

# Topic Scene Graph Generation by Attention Distillation from Caption

Wenbin Wang<sup>1,2</sup>, Ruiping Wang<sup>1,2,3</sup>, Xilin Chen<sup>1,2</sup>

<sup>1</sup>Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),  
Institute of Computing Technology, CAS, Beijing, 100190, China

<sup>2</sup>University of Chinese Academy of Sciences, Beijing, 100049, China

<sup>3</sup>Beijing Academy of Artificial Intelligence, Beijing, 100084, China

wenbin.wang@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

## Abstract

If an image tells a story, the image caption is the briefest narrator. Generally, a scene graph prefers to be an omniscient “generalist”, while the image caption is more willing to be a “specialist”, which outlines the gist. Lots of previous studies have found that a scene graph is not as practical as expected unless it can reduce the trivial contents and noises. In this respect, the image caption is a good tutor. To this end, we let the scene graph borrow the ability from the image caption so that it can be a specialist on the basis of remaining all-around, resulting in the so-called **Topic Scene Graph**. What an image caption pays attention to is distilled and passed to the scene graph for estimating the importance of partial objects, relationships, and events. Specifically, during the caption generation, the attention about individual objects in each time step is collected, pooled, and assembled to obtain the attention about relationships, which serves as weak supervision for regularizing the estimated importance scores of relationships. In addition, as this attention distillation process provides an opportunity for combining the generation of image caption and scene graph together, we further transform the scene graph into linguistic form with rich and free-form expressions by sharing a single generation model with image caption. Experiments show that attention distillation brings significant improvements in mining important relationships without strong supervision, and the topic scene graph shows great potential in subsequent applications.

## 1. Introduction

A picture is worth a thousand words. However, only a few person prefers to know all of the “thousand words”, while others would like to be informed the “topic words”. Therefore, the scene graph and the image caption are used for conveying the image contents out of different purposes.

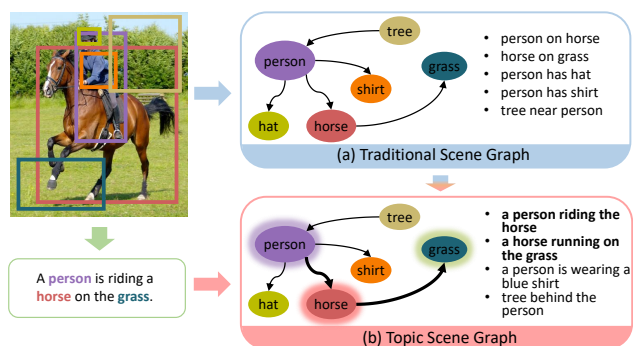


Figure 1. Comparison of the (a) traditional general scene graph, and the (b) topic scene graph generated under the guidance of attention in the image caption, which gives priority to the important relationships (highlighted nodes and edges), and expresses a relationship in the style of natural language.

Concretely, the scene graph [20] consists of objects in an image and the relationships between pairs of objects. A series of studies have tried to generate the scene graph and realize its potential in advanced intelligence tasks, *e.g.*, visual Q&A [2, 47], visual reasoning [44], and vision-and-language navigation (VLN) [53], *etc.* Nevertheless, as pointed in [25, 35, 52], the scene graph is helpful only if it is informative, while the current generated scene graph with such a lot of noises does not meet this standard. This is mainly due to the explosive combination possibility of two objects [52, 65], which brings the double-edge effect that the scene graph is comprehensive but the key information is overwhelmed by massive trivial details. It is necessary and practical to make the scene graph well-circumscribed between important and trivial contents. Fortunately, the image caption exactly shows this ability and is a good teacher from which a scene graph should learn.

