# Learning to Generate Language-supervised and Open-vocabulary Scene Graph using Pre-trained Visual-Semantic Space

Yong Zhang[†], Yingwei Pan[‡], Ting Yao[‡], Rui Huang[†*], Tao Mei[‡], and Chang-Wen Chen[§]
[†] The Chinese University of Hong Kong, Shenzhen [‡] HiDream.ai Inc. [§] The Hong Kong Polytechnic University
yongzhang@link.cuhk.edu.cn, {panyw.ustc, tingyao.ustc}@gmail.com, ruihuang@cuhk.edu.cn,
tmei@hidream.ai, changwen.chen@polyu.edu.hk

## Abstract

*Scene graph generation (SGG) aims to abstract an image into a graph structure, by representing objects as graph nodes and their relations as labeled edges. However, two knotty obstacles limit the practicability of current SGG methods in real-world scenarios: 1) training SGG models requires time-consuming ground-truth annotations, and 2) the closed-set object categories make the SGG models limited in their ability to recognize novel objects outside of training corpora. To address these issues, we novelly exploit a powerful pre-trained visual-semantic space (VSS) to trigger language-supervised and open-vocabulary SGG in a simple yet effective manner. Specifically, cheap scene graph supervision data can be easily obtained by parsing image language descriptions into semantic graphs. Next, the noun phrases on such semantic graphs are directly grounded over image regions through region-word alignment in the pre-trained VSS. In this way, we enable open-vocabulary object detection by performing object category name grounding with a text prompt in this VSS. On the basis of visually-grounded objects, the relation representations are naturally built for relation recognition, pursuing open-vocabulary SGG. We validate our proposed approach with extensive experiments on the Visual Genome benchmark across various SGG scenarios (i.e., supervised / language-supervised, closed-set / open-vocabulary). Consistent superior performances are achieved compared with existing methods, demonstrating the potential of exploiting pre-trained VSS for SGG in more practical scenarios.*

## 1. Introduction

Scene graph [10] is a structured representation for describing image semantics. It abstracts visual objects as graph nodes and represents their relations as labeled graph edges. The task of scene graph generation (**SGG**) [6, 14, 20, 26, 40, 47, 48, 50, 51, 57, 60, 63, 64] plays an important role for fine-grained visual understanding, which has shown promising results in facilitating various downstream applications, such as image-text retrieval [24, 38, 49], image captioning [2, 22, 32, 35, 52, 54, 55, 66], cross-media knowledge graph construction [18, 45] and robot planning [1].

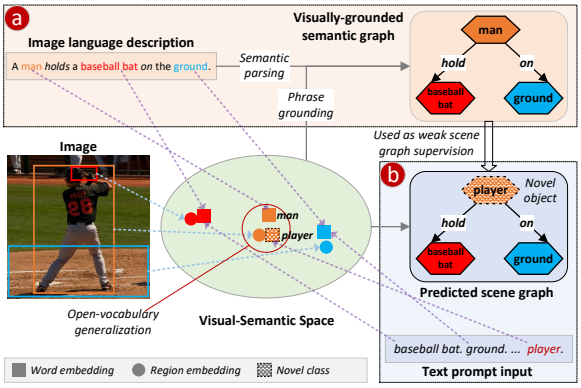


Figure 1. An illustration of exploiting a pre-trained visual-semantic space (VSS) to trigger language-supervised and open-vocabulary scene graph generation (SGG). (a) We acquire weak scene graph supervision by semantically parsing the image language description and grounding noun phrases on image regions via VSS. (b) At SGG inference time, thanks to the open-vocabulary generalization naturally rooted in VSS, the novel object name (e.g., player) in the text prompt input can be well aligned to one image region, which is regarded as its detection.

Though great effort has been made, SGG of the current stage still faces two knotty obstacles that limit its practicability in real-world scenarios. 1) Training SGG models requires massive ground-truth scene graphs that are expensive for manual annotation. Annotators have to draw bounding boxes for all objects in an image and connect possible interacted object pairs, and assign object/relation labels. Since assigned labels might be ambiguous, further verification and canonicalization processing are usually required [14]. Finally, a scene graph in the form of a set of $\langle subject, predicate, object \rangle$ triplets with subject and object bounding boxes is constructed. Such annotating pro-

*Corresponding author

cess is time-consuming and tedious, costing much human labor and patience. 2) Almost all existing SGG methods [20,21,26,47,48,50,51,60] involve a pre-defined closed set of object categories, making them limited in recognizing novel objects outside of training corpora. However, real-world scenes contain a boarder set of visual concepts than any pre-defined category pool. It is very likely to encounter unseen/novel categories. When this happens, current SGG models either classify novel objects to a known category or fail to detect them like background regions. Accordingly, the prediction of their interactions/relations with other objects is negatively affected or just neglected. This may lead to problems. For example, a real-world robot may take inappropriate actions using such closed-set SGG models [1,42].

Recently, there is a trend of leveraging free-form language supervision for benefiting visual recognition tasks via large-scale language-image pre-training [7, 15, 17, 36, 53, 59, 67]. These methods (e.g., CLIP [36]) perform pre-training on massive easily-obtained image-text pairs to learn a visual-semantic space (VSS), and have demonstrated great zero-shot transferability. Especially, the recent grounded language-image pre-training (GLIP) [17] has learned an object-level and semantic-rich VSS. Based on the learned VSS, it has established new state-of-the-art performances in phrase grounding and zero-shot object detection. This indicates such pre-trained VSS has powerful multi-modal alignment ability (i.e., image regions and text phrases that have similar semantics get close embeddings) and open-vocabulary generalization ability (i.e., covering virtually any concepts in the pre-training image-text corpus). This inspires our thought of addressing the aforementioned obstacles in SGG using the pre-trained VSS. On the one hand, taking advantage of its multi-modal alignment ability, we can cheaply acquire scene graph supervision from an image description (e.g., retrieving image regions aligned with noun phrases and re-arranging the description into a scene-graph-like form). On the other hand, by leveraging its open-vocabulary generalization ability, it is promising to enable novel category prediction in SGG.

