between the predicted node and the ground-truth, which generally consider both the location (*i.e.*, bounding box) and category information. $\boldsymbol{M} \in \mathbb{R}^{N \times K}$ is a binary mask where the element $\boldsymbol{M}_{ij} = 1$ indicates a match between node v_i and node $\tilde{v}_j$. Conversely, a value of 0 indicates no match. For any matched pair $(v_i, \tilde{v}_j)$, we directly maximize its similarity, in which the distance between bounding boxes is determined by the L1 and GIoU losses, and category similarity is described as

$$\mathrm{sim}_{\mathrm{cat}}(v_i, \tilde{v}_j) = \sigma(< \boldsymbol{w}_{v_i}, \boldsymbol{v}_j >) \tag{1}$$

where $\boldsymbol{w}_{v_i}$ is the word embedding for node v_i, $\boldsymbol{v}_j$ is the visual representation for predicted node $\tilde{v}_j$, $< \cdot, \cdot >$ refers to the dot product of two vectors, and σ refers to the sigmoid function. This Eq. (1) seeks to align visual features for nodes with their prototypes in text space.

To extend relation recognition from closed-set to open vocabulary, one intuitive idea is to learn a visual semantic space in which visual features and text features for relations are aligned. Specifically, given a text input $\boldsymbol{t}$ and a text encoder

E_t, a relation feature $\boldsymbol{e}$, the alignment score is defined as $s(\boldsymbol{e}) =< \boldsymbol{e}, f(E_t(\boldsymbol{t})) >$, where f is one fully connected layer, and $< \cdot, \cdot >$ refers to the dot product of two vectors. Once the alignment score computed, we can calculate a binary cross entropy loss with given ground truths. The loss can be formulated as

$$\mathcal{L}_{\text{bce}} = \frac{1}{|\mathcal{P}| + |\mathcal{N}|} \sum_{\boldsymbol{e} \in \mathcal{P} \cup \mathcal{N}} \{-y_{\boldsymbol{e}} \log \sigma(s(\boldsymbol{e})) - (1 - y_{\boldsymbol{e}}) \log(1 - \sigma(s(\boldsymbol{e})))\} \tag{2}$$

where σ refers to sigmoid function, $y_{\boldsymbol{e}}$ is a one hot vector where "1" index positive tokens, and $\mathcal{P}$, $\mathcal{N}$ refer to positive and negative samples set for relations.

Learning such visual-concept alignment is non-trivial as there is a lack of relation-aware pre-trained models on large-scale datasets. In contrast, object-language alignment can be beneficial from pre-trained models such as CLIP [30] and GLIP [15]. On the other hand, manual annotation of scene graphs is time-consuming and expensive, which makes it hard to obtain large-scale SGG datasets. To tackle this problem, we leverage image-caption data as a weak supervision for relation-aware pre-training. Specifically, given an image-caption pair without bounding boxes annotation, we utilize an off-the-shelf language parser [25] to parse relation triplets from the caption. These relation triplets are associated with predicted nodes by optimizing Sec. 3.2, and only triplets with high confidence (*e.g.*, object score is greater than 0.25 for both subject and object) are reserved in scene graphs as pseudo labels. Utilizing these pseudo labels as a form of weak supervision, the model is enabled to learn rich concepts for objects and relations with image-caption data.

3.3 Visual-Concept Retention with Knowledge Distillation

Through learning a visual-concept alignment as described in Sec. 3.2, the model is expected to recognize rich objects and relations beyond a fixed small set. However, we empirically find that directly optimizing the model by Eq. (2) on a new dataset will meet catastrophic forgetting even if we have a relation-aware pre-trained model. On the other hand, in *OvR-SGG* or *OvD+R-SGG* settings, unseen (or novel) relationships are removed from the graph, which increases the difficulty as the model is required to distinguish novel relations from "background". To mitigate this problem, we adopt a knowledge distillation strategy to maintain the consistency of learned semantic space. Specifically, we use the initialized model pre-trained on image caption data as the teacher. The teacher has learned a rich semantic space for relations, *e.g.*, there exist $\sim 2.5k$ relation categories parsed from COCO caption [3] data. The student's edge features should be as close as the teacher's for the same negative samples. Thus, the loss for relationship recognition can be formulated as

$$\mathcal{L}_{\text{distill}} = \frac{1}{|\mathcal{N}|} \sum_{\boldsymbol{e} \in \mathcal{N}} ||\boldsymbol{e}^s - \boldsymbol{e}^t||_1 \tag{3}$$

where $\boldsymbol{e}^s$ and $\boldsymbol{e}^t$ refer to the student's and teacher's edge features, respectively. The total loss is given as $\mathcal{L} = \mathcal{L}_{\text{bce}} + \lambda \mathcal{L}_{\text{distill}}$, where λ controls the ratio of ground truths supervision and distillation part.