In the context of scene graph generation, few researches devote endeavor to discovering the important relationships, which is a meaningful step for restricting the scale of the

scene graph when it is used for downstream tasks. The most popular approach is to keep the relationships with large products of the predicted subject, object, and predicate scores. However, this product measures the accuracy of prediction rather than the importance. Yang *et al.* [58] and Lv *et al.* [33] either use a light-weight relationship proposal network to extract some probably related pairs, or predict an attention score for each relationship, based on the perspective that annotated relationships are the important ones. This may be questionable because the mainstream scene graph datasets (*e.g.*, Visual Genome [22]) are suffered from serious long-tailed problem [6, 46] and the annotated pairs (head pairs) are usually trivial ones [52]. To more precisely define what is important relationship, image caption is found helpful because a caption almost exactly reveals what humans think important [14], as shown in Figure 1. Consequently, Yu *et al.* [65] and Wang *et al.* [52] learn to mine the important relationships under the strong supervision from the important relationship annotation which is obtained under the guidance of captions. But this process need to transform the captions into triplets first and then align two groups of heterologous triplets, which is so expensive and complicates the scene graph generation.

In this work, we propose to let the scene graph learn the important relationships from the image caption in an economical way, resulting in the **Topic Scene Graph**. The importance of relationship is estimated by distilling the visual attention during the image caption generation, which is treated as the *weak* supervision. Specifically, most advanced image captioners are able to fix its gaze on the correct object regions. We apply an image captioner and collect the *first-order attention* information with respect to the object regions, which are used for assembling the *second-order attention* about the relationships. In this way, we actually transform humans' attention into a new form, converting it from the concern about individual to that about relational events. The second-order attention is used as the weak supervision for guiding the estimation of the importance of the relationship. In this way, strong supervision is no longer necessary.

Furthermore, as the attention distillation process makes it possible to generate the image caption and scene graph simultaneously and both of them are the description of an image, why not generating them with a single model? It is noted that the most popular scene graph dataset, Visual Genome (VG) [22] of world scale contains more than 40,000 types of relationships which are originally extracted from humans' language, while the traditional definition of scene graph treats the relationship recognition as predicate classification and makes most of the relationships filtered, which is harmful to the diversity of relationship description. What is worse, there exist huge interior differences in some certain predicate classes, *e.g.*, the appearance of

two relationship triplet instances for the predicate class *riding*, *person-riding-horse* and *dog-riding-skateboard* are totally different. It is difficult to clearly define the semantic boundaries between different predicates. Inspired by [21], we redefine the scene graph as the set of short relational sentences. In this way, a shared captioning module can be used for the so-called **linguistic scene graph** generation and image captioning at the same time.

## 2. Related Work

**Scene Graph Generation (SGG) and Visual Relationship Detection (VRD)** focus on understanding the relationships between objects. Early studies [10, 42] treat each distinct combination of object categories and relationship predicates as a distinct class. Lu *et al.* [32] formally define the VRD task and address the object and predicate classification separately. Recent state-of-the-art VRD works [8, 18, 26, 34, 38, 64, 66, 71, 72, 73] pay attention on the prediction of each relationship triplet independently. The scene graph which describes the image faithfully from a bird's-eye view is proposed in [20]. After that, a batteries of studies contribute to generation of a high-quality scene graph. Message passing mechanism [55] has been proved effective and its variants are widely adopted in [27, 28]. The latest essential practice achieves more promising results through constructing reasonable context among objects and visual relationships [31, 39, 47, 51, 58, 70], or taking the advantage of external knowledge and common-sense [6, 13, 67, 69]. Besides, Zhang *et al.* [74] propose contrastive losses to resolve the related pair configuration ambiguity. Zareian *et al.* [68] creatively consider SGG as an edge role assignment problem. Tang *et al.* [46] diversify the predicted relationships through addressing the causal effect. Most of these works struggle to fit the VG dataset but always overlook the fact that the scene graph annotation suffers from serious long-tailed problem and the valuable relationships are usually overwhelmed by trivial ones. This problem is general because of the reporting bias [11] and should not be blamed on a particular dataset. A growing number of works are considering how to make the scene graph more practical. Liang *et al.* [29] prune the dominant and easy-to-predict relations while keeping the visually relevant relations in VG. Lv *et al.* [33] estimate the importance of relationships with an attention module, but actually they still think that the annotated relationships are semantically important, which may not be true. Yu *et al.* [65] and Wang *et al.* [52] provide annotations with relationships of humans interest under guidance of image caption and explicitly use them as supervision. However, only semantically important relationships are detected in [65], which is not enough for a comprehensive scene graph. In this work, we distill the attention from image caption as weak supervision rather than constructing the high-cost annotation of important re-

relationships, and reasonably estimate the importance of all relationships so that the number of remaining relationships are controllable.