In this work, we investigate the opportunity of fully exploiting the VSS learned by language-image pre-training to trigger language-supervised and open-vocabulary SGG. Specifically, we obtain weak scene graph supervision by semantically parsing an image language description into a semantic graph, then grounding its noun phrases over image regions through region-word alignment in the pre-trained VSS (Figure 1 (a)). Moreover, we propose a novel SGG model, namely Visual-Semantic Space for Scene graph generation (**VS$^3$**). It takes a raw image and a text prompt containing object category names as inputs, and projects them into the shared VSS as embeddings. Next, VS$^3$ performs object detection by aligning the embeddings of category names and image regions. Based on high-confidence detected objects, VS$^3$ builds relation representations for object pairs with a devised relation embedding module that fully mines relation patterns from visual and spatial perspectives. Finally, a relation prediction module takes relation representations to infer relation labels. The predicted scene graph is composed by combining object detections and inferred relation labels. During training, visually-grounded semantic graphs parsed from image descriptions could be used as weak scene graph supervision, achieving language-supervised SGG. At SGG inference time, when using a text prompt input containing novel categories, VS$^3$ manages to detect novel objects thanks to the open-vocabulary generalization ability natually rooted in VSS, hence allowing for open-vocabulary SGG (Figure 1 (b)).

In summary, we have made the following contributions: (1) the exploitation of a pre-trained VSS provides an elegant solution for addressing obstacles to triggering both language-supervised and open-vocabulary SGG, making a solid step toward real-world usage of SGG. (2) The proposed VS$^3$ model is a new and versatile framework, which effectively transfers language-image pre-training knowledge for benefiting SGG. (3) We fully validate the effectiveness of our approach through extensive experiments on the Visual Genome benchmark, and have set new state-of-the-art performances spanning across all settings (i.e., supervised / language-supervised, closed-set / open-vocabulary).

## 2. Related Work

**Fully supervised SGG.** The concept of scene graph as a structured image representation is first introduced in [10]. Next, the Visual Genome benchmark [14] is manually annotated with large-scale scene graphs on images. Such annotated dataset triggers a series of innovations [6,20,26,40,47, 48, 50, 51, 57, 60] for the fully supervised SGG task. Typically, an object detector (e.g., Faster-RCNN [37]) is trained to retrieve image regions as scene graph nodes. Then, relation representations of object pairs are constructed from visual, spatial and language perspectives, and are used for relation classification to label scene graph edges. To achieve desirable SGG, researchers have devised message-passing mechanisms [20, 26, 48, 50, 60] to exploit contextual information, derived contrastive loss functions [62] or incorporated external knowledge [6, 40, 57]. However, all these methods rely on training with expensive scene graph annotations. Our proposed VS$^3$ model is compatible with fully supervised SGG, but we seek to make SGG training cheaper, which is more practical in real-world applications.

**Language-supervised SGG.** This task aims to train SGG models using language descriptions. It has recently attracted increasing attention [21,41,56,58,65], which is also referred as weakly supervised SGG in [21, 41, 58]. Particularly, based on a graph alignment algorithm, VSPNet [58] first proposes to supervise SGG training with scene graphs

that have no object locations. Subsequent works [21,41,65] extract entities and relations from image captions to compose such unlocalized scene graph, which is achieved via an off-the-shelf language parser [31, 39]. They next follow a common paradigm: first grounding text entities on image regions, and then leveraging grounded scene graphs as pseudo labels to train standard SGG models. To acquire entity groundings, Shi *et al* [41] devise an efficient graph matching module optimized via contrastive learning; Zhong *et al* [65] simply match text entity names with predicted object labels from a pre-trained object detector using semantic rules such as WordNet [33] synsets matching. More recently, Li *et al* [21] integrate interaction-aware knowledge distilled from pre-trained language-image models [16] for enhancing grounding reliability. Instead, we propose to obtain groundings through region-word alignment in a pre-trained VSS, which is much simple yet more effective to collect scene graph supervision from language.

**Language-image pre-training.** This has been shown effective for boosting various vision-language downstream tasks [11, 16, 23, 29, 30, 34, 43], e.g., image-text retrieval, image captioning. Also, recent studies present remarkable results on transferring pre-trained language-image knowledge to solve vision recognition problems, such as zero-shot image classification [9,36], open-vocabulary object detection [7, 17, 53, 59, 67] and zero-shot semantic segmentation [15]. For example, CLIP [36] and ALIGN [9] learn separate encoders to embed image and text into a shared space by pre-training on massive image-text pairs using a contrastive loss. They have demonstrated remarkable generalization ability on zero-shot image classification after pre-training. Distinct from CLIP and ALIGN that learn image-level representations, GLIP [17] focuses on learning object-level visual representations through region-word alignment. It has attained strong zero-shot and few-shot transferability to various object-level recognition tasks such as object detection and phrase grounding. Most recently, He *et al* [8] investigate a visual-relation pre-training and prompt-based fine-tuning method for open-vocabulary SGG. However, its image encoder relies on a pre-trained region proposal extractor, which is a bottleneck for achieving open-vocabulary SGG under the SGDET protocol. Unlike [8], our proposed VS$^3$ directly encodes an image into region tokens, avoiding the bottleneck of region proposals. More importantly, our approach addresses obstacles to achieving both language-supervised and open-vocabulary SGG using a unified framework, while He *et al* [8] only focus on the latter.

# 3. Approach

## 3.1. Notation & Overview

The task of scene graph generation (SGG) aims to map an image into an abstract graph $SG = \{O, R\}$, where graph nodes $O = \{o_1, ..., o_N\}$ correspond to image objects, and graph edges $R = \{r_1, ..., r_M\}$ represents their relations. Each object $o_i = \{\mathbf{b}_i, l_i\} \in O$ contains the bounding box coordinates $\mathbf{b}_i \in \mathbb{R}^4$ and its class label information $l_i \in \mathcal{C}_o$, where $\mathcal{C}_o$ denotes the set of object categories. Each relation $r_m \in R$ is a $\langle subject, predicate, object \rangle$ triplet, and we represent it as $r_m = r_{i \to j} = \{o_i, p_{ij}, o_j\}$, in which $p_{ij}$ is the predicate/relation label belonging to category set $\mathcal{C}_r$.

| $SG$ \ $\mathcal{C}_o^{target}$ | **Not containing novel object classes** | **Containing novel object classes** |
|---|---|---|
| **Manually annotated** | fully supervised & closed-set | fully supervised & open-vocabulary |
| **Automatically parsed from image descriptions** | language-supervised & closed-set | language-supervised & open-vocabulary |

Table 1. Definitions of different SGG settings, according to scene graph supervision $SG$ and the object set at inference $\mathcal{C}_o^{target}$.