Table 1: Experimental results of *Closed-set SGG* on VG150 test set. "40M/177M" in Params. refers to 40M trainable parameters and 177M total parameters. Inference time is benchmarked on an NVIDIA RTX 3090 GPU with batch size 1 and an input resolution 1000×600. Time for SGNLS [49] is benchmarked on an NVIDIA A100 GPU (80G) due to memory out of usage.

SGG model	Backbone	Detector	Params.	R@20/50/100			mR@20/50/100			Time (s)
IMP [41]	RX-101		146M/308M	17.7	25.5	30.7	2.7	4.1	5.3	0.25
MOTIFS [46]	RX-101	Faster	205M/367M	25.5	32.8	37.2	5.0	6.8	7.9	0.27
VCTREE [35]	RX-101	R-CNN	197M/358M	24.7	31.5	36.2	-	-	-	0.38
SGNLS [49]	RX-101		165M/327M	24.6	31.8	36.3	-	-	-	> 7
HL-Net [21]	RX-101		220M/382M	26.0	33.7	38.1	-	-	-	0.10
FCSGG [22]	HRNetW48	-	87M/87M	13.6	18.6	22.5	2.3	3.2	3.9	0.13
SGTR [16]	R-101	DETR	36M/96M	-	24.6	28.4	-	-	-	0.21
VS^3 [48]	Swin-T	-	93M/233M	26.1	34.5	39.2	-	-	-	0.16
VS^3 [48]	Swin-L	-	124M/432M	27.3	36.0	40.9	4.4	6.5	7.8	0.24
OvSGTR	Swin-T	DETR	41M/178M	**27.0**	**35.8**	**41.3**	**5.0**	**7.2**	**8.8**	0.13
OvSGTR	Swin-B	DETR	41M/238M	**27.8**	**36.4**	**42.4**	**5.2**	**7.4**	**9.0**	0.19

4 Experiments

4.1 Datasets and Experiment setup

Datasets. The widely used VG150 dataset [41] contains 150 object and 50 relation categories annotated by humans. Of its 108,777 images, 70% are used for training, 5,000 for validation, and the rest for testing. Following VS^3 [48], we exclude images used in pre-trained object detector Grounding DINO [23] , retaining 14,700 test images. And we use an off-the-shelf language parser [25] to parse relation triplets from image caption, which yields $\sim 117k$ images with $\sim 44k$ phrases and $\sim 2.5k$ relations for COCO caption training set. To showcase the scalability of our model, we concat COCO caption data [3], Flickr30k [29], and SBU Captions [28] to construct a large-scale dataset for scene graph pre-training, resulting in $\sim 569k$ images with $\sim 198k$ type phrases and $\sim 5k$ relations.
Metrics. We adopt the **SGDET** [34,41] protocol (alias: **SGGen**) for fair comparison and report the performance on Recall@K (K=20/50/100) for each settings. Mean Recall@K (mR@K, K=20/50/100) and inference speed are reported under the setting of *Closed-set SGG*.
Implementation details. We use pre-trained Grounding DINO [23] models to initialize our model, and keep the visual backbone (*i.e.*, Swin-T or Swin-B) and text encoder (*i.e.*, BERT-base [6]) as frozen. Other modules like relation-aware embedding are initialized randomly. And 100 object detections per image are selected for pairwise relation recognition.

4.2 Compared with State-of-the-arts

Closed-set SGG Benchmark. The *Closed-set SGG* setting follows previous works [16,34,41,46,48], utilizing the VG150 dataset [41] with full manual annotations for training and evaluation. Experimental results on the VG150 test set are

Table 2: Experimental results (R@50/100) of *OvD-SGG* setting on VG150 test set. Following VS^3 [48], OvSGTR chooses image regions that best match the ground-truth objects in post-processing for PREDCLS.

Method	Base+Novel (Object)		Novel (Object)	
	PREDCLS	SGDET	PREDCLS	SGDET
IMP [41]	40.02 / 43.40	2.85 / 3.43	37.01 / 39.46	0.00 / 0.00
MOTIFS [46]	41.14 / 44.70	3.35 / 3.86	39.53 / 41.14	0.00 / 0.00
VCTREE [35]	42.56 / 45.84	3.56 / 4.05	41.27 / 42.52	0.00 / 0.00
TDE [34]	38.29 / 40.38	3.50 / 4.07	34.15 / 36.37	0.00 / 0.00
GCA [13]	43.48 / 46.26	-	42.56 / 43.18	-
EBM [33]	44.09 / 46.95	-	43.27/44.03	-
SVRP [10]	47.62 / 49.94	-	45.75 / 48.39	-
VS^3 [48] (Swin-T)	50.10 / 52.05	15.07 / 18.73	46.91 / 49.13	10.08 / 13.65
OvSGTR (Swin-T)	**60.58 / 62.10**	**18.14 / 23.20**	**59.01 / 60.65**	**12.06 / 16.49**
OvSGTR (Swin-B)	**60.83 / 62.33**	**21.35 / 26.22**	**59.30 / 60.95**	**15.58 / 19.96**

Table 3: Experimental results of *OvR-SGG* setting on VG150 test set. † refers to *w.o.* distillation.

Method	Base+Novel (Relation)		Novel (Relation)	
	R@50	R@100	R@50	R@100
IMP [41]	12.56	14.65	0.00	0.00
MOTIFS [46]	15.41	16.96	0.00	0.00
VCTREE [35]	15.61	17.26	0.00	0.00
TDE [34]	15.50	17.37	0.00	0.00
VS^3 [48] (Swin-T)	15.60	17.30	0.00	0.00
OvSGTR$_{Swin-T}$†	17.71	20.00	0.34	0.41
OvSGTR$_{Swin-T}$	**20.46**	**23.86**	**13.45**	**16.19**
OvSGTR$_{Swin-B}$†	18.58	20.84	0.08	0.10
OvSGTR$_{Swin-B}$	**22.89**	**26.65**	**16.39**	**19.72**

reported in Tab. 1, demonstrating that the proposed model outperforms all competitors. Notably, when compared to the recent VS^3 [48], *OvSGTR* (*w.* Swin-T) shows a performance gain of up to 3.8% for R@50 and 5.4% for R@100. The performance gain regarding mR@K reflects that our model handles the long-tail bias better than others.

Moreover, while many previous works rely on a complex message-passing mechanism to extract relation features, our model achieves strong performance with a simpler relation head consisting of only two MLP layers. For example, *OvSGTR* (*w.* Swin-T) achieves a comparable, even better result than VS^3 (*w.* Swin-L). At the same time, our model has fewer trainable parameters (41M vs. 124M) and lower inference latency (0.13 s vs. 0.24 s).

OvD-SGG Benchmark. Following previous works [10, 48], the *OvD-SGG* setting requires the model cannot see novel object categories during training. Specifically, 70% selected object categories of VG150 are regarded as base categories, and the remaining 30% object categories are acted as novel categories. The

Table 4: Experimental results of *OvD+R-SGG* setting on VG150 test set. † refers to *w.o.* distillation.

Method	Joint Base+Novel		Novel (Object)		Novel (Relation)	
	R@50	R@100	R@50	R@100	R@50	R@100
IMP [41]	0.77	0.94	0.00	0.00	0.00	0.00
MOTIFS [46]	1.00	1.12	0.00	0.00	0.00	0.00
VCTREE [35]	1.04	1.17	0.00	0.00	0.00	0.00
TDE [34]	1.00	1.15	0.00	0.00	0.00	0.00
VS^3 [48] (Swin-T)	5.88	7.20	6.00	7.51	0.00	0.00
OvSGTR$_{Swin-T}$†	7.88	10.06	6.82	9.23	0.00	0.00
OvSGTR$_{Swin-T}$	**13.53**	**16.36**	**14.37**	**17.44**	**9.20**	**11.19**
OvSGTR$_{Swin-B}$†	11.23	14.21	13.27	16.83	1.78	2.57
OvSGTR$_{Swin-B}$	**17.11**	**21.02**	**17.58**	**21.72**	**14.56**	**18.20**

experiments under this setting are as same as *Closed-set SGG* except that novel object categories are removed in labels. After excluding unseen object nodes, the training set of VG150 contains 50,107 images. We report the performance of *OvD-SGG* setting in terms of "Base+Novel (Object)" and " Novel (Object)" in Tab. 2. It can be found that the proposed model significantly excel previous methods. Compared to VS^3 [48], the performance gain on novel categories is up to 19.6% / 20.8% for R@50 / R@100, which demonstrate the proposed model has more powerful open vocabulary-aware and generalization ability. Since the *OvD-SGG* setting only removes nodes with novel object categories, learning process of relations will not be affected; This indicates that the performance is more dependent on the open-vocabulary ability of an object detector.