Most existing SGG methods require expensive manually annotated scene graphs $SG$ as supervision. And they involve a closed set of object categories $\mathcal{C}_o$ in both training and inference. These issues limit SGG for practical usage. In this work, we propose to fully exploit a pre-trained VSS to push SGG towards language-supervised and open-vocabulary scenarios. We illustrate the definitions of different SGG settings in Table 1. Concretely, from the perspective of scene graph supervision $SG$, SGG is categorized into fully supervised and language-supervised, using $SG$ from manual annotation and language respectively. From the other perspective of object categories $\mathcal{C}_o^{target}$ at inference, it is referred to as open-vocabulary or closed-set according to whether or not $\mathcal{C}_o^{target}$ contains novel objects.

Next, in Section 3.2, we present a new SGG model named VS$^3$, which is versatile to handle with all SGG settings in Table 1. In Section 3.3, we devise a scheme to obtain scene graph supervision from language descriptions, allowing for language-supervised SGG. Finally, Section 3.4 details the strategy of transferring the proposed VS$^3$ to pursue open-vocabulary SGG.

## 3.2. The Proposed VS$^3$ Model

We propose the VS$^3$ model for tackling the SGG task by extending the GLIP [17] framework with relation recognition modules, as shown in Figure 2.

**Preliminary.** GLIP unifies object detection and phrase grounding into one framework. It has an image encoder $\text{Enc}_I$ (e.g., Swin Transformer backbone [28]) and a text encoder $\text{Enc}_L$ (e.g., BERT [12]). $\text{Enc}_I$ extracts region/box features $\bar{O} \in \mathbb{R}^{\bar{N} \times d}$ from an input image, where $\bar{N}$ is the number of regions and $d$ is the feature dimension. $\text{Enc}_L$ encodes a text input into contextualized word/token embeddings $\bar{P} \in \mathbb{R}^{\bar{T} \times d}$, where $\bar{T}$ is the text length. Further, GLIP uses a cross-modal fusion module to achieve feature communication between $\bar{O}$ and $\bar{P}$, resulting in enriched region embeddings $\tilde{O} \in \mathbb{R}^{\bar{N} \times d}$ and word embeddings $\tilde{P} \in \mathbb{R}^{\bar{T} \times d}$. Finally, the region-word alignment

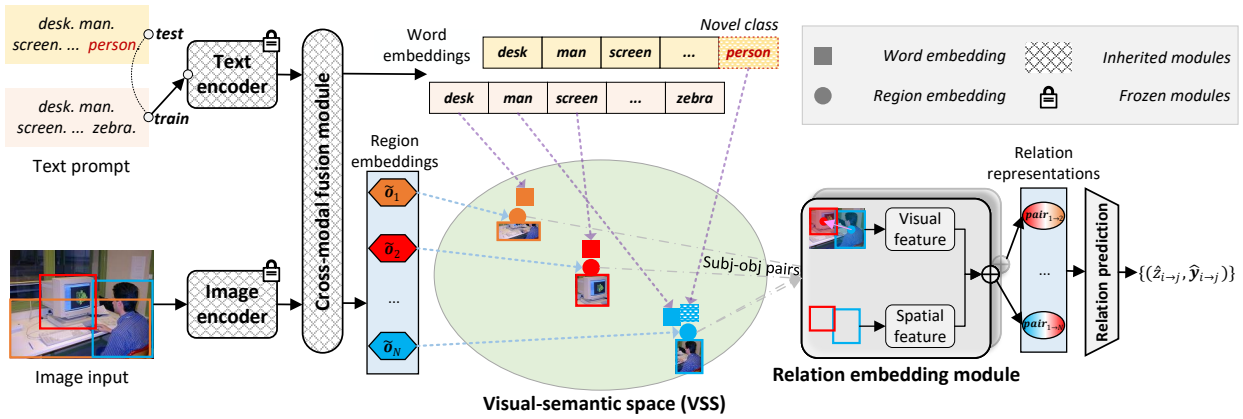


Figure 2. An overview of the proposed Visual-Semantic Space for Scene graph generation (VS$^3$) model. It inherits the image encoder, the text encoder and the cross-modal fusion module from GLIP [17], so as to project image regions and text prompt words in a pre-trained visual-semantic space (VSS). Object regions are detected by aligning the embeddings of category names and image regions in VSS. Next, high-confidence detected results are retained to compose subject-object pairs. After that, the relation embedding module constructs their relation representations by extracting visual and spatial features, on which relation prediction is performed. At test time, thanks to the open-vocabulary generalization ability of VSS, VS$^3$ manages to detect novel objects by switching to a text prompt containing novel classes.

scores $\hat{S}_{ground} = \tilde{O}\tilde{P}^\top \in \mathbb{R}^{\bar{N}\times\bar{T}}$ and the predicted locations $\hat{B} = box\_predictor(\tilde{O}) \in \mathbb{R}^{\bar{N}\times 4}$ are supervised by ground-truth text grounding data. After large-scale pre-training, $\text{Enc}_I$ and $\text{Enc}_L$ embed the input image and text into a joint VSS, which aligns multi-modal embeddings and covers open-vocabulary concepts. VS$^3$ inherits $\text{Enc}_I$, $\text{Enc}_L$ and the cross-module fusion module from GLIP.

**Text prompt.** Considering that object detection has been reformulated as phrase grounding, VS$^3$ also requires a text prompt input except for the image input. Following GLIP's design, we set the text prompt for object detection in the form of "$name(c_1)$. $name(c_2)$. ... $name(c_{|\mathcal{C}_o|})$.", where $c_i \in \mathcal{C}_o$ and $name(c_i)$ gets the category name of $c_i$ (e.g., $person$). Hence, an object is detected according to the alignment score between a region embedding $\tilde{\mathbf{o}}_i \in \tilde{O}$ (the $i$th row) and the category name embeddings $\tilde{P}$.

**Relation embedding module.** To further enable VS$^3$ with relation recognition ability, we devise the relation embedding module to build relation representations. Based on the region/box features $\tilde{O}$ after cross-modal fusion, we first sample a subset of regions $\tilde{O}' \in \mathbb{R}^{N'\times d}$ that are most likely to be valid objects. This is achieved by matching predicted bounding boxes $\hat{B}$ with ground-truth objects during training, and by retaining top-$N'$ regions with the highest confidence scores after non-maximum suppression (NMS) at inference. Next, we construct relation representations for all possible subject-object pairs. Given an object pair $(\tilde{\mathbf{o}}_i, \tilde{\mathbf{o}}_j)$ and their normalized bounding boxes $(\mathbf{b}_i, \mathbf{b}_j)$, the pairwise relation representation is represented as $\mathbf{pair}_{i\to j} = cat[\mathbf{pair}^{visual}_{i\to j}, \mathbf{pair}^{spatial}_{i\to j}]$. This is the concatenation of features mined from the visual and spatial perspectives. The visual feature is computed by

$$\mathbf{pair}^{visual}_{i\to j} = \boldsymbol{f}_{diff}(\tilde{\mathbf{o}}_i - \tilde{\mathbf{o}}_j) + \boldsymbol{f}_{sum}(\tilde{\mathbf{o}}_i + \tilde{\mathbf{o}}_j), \quad (1)$$

where $\boldsymbol{f}_{diff}$ and $\boldsymbol{f}_{sum}$ are two mapping functions implemented as 2-layer MLPs (multi-layer perceptron). By defining the normalized center coordinates of two involved objects as $(ct^x_i, ct^y_i)$ and $(ct^x_j, ct^y_j)$, the spatial feature is measured as

$$\mathbf{pair}^{spatial}_{i\to j} = cat[\mathbf{b}_i, \mathbf{b}_j, dx, dy, dis, \theta, A_i, A_j, I, U], \quad (2)$$

where $dx = ct^x_i - ct^x_j, dy = ct^y_i - ct^y_j, dis = \sqrt{dx^2 + dy^2}, \theta = arctan(\frac{dy}{dx})$. $A_i, A_j, I, U$ denote the areas of the subject, the object, their intersection, and union boxes, respectively.

**Relation prediction.** Conditioned on the relation representation $\mathbf{pair}_{i\to j}$ of each object pair, we predict a relateness score $\hat{z}_{i\to j} = f_{relateness}(\mathbf{pair}_{i\to j}) \in [0, 1]$ and a semantic label probability $\hat{\boldsymbol{y}}_{i\to j} = \boldsymbol{f}_{semantic}(\mathbf{pair}_{i\to j}) \in [0, 1]^{|\mathcal{C}_r|}$. The relateness $\hat{z}_{i\to j}$ represents the probability that relations exist between the object pair. $f_{relateness}$ is implemented with an MLP coupled with Sigmoid activation. $\boldsymbol{f}_{semantic}$ is implemented with another MLP using Softmax activation. The total loss for relation recognition $L_{rel\_rcg}$ is measured as

$$L_{relateness} = FL(\hat{z}_{i\to j}, z_{i\to j}), \quad (3)$$
$$L_{semantic} = CE(\hat{\boldsymbol{y}}_{i\to j}, \boldsymbol{y}_{i\to j}), \quad (4)$$
$$L_{rel\_rcg} = L_{relateness} + L_{semantic}, \quad (5)$$

where $FL$ and $CE$ denote focal loss [25] and cross-entropy loss functions. $z_{i\to j}$ and $\boldsymbol{y}_{i\to j}$ represent the ground-truth relateness label and predicate category label respectively.

**Training & inference.** During training, we initialize parameters from pre-trained GLIP models for inherited modules in VS$^3$. To ease training difficulty, we freeze the image encoder and text encoder, and only fine-tune the

cross-modal fusion module and devised modules for relation recognition. This also avoids the degeneration of the pre-trained VSS. At inference, by retaining high-confidence detected objects and further predicting their relations, we generate an image scene graph representation.

### 3.3. Obtaining Language Scene Graph Supervision

Ground-truth scene graphs are time-consuming to annotate. Alternatively, we can parse semantic graphs from image language descriptions, and obtain noun phrase groundings through region-word alignment in the pre-trained VSS (implemented with an off-the-shelf GLIP). This is a much cheaper way to obtain weak scene graph supervision.

**Semantic graph parsing.** Concretely, for each image language description, we parse it into a semantic graph $SG^{text} = \{O^{text}, R^{text}\}$ using the Standard Scene Graph Parser based on [39]. The parser not only extracts noun phrases as entities/objects ($O^{text}$), but also extracts the words describing their relations ($R^{text}$). For example, the sentence "*a woman is playing the piano in the room.*" is parsed to the $SG^{text}$, of which $O^{text} = \{woman, piano, room\}$ and $R^{text} = \{\langle 0, playing, 1\rangle, \langle 0, in, 2\rangle\}$ (numbers denote object indices). Considering that parsed object/relation words are free-form, we map them to our concerning categories (e.g., VG150 object/relation categories in experiments) by rules such as direct string matching and WordNet [33] synsets matching following [65].

**Semantic graph grounding.** Note that each element of $O^{text}$ only contains a text label name so far, and its bounding box information is still missing. To obtain grounding boxes, we construct a text prompt using triplets in $SG^{text}$, e.g., "*woman playing piano. woman in room.*". Then, we feed such text prompt together with the raw image into a pre-trained GLIP, in order to acquire grounding boxes of $O^{text}$. Specifically, for each element in $O^{text}$, we select the image region that has the highest alignment score with its category name as its grounding box. Since there might be multiple objects in $O^{text}$ that actually refer to the same object, we perform a post-processing NMS to merge boxes with the same label and high IoU (intersection over union) scores ($\geq 0.9$). Finally, with box information, the visually-grounded $SG^{text}$ is ready to be used as weak supervision for training SGG models, e.g., the proposed VS$^3$.

### 3.4. Transferring to Open-vocabulary SGG

Open-vocabulary SGG [8] aims to train SGG models that can recognize objects of novel categories and their involved relations. Formally, we train the SGG model with scene graphs containing objects in the base category set $\mathcal{C}_o^{base}$. At inference, the object category set is $\mathcal{C}_o^{target}$, which contains novel categories in $\mathcal{C}_o^{novel} = \mathcal{C}_o^{target} \backslash \mathcal{C}_o^{base} \neq \emptyset$.

Back to our proposed VS$^3$, an open-vocabulary VSS is maintained by freezing the image and text encoders. Taking this advantage, we devise a scheme to adapt VS$^3$ for open-vocabulary SGG. Concretely, during training, we set the text prompt as "$name(c_1)$. $name(c_2)$. ... $name(c_{|\mathcal{C}_o^{base}|})$.", where $c_i \in \mathcal{C}_o^{base}$. And only relation triplets involving base object categories are kept for training. At inference, the text prompt is switched to be "$name(c_1)$. $name(c_2)$. ... $name(c_{|\mathcal{C}_o^{target}|})$.", where $c_i \in \mathcal{C}_o^{target}$. In this way, a novel object class (e.g., $lady$) may have an embedding close to a base category (e.g., $woman$) embedding. This makes the novel class also able to find well-aligned image regions. Note that relation representations are constructed from visual and spatial cues, which are usually class-agnostic. Hence, the following relation recognition in VS$^3$ will not be affected when encountering novel objects.