OvR-SGG Benchmark. Different from *OvD-SGG* which removes all unseen nodes , *OvR-SGG* only removes all unseen edges but keep original nodes. Considering VG150 has 50 relation categories, we randomly select 15 of them as unseen (novel) relation categories. During training, only base relation annotation is available. After removing unseen edges, there exists 44,333 images of VG150 for training. Similar as *OvD-SGG*, Tab. 3 reports the performance of *OvR-SGG* in terms of "Base+Novel (Relation)" and "Novel (Relation)". From Tab. 3, the proposed *OvSGTR* notably outperforms other competitors even without distillation. However, a marked decline in performance is observed across all techniques, inclusive of *OvSGTR* without distillation, within the "Novel (Relation)" categories, underscoring the intrinsic difficulties associated with discerning novel relations in the *OvR-SGG* paradigm. Nevertheless, with visual-concept retention, the performance of OvSGTR (*w.* Swin-T) on novel relations has been significantly improved from 0.34 (R@50) to 13.45 (R@50).

OvD+R-SGG Benchmark. This benchmark augments the SGG from a closed-set setting to a fully open vocabulary domain, where both novel object and relation categories are omitted during the training phase. For its construction, we combine the split of *OvD-SGG* and *OvR-SGG* and use their base object categories and base relation categories, resulting in 36,425 images of VG150 for training. We report the performance of *OvD+R-SGG* in Tab. 4 regarding

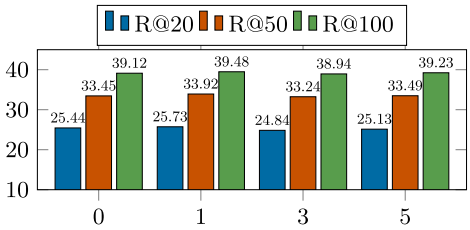

Fig. 3: Ablation study of relation queries on VG150 validation set (*Closed-set SGG*).

"Joint Base+Novel" (*i.e.*, all object and relation categories considered), "Novel (Object)" (*i.e.*, only novel object categories considered), and "Novel (Relation)" (*i.e.*, only relation categories considered). From Tab. 4, the catastrophic forgetting still occurred in *OvD+R-SGG* as same as *OvR-SGG*, which is alleviated by visual-concept retention in a significant degree. When juxtaposed with other methods, our model achieves significant performance gain on all metrics.

Overall Analysis. Experimental results present distinct challenges and difficulties in these four settings. Based on these experiments, **1)** many previous methods rely on a two-stage object detector, Faster R-CNN [31], and complicated message-passing mechanism. Nevertheless, our model showcases that a one-stage DETR-based framework can significantly surpass R-CNN-like architecture even with only one MLP to obtain feature representation for relations. **2)** previous methods with a closed-set object detector struggle to discern objects without textual information under the object-involved open vocabulary SGG (*i.e.*, *OvD-SGG* and *OvD+R-SGG*). **3)** the performance drop compared to previous settings reveals that *OvD+R-SGG* is much more challenging than others, indicating much room for extensive exploration toward fully open vocabulary SGG.

4.3 Ablation Study

Effect of Relation Queries. We first consider remove relation query embedding. The relation feature is given by $\boldsymbol{e}_{s_i \to o_j} = f_\theta([\boldsymbol{v}_{s_i}, \boldsymbol{v}_{o_j}])$, which only encodes hidden features for the subject and object node. Further, we extend the Sec. 3.1 to a more general form as $\boldsymbol{e}_{s_i \to o_j} = \frac{1}{M}\sum_{n=1}^{M} f_\theta([\boldsymbol{v}_{s_i}, \boldsymbol{v}_{o_j}, \boldsymbol{r}_n])$, which averages multiple relation query results. As shown in Fig. 3, the model achieves the best performance when the number of relation queries is set to 1. This can be interpreted from two aspects. On the one hand, the relation queries interact with all edges during training, which captures global information for the whole dataset. On the other hand, increasing the number of relation-aware queries does not introduce specific supervision yet heavy the optimization burden.