## 4. Experiments

### 4.1. Datasets and Experimental Settings

**Datasets.** To evaluate the SGG task, we adopt the widely-used ***VG150*** version [50] of the Visual Genome (VG) dataset [14]. VG150 retains the most frequent 150 categories and 50 relation/predicate categories in VG. It contains ∼108K images, of which 70% images are used for training (including 5K for validation), and the remaining 30% images are used for testing. The annotated scene graph of each image has 11.5 objects and 6.2 relation triplets on average. In addition, images of VG are densely annotated with region descriptions, about 50 descriptions for each image. We refer to these descriptions as ***VG caption***, which provides a text source for evaluating the language-supervised SGG setting. Moreover, we consider the challenging setting of using image-text pairs in ***COCO caption*** [4] for training SGG models. This dataset contains 123K images in total. Each image has 5 human-annotated captions. We keep ∼106k images by filtering out those images that also exist in the VG150 test split.

**Evaluation protocols and metrics.** We mainly adopt the SGDET [47,50] protocol, which generates a scene graph from the input image without any given box information. We report the performance on Recall@K (K=20/50/100) following previous works [21, 41, 47, 50, 56, 65], which measures the fraction of correctly predicted relation triplets in top $K$ predictions. A triplet prediction is considered as correct when its subject, object, predicate labels and both the subject and object regions match with (same label or IoU>0.5) a ground-truth triplet. Note that we obtain triplet predictions using graph constraint, which limits each subject-object pair to have only the most confident predicate. All recall metrics across different SGG settings in experiments are computed over VG150 test images. Considering that the adopted GLIP pre-trained VSS has seen part of images in the original VG150 test split (∼26k) during

pre-training, we exclude these images and get a new split of ∼15k test images. We have validated that such VG150 test split is sufficiently large for computing stable metrics as the original, by comparing computed metrics of several SGG models (in codebase [46]) on these two splits ($< 0.15$ points variation, see supplementary materials).

**Implementation details.** We initialize VS$^3$ from pre-trained GLIP [17] models, i.e., the GLIP-T and the larger GLIP-L trained with more data. Both construct a VSS of dimension $d = 256$. We retain the top 36 object detections per image for pairwise relation recognition. The whole framework is fine-tuned on 8 Nvidia 2080Ti GPUs with AdamW optimizer. During fine-tuning, we freeze the parameters of the image and text encoder; and set the learning rate for the cross-modal fusion module as 1e-5 and 10x larger learning rates for the relation embedding and prediction modules. The maximum fine-tuning epoch number is 10, with learning rates dropping by 10x after 6 epochs.

### 4.2. Fully Supervised SGG

**Setup.** We first evaluate our proposed VS$^3$ under the conventional fully supervised SGG setting. This setting trains SGG models using manually annotated scene graphs, consisting of object labels coupled with bounding boxes, and relation labels. We adopt VG150 for training and evaluation following previous methods [3, 19, 26, 27, 48, 50, 60, 61, 65]. All these methods involve a closed set of object categories. Specifically, the text prompt input of VS$^3$ is constructed from VG150 object category names, i.e., "*airplane. animal. ... zebra.*". We train VS$^3$ by fine-tuning over two GLIP variants: GLIP-T with the Swin-T [28] backbone, GLIP-L with the Swin-L [28] backbone.

**Comparison with state-of-the-arts.** The results are summarized in Table 2. Our proposed VS$^3$ model using the Swin-T backbone already achieves competitive recall metrics. When upgrading to the larger Swin-L variant, the performance improvements become significant (1.8 to 3.4 points improvement than the previous best results). Note that previous methods [26, 50, 60] build their models upon an off-the-shelf object detector, and they usually design heavy message-passing modules to incorporate context information. Instead, VS$^3$ devises a light-weighted relation recognition head (including the relation embedding and prediction modules) over a pre-trained VSS. The superior performances clearly suggest the merits of transferring language-image pre-trained models for boosting SGG.

**Ablation on relation representation.** Next, we carry out ablation studies on relation representation construction in the relation embedding module. As shown in Table 2, by removing visual and spatial feature components, the relation triplet recalls drop accordingly. Also notice that the removal of visual features which is built from subject and object region embeddings leads to relatively larger performance drops than spatial. This suggests the region embeddings in the pre-trained VSS provide strong cues for relation recognition. Overall, these observations validate the effectiveness of our design to mine relation patterns.

| SGG model | Detector | Backbone | R@20 | R@50 | R@100 |
|---|---|---|---|---|---|
| FCSGG [27] | - | HRNetW48 | 16.1 | 21.3 | 25.1 |
| SGTR [19] | DETR | R-101 | - | 24.6 | 28.4 |
| IMP [50] | Faster-RCNN | VGG-16 | 14.6 | 20.7 | 24.5 |
| KERN [3] | Faster-RCNN | VGG-16 | - | 27.1 | 29.8 |
| MOTIFS [60] | Faster-RCNN | VGG-16 | 21.4 | 27.2 | 30.3 |
| RelDN [62] | Faster-RCNN | VGG-16 | 21.1 | 28.3 | 32.7 |
| VTransE [61] | Faster-RCNN | RX-101 | 23.0 | 29.7 | 34.3 |
| MOTIFS [60] | Faster-RCNN | RX-101 | 25.1 | 32.1 | 36.9 |
| VCTREE [48] | Faster-RCNN | RX-101 | 24.7 | 31.5 | 36.2 |
| SGNLS [65] | Faster-RCNN | RX-101 | 24.6 | 31.8 | 36.3 |
| HL-Net [26] | Faster-RCNN | RX-101 | 26.0 | 33.7 | 38.1 |
| VS$^3$ | - | Swin-T | 26.1 | 34.5 | 39.2 |
| *w/o visual* | - | Swin-T | 23.1 | 31.6 | 36.7 |
| *w/o spatial* | - | Swin-T | 24.3 | 32.8 | 37.8 |
| VS$^3$ | - | Swin-L | **27.8** | **36.6** | **41.5** |

Table 2. Experimental results of fully supervised SGG. *w/o visual* and *w/o spatial* indicate removing spatial and visual features in the relation embedding module for relation representation. All metrics are computed under the SGDET protocol on VG150 test images.

### 4.3. Language-supervised SGG

**Setup.** Language-supervised SGG [21, 41, 56, 65] explores to train SGG models with language descriptions of images. Concretely, we parse each image description into a semantic graph, in the form of a set of $\langle subject, predicate, object \rangle$ triplets. Note that parsed object/relation phrases from language descriptions are free-form, we map them to VG150 categories by semantic rules following [65], such as WordNet [33] synsets matching. This makes the learned SGG model compatible for evaluating on VG150. Next, the parsed semantic graph is grounded to image regions using grounding methods, i.e., the pre-trained GLIP-L [17] in our approach. Finally, the visually-grounded semantic graphs are used as weak supervision to train our proposed VS$^3$ like the fully supervised setting.

Particularly, we have trained VS$^3$ with text triplets parsed from three different sources of text following [21, 65]. 1) The *unlocalized graph* setting uses ground-truth triplet annotations in VG. 2) The *VG caption* setting uses triplets that are automatically parsed from natural image descriptions in VG. 3) The *COCO caption* setting leverages triplets parsed from captions in COCO. This setting is the most challenging since COCO captions are image-level descriptions. Such captions are different from the region-level descriptions in VG, which focus on describing object interactions. Also, note that the number of annotated captions for each COCO image (average 5) is much less than the number for each VG image (average ∼50).

**Comparison with state-of-the-arts.** The experimental results compared with previous methods are presented in Table 3. All evaluation metrics are com-

| | SGG model | Grounding | R@20 | R@50 | R@100 |
|---|---|---|---|---|---|
| Unlocalized graph | VSPNet [58] | - | - | 4.70 | 5.40 |
| | LSWS [56] | - | - | 7.30 | 8.73 |
| | MOTIFS [60] | WSGM [41] | 4.12 | 5.59 | 6.45 |
| | MOTIFS [60] | SGNLS [65] | 7.23 | 9.28 | 10.71 |
| | MOTIFS [60] | Li et.al [21] | 9.09 | 11.39 | 12.89 |
| | Uniter† [5] | SGNLS [65] | 7.81 | 10.03 | 11.50 |
| | Uniter† [5] | Li et.al [21] | 9.57 | 11.80 | 13.15 |
| | VS$^3$ *(Swin-T)* | GLIP-L [17] | 18.02 | 23.89 | 28.19 |
| | VS$^3$ *(Swin-T+FreqBias)* | GLIP-L [17] | 20.06 | 26.72 | 31.75 |
| | VS$^3$ *(Swin-L+FreqBias)* | GLIP-L [17] | **22.18** | **29.81** | **34.96** |
| VG caption | LSWS [56] | - | - | 3.85 | 4.04 |
| | MOTIFS [60] | SGNLS [65] | 6.31 | 8.05 | 9.21 |
| | MOTIFS [60] | Li et.al [21] | 8.25 | 10.50 | 11.98 |
| | Uniter† [5] | SGNLS [65] | - | 9.20 | 10.30 |
| | Uniter† [5] | Li et.al [21] | 8.90 | 10.93 | 12.14 |
| | VS$^3$ *(Swin-T)* | GLIP-L [17] | 11.78 | 16.25 | 19.7 |
| | VS$^3$ *(Swin-L)* | GLIP-L [17] | **13.01** | **17.38** | **20.54** |
| COCO caption | LSWS [56] | - | - | 3.28 | 3.69 |
| | MOTIFS [60] | Li et.al [21] | 5.02 | 6.40 | 7.33 |
| | Uniter† [5] | SGNLS [65] | - | 5.80 | 6.70 |
| | Uniter† [5] | Li et.al [21] | 5.42 | 6.74 | 7.62 |
| | VS$^3$ *(Swin-T)* | GLIP-L [17] | 5.59 | 7.30 | 8.62 |
| | VS$^3$ *(Swin-L)* | GLIP-L [17] | **6.04** | **8.15** | **9.90** |

Table 3. Comparison with state-of-the-art language-supervised SGG methods, using weak scene graph supervision from three different text sources: unlocalized scene graphs, VG caption and COCO caption. All metrics are computed under the SGDET protocol on VG150 images. († indicates adapted for SGG.)

puted on the VG150 test set under the SGDET protocol. Specifically, under the unlocalized scene graphs setting, VS$^3$ with the Swin-T backbone (VS$^3$*(Swin-T)*) obtains substantial improvements on recall metrics over existing best results ($R$@20/50/100 from 9.57/11.80/13.15 to 18.02/23.89/28.19). Since relation frequency statistics are available in this setting, we use them as frequency biases [60] in predicate classification, leading to further performance gains (VS$^3$*(Swin-T+FreqBias)*). When using the stronger Swin-L backbone (VS$^3$*(Swin-L+FreqBias)*), we attain the highest performances ($R$@20/50/100 = 22.18/29.81/34.96), which even outperform many fully supervised methods (see Table 2). The VG caption setting provides weaker scene graph supervision via language parsing. We observe that our approach also outperforms previous state-of-the-art methods significantly. As for the most challenging COCO caption setting, it suffers from the additional domain shift problem since it trains on COCO but evaluates on VG150. As expected, the performances are lower than in the two aforementioned settings. But when comparing with previous works using the same text source, our approach still manages to achieve better performances. Overall, our approach consistently surpasses previous methods for language-supervised SGG. This demonstrates the benefits brought by pre-trained language-image models in terms of both grounding box acquirement and task transferring to tackle SGG.

**Ablation on scene graph parsing strategy.** We also conduct ablation studies on different scene graph parsing strategies for obtaining language SGG supervision. The results are shown in Table 4. Note that each image in COCO is annotated with 5 captions, and these captions are usually complimentary in describing image content. At first, we compare the performances between training with triplets from a single caption and all captions. We see the recalls achieve relative 10% performance boosts by replacing triplets from a single caption with all captions. This suggests the completeness of extracted scene graphs from image descriptions is a non-negligible factor for training a high-quality SGG model.

| *SG* from | *SG* parser | R@20/50/100 |
|---|---|---|
| Single caption | Simple | 5.07 / 6.25 / 7.36 |
| All captions | Simple | 5.42 / 6.82 / 7.93 |
| All captions | Advanced | **5.59 / 7.30 / 8.62** |

Table 4. Ablation on scene graph parsing strategies for language-supervised SGG. Results are obtained with VS$^3$*(Swin-T)* trained on scene graph supervision parsed from COCO captions.