Relation-aware Pre-training. We compare *OvSGTR* trained on image caption data with others in Tab. 5. From the result, the *OvSGTR* (*w.* Swin-T) with COCO captions outperforms others, scoring 6.61, 8.92, and 10.90 for R@20, R@50, and R@100, respectively. When integrated with COCO Captions [3],

Table 5: Comparison with others trained on image captions (which is referred to as *language-supervised SGG* in [48]). All models are trained on image-caption data and test on VG150 test set directly. Our models trained on COCO Captions are used as the pre-training models for *OvR-SGG* and *OvD+R-SGG* settings.

SGG model	Training Data	Grounding	R@20	R@50	R@100
LSWS [44]	COCO	-	-	3.28	3.69
MOTIFS [46]	COCO	Li *et al.* [19]	5.02	6.40	7.33
Uniter [4]	COCO	SGNLS [49]	-	5.80	6.70
Uniter [4]	COCO	Li *et al.* [19]	5.42	6.74	7.62
VS^3 [48] (Swin-T)	COCO	GLIP-L [15]	5.59	7.30	8.62
VS^3 [48] (Swin-L)	COCO	GLIP-L [15]	6.04	8.15	9.90
VS^3 [48] (Swin-L)	VG Caption	GLIP-L [15]	10.98	15.51	19.75
Ours (Swin-B)	VG Caption	GLIP-L [15]	16.36	22.14	26.20
Ours (Swin-T)	COCO	Grounding DINO-B [23]	**6.61**	**8.92**	**10.90**
Ours (Swin-B)	COCO	Grounding DINO-B [23]	**6.88**	**9.30**	**11.48**
Ours (Swin-T)	COCO+ Flickr30k+SBU	Grounding DINO-B [23]	**7.01**	**9.43**	**11.43**

Table 6: Impact of hyper-parameter λ for distillation loss on VG150 validation set under the setting of *OvR-SGG*. $a \rightarrow b$ refers to the performance shift from a (initial checkpoint's performance) to b during training.

λ	Base+Novel		Novel	
	R@50	R@100	R@50	R@100
0	7.25 → 13.74	8.98 → 16.11	10.78 → 0.32	13.24 → 0.38
0.1	7.25 → **16.00**	8.98 → **19.20**	10.78 → **11.54**	13.24 → **13.94**
0.3	7.25 → 14.35	8.98 → 17.04	10.78 → 10.71	13.24 → 12.71
0.5	7.25 → 13.34	8.98 → 16.08	10.78 → 10.90	13.24 → 13.22

Flickr30k [29], and SBU Captions [28] , its performance peaks at 7.01, 9.43, and 11.43 for the respective metrics. The results clearly indicate the effectiveness of the proposed method, particularly when using the more lightweight Swin-B backbone compared to Swin-L; For reference, the zero-shot performance on COCO validation set of GLIP-L [15] (*w.* Swin-L) and Grounding DINO-B (*w.* Swin-B) [23] stands at 49.8 AP and 48.4 AP respectively.

Hyper-parameter λ for Distillation. Tab. 6 illustrates the impact of varying hyper-parameter λ . From the results, when $\lambda = 0.1$, the model with distillation achieves the best performance. By contrast, without distillation, a significant decline in performance for novel categories exists, showing the model struggles to retain knowledge inherited from pre-trained models for novel categories.

4.4 Visualization and Discussion

We present qualitative results of our model trained under *OvD+R-SGG* setting as well as *Closed-set SGG* setting, as shown in Fig. 4. From the figure, the model trained on *Closed-set SGG* tends to generate more dense scene graphs as the whole object and relationship categories are available during training. Despite