Moreover, we compare two language parsers for extracting $\langle subject, predicate, object \rangle$ triplets: the simple SG parser [31], and the advanced SG parser [39]. Both parsers apply pre-defined rules to extract object and relation concepts from the semantic graphs of image language descriptions. Compared with the simple SG parser, the advanced SG parser covers additional features for dealing with complex quantificational modifiers (e.g., *a lot of*), resolving pronouns (e.g., *it*) and handling plural nouns (e.g., *three men*). The performance boosts of the advanced SG parser over the simple one (recalls from 5.42/6.82/7.93 to 5.59/7.30/8.62), indicates that the quality of semantic parsing is also important for language-supervised SGG.

### 4.4. Open-vocabulary SGG

**Setup.** Following [8], we train the proposed VS$^3$ with the same 70% object categories of VG150 as base categories. With the aid of the pre-trained VSS, we hope VS$^3$ can generalize to recognize the remaining 30% novel objects and their involved relations at inference. Concretely, we compute evaluation metrics over two object category sets: 70% base + 30% novel objects (dubbed as open-vocabulary SGG (Ov-SGG) evaluation), and 30% novel objects (dubbed as zero-shot SGG (ZsO-SGG) evaluation).

In addition, we adopt the PREDCLS and SGDET evaluation protocols [50]. PREDCLS assumes object information given, yet SGDET generates scene graphs from the raw image using predicted objects. Since VS$^3$ detects objects in a one-stage manner, we implement PREDCLS by selecting image regions that best match the ground-truth objects in post-processing, then performing relation recognition. We neglect the SGCLS protocol that assumes bounding box information given. This is because given bounding boxes can be directly used as region proposals in two-stage detectors, while the adopted one-stage manner in VS$^3$ has no region proposal counterpart.

| Method | Ov-SGG (70%+30%) | | ZsO-SGG (30%) | |
|---|---|---|---|---|
| | PREDCLS | SGDET | PREDCLS | SGDET |
| IMP [50] | 40.02 / 43.40 | - | 37.01 / 39.46 | - |
| MOTIFS [60] | 41.14 / 44.70 | - | 39.53 / 41.14 | - |
| VCTREE [48] | 42.56 / 45.84 | - | 41.27 / 42.52 | - |
| TDE [47] | 38.29 / 40.38 | - | 34.15 / 36.37 | - |
| GCA [13] | 43.48 / 46.26 | - | 42.56 / 43.18 | - |
| EBM [44] | 44.09 / 46.95 | - | 43.27 / 44.03 | - |
| SVRP [8] | 47.62 / 49.94 | - | 45.75 / 48.39 | - |
| $VS^3$ *(Swin-T)* | 50.10 / 52.05 | 15.07 / 18.73 | 46.91 / 49.13 | 10.08 / 13.65 |
| $VS^3$ *(Swin-L)* | **55.88 / 58.18** | **23.13 / 28.49** | **54.44 / 57.35** | **21.51 / 27.62** |

Table 5. Evaluation results (R@50/100) of fully supervised open-vocabulary SGG. Ov-SGG evaluates on 70% base categories + 30% novel categories in VG150, while ZsO-SGG only evaluates on 30% novel categories.

**Fully supervised results.** We first conduct experiments using manually annotated scene graphs. The results are presented in Table 5. For both Ov-SGG and ZsO-SGG, $VS^3$ achieves substantial performance improvements under PREDCLS. When upgrading to the stronger backbone Swin-L, more significant improvements are obtained. More importantly, we report performances for the challenging and more practical SGDET, which are neglected by all previous methods since their used object detector cannot handle open-vocabulary detection [8]. The SGDET performances ($R@50/100 = 10.08/13.65$) of ZsO-SGG using $VS^3$ are even higher than SGCLS metrics of SVRP ($R@50/100 = 9.30/11.32$ in [8]). This reveals the superiority of our approach to recognizing novel objects thanks to the open-vocabulary generalization ability of the pre-trained VSS.

**Language-supervised results.** Next, we evaluate the most challenging setting, i.e., open-vocabulary SGG using language supervision. To our knowledge, we are the first to propose such a new and practical SGG setting, and present the benchmark performances in Table 5. Not surprisingly, the recalls obtained via language-supervised training (i.e., $SG$ from VG caption or COCO caption) are lower than supervised results (i.e., $SG$ from annotated). When comparing $VS^3$ *(Swin-T)* and $VS^3$ *(Swin-L)* that is transferred from a stronger pre-trained model, the latter gets substantially higher Ov-SGG and ZsO-SGG performances. More importantly, we observe the performance gap between Ov-SGG and ZsO-SGG get closer in $VS^3$ *(Swin-L)*, e.g., the $R@50$ gap under the VG caption setting becomes *12.98-10.71=2.27* from *7.61-4.06=3.55*. This is due to the better generalization ability for recognizing novel classes. Moreover, the superior performances obtained by VG caption over COCO caption, indicate that using dense region-level descriptions and avoiding domain shift will help improve language-supervised open-vocabulary SGG in practice.

**Qualitative analysis.** We further showcase qualitative results of open-vocabulary SGG in Figure 3. The results demonstrate that our approach manages to detect novel objects and their relations with other objects. We also find that, compared with the fully supervised setting, the language-supervised results bias to predict simple relations such as

| Method | $SG$ supervision | Ov-SGG (70%+30%) | ZsO-SGG (30%) |
|---|---|---|---|
| $VS^3$ *(Swin-T)* | Manual annotation | **15.07 / 18.73** | **10.08 / 13.65** |
| | VG caption | 7.61 / 9.60 | 4.06 / 5.58 |
| | COCO caption | 4.39 / 5.63 | 3.65 / 4.73 |
| $VS^3$ *(Swin-L)* | Manual annotation | **23.13 / 28.49** | **21.51 / 27.62** |
| | VG caption | 12.98 / 16.29 | 10.71 / 13.70 |
| | COCO caption | 6.76 / 8.45 | 6.26 / 7.89 |

Table 6. Evaluation results (R@50/100) of open-vocabulary SGG using three different scene graph supervisions: manual annotation, VG caption and COCO caption (language-supervised). Ov-SGG evaluates on 70% base categories + 30% novel categories in VG150, while ZsO-SGG only evaluates on 30% novel categories.

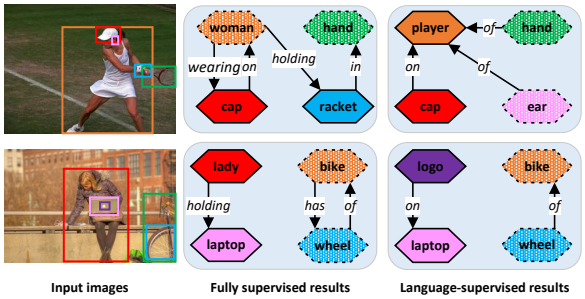


Figure 3. Qualitative results of open-vocabulary SGG, particularly from fully supervised and language-supervised (VG caption) settings. Note that dotted nodes denote novel objects. For clarity, we only show triplets among the top 20 predictions that depict relations of highlighted image regions (i.e., boxes on input images).

'on', 'of'. Presumably, it's because scene graph supervision parsed from language is more likely to extract such simple words as relation predicates.

## 5. Conclusion

In this work, we have proposed a novel approach to exploit a powerful pre-trained VSS for triggering language-supervised and open-vocabulary SGG. Particularly, we obtain cheap scene graph supervision by semantically parsing image language descriptions into semantic graphs and grounding the noun phrases through region-word alignment in the VSS. In addition, we devise the $VS^3$ model, which performs object detection as category name grounding in the VSS and naturally builds relation representations for relation recognition. Thanks to the open-vocabulary generalization ability of the VSS, $VS^3$ manages to detect novel objects and their relations with other objects, achieving open-vocabulary SGG. We validate our approach on the Visual Genome benchmark across supervised, language-supervised and open-vocabulary SGG settings, and have set new state-of-the-art performances. This demonstrates the merits of transferring pre-training knowledge to push SGG toward more practical scenarios.

**Acknowledgment.** This work was supported in part by the Shenzhen Science and Technology Program under Grant JCYJ20220818103006012 and ZDSYS20211021111415025.

## References


[1] Saeid Amiri, Kishan Chandan, and Shiqi Zhang. Reasoning with scene graphs for robot planning under partial observability. *IEEE Robotics and Automation Letters*, 7(2):5560–5567, 2022. 1, 2

[2] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *CVPR*, 2020. 1

[3] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *CVPR*, 2019. 6

[4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 7

[6] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019. 1, 2

[7] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2021. 2, 3

[8] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *ECCV*, 2022. 3, 5, 7, 8

[9] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICLR*, 2021. 3

[10] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 1, 2

[11] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 3

[12] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. 3

[13] Boris Knyazev, Harm de Vries, Cătălina Cangea, Graham W Taylor, Aaron Courville, and Eugene Belilovsky. Generative compositional augmentations for scene graph prediction. In *ICCV*, 2021. 8

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 1, 2, 5

[15] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2021. 2, 3

[16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021. 3

[17] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, et al. Grounded language-image pre-training. In *CVPR*, 2022. 2, 3, 4, 6, 7

[18] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. *arXiv preprint arXiv:2005.02472*, 2020. 1

[19] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *CVPR*, 2022. 6

[20] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*, 2021. 1, 2

[21] Xingchen Li, Long Chen, Wenbo Ma, Yi Yang, and Jun Xiao. Integrating object-aware and interaction-aware knowledge for weakly supervised scene graph generation. In *ACM MM*, 2022. 2, 3, 5, 6, 7

[22] Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, and Tao Mei. X-modaler: A versatile and high-performance codebase for cross-modal analytics. In *ACM MM*, 2021. 1

[23] Yehao Li, Yingwei Pan, Ting Yao, Jingwen Chen, and Tao Mei. Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network. In *AAAI*, 2021. 3

[24] Yongzhi Li, Duo Zhang, and Yadong Mu. Visual-semantic matching by exploring high-order attention and distraction. In *CVPR*, 2020. 1

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 4

[26] Xin Lin, Changxing Ding, Yibing Zhan, Zijian Li, and Dacheng Tao. Hl-net: Heterophily learning network for scene graph generation. In *CVPR*, 2022. 1, 2, 6

[27] Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *CVPR*, 2021. 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 6

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

[30] Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Hongyang Chao, and Tao Mei. Coco-bert: Improving video-language pre-training with contrastive cross-modal matching and denoising. In *ACM MM*, 2021. 3

[31] Jiayuan Mao. Scenegraphparser, 2019. https://github.com/vacancy/SceneGraphParser (Access date: 2022-8-11). 3, 7

[32] Victor Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? *arXiv preprint arXiv:2009.12313*, 2020. 1

[33] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 3, 5, 6

[34] Yingwei Pan, Yehao Li, Jianjie Luo, Jun Xu, Ting Yao, and Tao Mei. Auto-captions on gif: A large-scale video-sentence dataset for vision-language pre-training. In *ACM Multimedia*, 2022. 3

[35] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020. 1

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICLR*, 2021. 2, 3

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 2016. 2

[38] Brigit Schroeder and Subarna Tripathi. Structured query-based image retrieval using scene graphs. In *CVPR Workshops*, 2020. 1

[39] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, 2015. 3, 5, 7

[40] Sahand Sharifzadeh, Sina Moayed Baharlou, and Volker Tresp. Classification by attention: Scene graph classification with prior knowledge. *arXiv preprint arXiv:2011.10084*, 2020. 1, 2

[41] Jing Shi, Yiwu Zhong, Ning Xu, Yin Li, and Chenliang Xu. A simple baseline for weakly-supervised scene graph generation. In *ICCV*, 2021. 2, 3, 5, 6, 7

[42] Motoharu Sonogashira, Masaaki Iiyama, and Yasutomo Kawanishi. Towards open-set scene graph generation with unknown objects. *IEEE Access*, 10:11574–11583, 2022. 2

[43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 3

[44] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *CVPR*, 2021. 8

[45] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. Multi-modal knowledge graphs for recommender systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020. 1

[46] Kaihua Tang. A scene graph generation codebase in pytorch, 2020. https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch. 6

[47] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 1, 2, 5, 8

[48] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, 2019. 1, 2, 6, 8

[49] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, 2020. 1

[50] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 1, 2, 5, 6, 7, 8

[51] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, 2018. 1, 2

[52] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019. 1

[53] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022. 2, 3

[54] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018. 1

[55] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *ICCV*, 2019. 1

[56] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. In *CVPR*, 2021. 2, 5, 6, 7

[57] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *ECCV*, 2020. 1, 2

[58] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *CVPR*, 2020. 2, 7

[59] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 2, 3

[60] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018. 1, 2, 6, 7, 8

[61] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 6

[62] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, 2019. 2, 6

[63] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *CVPR*, 2022. 1

[64] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Boosting scene graph generation with visual relation saliency. *ACM TOMM*, 2023. 1

[65] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision. In *ICCV*, 2021. 2, 3, 5, 6, 7

[66] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *ECCV*, 2020. 1

[67] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 2, 3