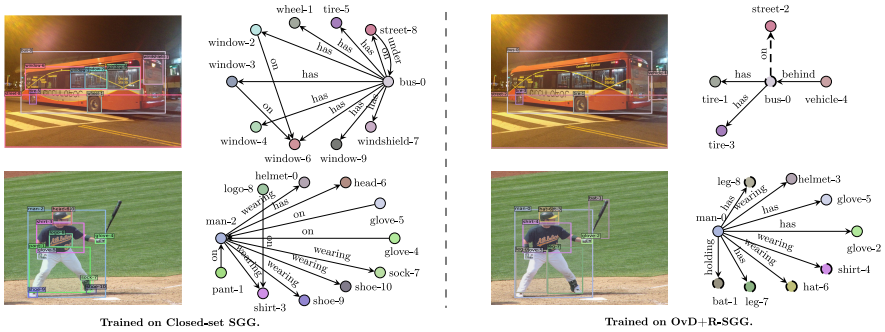

Fig. 4: Qualitative results of our model on VG150 test set (best view in color). For clarity, we only show triplets with high confidence in top-20 predictions. Dashed nodes or arrows refer to novel object categories or novel relationships.

lacking full supervision of novel categories, the model trained on *OvD+R-SGG* still can recognize novel objects like "bus", "bat" (which does not exist in VG150 dataset), and novel relationship like "on'.

Limitations & Future works. One latent limitation of this work is that we utilize an off-the-shelf language parser [25] to parse triplets from the caption. The accuracy of the parser will have a significant impact on the pre-training phase. Recently, LLM (large language model) has gained much attention. The naive parser can be replaced with a LLM to provide more accurate triplets. Moreover, it is worth discussing *Can LLMs benefit the SGG task with fewer manual annotations?* or *Can structured representations like scene graphs benefit for LLMs to alleviate hallucination?* In the future, we will try to answer these two questions.

5 Conclusion

This work advances the SGG task from a closed set to a fully open vocabulary setting based on the node and edge properties, categorizing SGG scenarios into four distinct settings including *Closed-SGG*, *OvD-SGG*, *OvR-SGG*, and *OvD+R-SGG*. Towards fully open vocabulary SGG, we design a unified framework named *OvSGTR* with transformers. The proposed framework learns to align visual features and concept information with not only base objects, but also relation categories and generalize on both novel object and relation categories. To obtain a transferable representation for relations, we utilize image-caption data as a weak supervision for relation-aware pre-training. In addition, visual-concept retention via knowledge distillation is adopted for alleviating the catastrophic forgetting problem in relation-involved open vocabulary SGG. We conduct extensive experiments on the VG150 benchmark dataset and have set up new state-of-the-art performances for all settings.

Acknowledgements

This research was supported in part by the Hong Kong Research Grants Council (GRF-15229423), the Chinese National Natural Science Foundation Projects (U23B2054, 62306313), and the InnoHK program.

References

1. Chen, S., Jin, Q., Wang, P., Wu, Q.: Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In: CVPR. pp. 9959–9968 (2020)
2. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: CVPR. pp. 6163–6171 (2019)
3. Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. CoRR **abs/1504.00325** (2015)
4. Chen, Y., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: universal image-text representation learning. In: ECCV. pp. 104–120 (2020)
5. Chiou, M., Ding, H., Yan, H., Wang, C., Zimmermann, R., Feng, J.: Recovering the unbiased scene graphs from the biased ones. In: ACMMM. pp. 1581–1590 (2021)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. pp. 4171–4186 (2019)
7. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: CVPR. pp. 14064–14073 (2022)
8. Ghiasi, G., Gu, X., Cui, Y., Lin, T.: Scaling open-vocabulary image segmentation with image-level labels. In: ECCV. pp. 540–557 (2022)
9. Gu, J., Joty, S.R., Cai, J., Zhao, H., Yang, X., Wang, G.: Unpaired image captioning via scene graph alignments. In: ICCV. pp. 10322–10331 (2019)
10. He, T., Gao, L., Song, J., Li, Y.: Towards open-vocabulary scene graph generation with prompt-based finetuning. In: ECCV. pp. 56–73 (2022)
11. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: CVPR. pp. 1219–1228 (2018)
12. Kenfack, F.K., Siddiky, F.A., Balint-Benczedi, F., Beetz, M.: Robotvqa - A scene-graph- and deep-learning-based visual question answering system for robot manipulation. In: IROS. pp. 9667–9674 (2020)
13. Knyazev, B., de Vries, H., Cangea, C., Taylor, G.W., Courville, A.C., Belilovsky, E.: Generative compositional augmentations for scene graph prediction. In: ICCV. pp. 15807–15817 (2021)
14. Lee, S., Kim, J., Oh, Y., Jeon, J.H.: Visual question answering over scene graph. In: GC. pp. 45–50 (2019)
15. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J., Chang, K., Gao, J.: Grounded language-image pre-training. In: CVPR. pp. 10955–10965 (2022)
16. Li, R., Zhang, S., He, X.: Sgtr: End-to-end scene graph generation with transformer. In: CVPR. pp. 19464–19474 (2022)
17. Li, R., Zhang, S., Lin, D., Chen, K., He, X.: From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In: CVPR. pp. 28076–28086 (2024)

18. Li, R., Zhang, S., Wan, B., He, X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In: CVPR. pp. 11109–11119 (2021)
19. Li, X., Chen, L., Ma, W., Yang, Y., Xiao, J.: Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In: ACMMM. pp. 4204–4213 (2022)
20. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2999–3007 (2017)
21. Lin, X., Ding, C., Zhan, Y., Li, Z., Tao, D.: Hl-net: Heterophily learning network for scene graph generation. In: CVPR. pp. 19454–19463 (2022)
22. Liu, H., Yan, N., Mortazavi, M.S., Bhanu, B.: Fully convolutional scene graph generation. In: CVPR. pp. 11546–11556 (2021)
23. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. CoRR **abs/2303.05499** (2023)
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 9992–10002 (2021)
25. Mao, J.: Scene graph parser. `https://github.com/vacancy/SceneGraphParser` (2022)
26. Nguyen, K., Tripathi, S., Du, B., Guha, T., Nguyen, T.Q.: In defense of scene graphs for image captioning. In: ICCV. pp. 1387–1396 (2021)
27. Nuthalapati, S.V., Chandradevan, R., Giunchiglia, E., Li, B., Kayser, M., Lukasiewicz, T., Yang, C.: Lightweight visual question answering using scene graphs. In: CIKM. pp. 3353–3357 (2021)
28. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: NeurIPS. pp. 1143–1151 (2011)
29. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: ICCV. pp. 2641–2649 (2015)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
31. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. NeurIPS **28** (2015)
32. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I.D., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–666 (2019)
33. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G.G., Sigal, L.: Energy-based learning for scene graph generation. In: CVPR. pp. 13936–13945 (2021)
34. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: CVPR. pp. 3713–3722 (2020)
35. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: CVPR. pp. 6619–6628 (2019)
36. Teney, D., Liu, L., van den Hengel, A.: Graph-structured representations for visual question answering. In: CVPR. pp. 3233–3241 (2017)
37. Wang, D., Beck, D., Cohn, T.: On the role of scene graphs in image captioning. In: LANTERN@EMNLP-IJCNLP. pp. 29–34 (2019)
38. Wang, M., Xing, J., Liu, Y.: Actionclip: A new paradigm for video action recognition. CoRR **abs/2109.08472** (2021)

39. Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., et al.: Towards open vocabulary learning: A survey. IEEE TPAMI (2024)
40. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. In: CVPR. pp. 15254–15264 (2023)
41. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: CVPR. pp. 3097–3106 (2017)
42. Yang, L., Huang, Z., Song, Y., Hong, S., Li, G., Zhang, W., Cui, B., Ghanem, B., Yang, M.: Diffusion-based scene graph to image generation with masked contrastive pre-training. CoRR **abs/2211.11138** (2022)
43. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: CVPR. pp. 10685–10694 (2019)
44. Ye, K., Kovashka, A.: Linguistic structures as weak supervision for visual scene graph generation. In: CVPR. pp. 8289–8299 (2021)
45. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.: Open-vocabulary object detection using captions. In: CVPR. pp. 14393–14402 (2021)
46. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: CVPR. pp. 5831–5840 (2018)
47. Zhang, J., Shih, K.J., Elgammal, A., Tao, A., Catanzaro, B.: Graphical contrastive losses for scene graph parsing. In: CVPR. pp. 11535–11543 (2019)
48. Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., Chen, C.W.: Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space. In: CVPR. pp. 2915–2924 (2023)
49. Zhong, Y., Shi, J., Yang, J., Xu, C., Li, Y.: Learning to generate scene graph from natural language supervision. In: ICCV. pp. 1823–1834 (2021)
50. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., Gao, J.: Regionclip: Region-based language-image pretraining. In: CVPR. pp. 16772–16782 (2022)
51. Zhu, C., Chen, L.: A survey on open-vocabulary detection and segmentation: Past, present, and future. CoRR **abs/2307.09220** (2023)
52. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. In: ICLR (2021)