**Image captioning and scene graph.** Compared to the scene graph, the image caption is usually treated as the final presentation to humans (for interaction). Restricted by the length, an image caption usually contains the most important contents in an image [14], but misses details. Some researchers propose the dense captioning task [19] which generates diverse but aimless region-level descriptions passively. Our proposed topic scene graph is naturally a structured representation of an image, and is especially able to actively estimate the importance of the image contents.

Early studies on image captioning are rule or template-based [45, 61]. Modern captioning models have achieved great progress benefiting from encoder-decoder framework [49], attention technique [4, 7, 15, 17, 24, 37, 48, 56, 60] and RL-based training objective [41]. Our work distills the attention from image caption. Besides, as the scene graph contains much semantic information, lots of works have tried to incorporate it into captioning models [5, 12, 25, 35, 57, 59, 62, 63]. Inspired by this direction, we propose to generate linguistic scene graph, and creatively let the scene graph benefit from image caption.

### 3. Approach

#### 3.1. Overview

Given an image  $I$ , its scene graph  $G$  consists of a set of objects (nodes)  $O = \{o_i \mid i \in [1, n]\}$  with the assigned class labels  $C = \{c_i \mid i \in [1, n], c_i \in \mathcal{C}\}$  and the corresponding bounding boxes  $B = \{b_i \mid i \in [1, n], b_i \in \mathbb{R}^4\}$ , and a set of relationships (edges)  $R = \{r_k \mid k \in [1, m]\}$ . Conventionally, each relationship  $r_k$  is a triplet of the start node  $o_i$ , the end node  $o_j$ , and the relationship label  $x_{ij} \in \mathcal{R}$  where  $\mathcal{R}$  is the set of predicate types. These relationships are disordered. Thinking of the limitation of this representation, we redefine the relationship  $r_k$  as a relational caption in the form of word sequence  $(y_{kt})_{t=1}^{T_R}$ , where  $y_{kt} \in \mathcal{V}$  and  $\mathcal{V}$  denotes the vocabulary.  $t$  is the positional index of the word in the sequence and  $T_R$  is the sequence length. More importantly, these relationships are sorted according to their importance. Specifically, as depicted in Figure 1 (b), the detected objects are used for generating the relational captions and the image caption  $(w_t)_{t=1}^{T_C}$  ( $w_t \in \mathcal{V}$  and  $T_C$  is the caption length), during which the subjective interest (attention)  $\alpha$  is collected from the image caption. The  $\alpha$  is used for the estimation of importance scores of the relationships.

In the following parts, we will describe the generation of the image caption (Sec. 3.2) and the linguistic scene graph (Sec. 3.3) using a shared captioning module. Then we will elaborate on importance score estimation and attention distillation for obtaining the final topic scene graph (Sec. 3.4).

#### 3.2. Captioning Module

In this work, we adopt two types of state-of-the-art captioning models, the Up-Down model [1] based on LSTM [16], and the Transformer [48]. The Up-Down model comprises of an attention LSTM layer and a language LSTM layer. Specifically, the detected objects are represented by their visual features  $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n] \in \mathbb{R}^{d_v \times n}$  and bounding boxes  $[b_1, \dots, b_n] \in \mathbb{R}^{4 \times n}$ . The object visual features are firstly transformed to  $\mathbf{V}' = [\mathbf{v}'_1, \dots, \mathbf{v}'_n] \in \mathbb{R}^{d_i \times n}$  with a lower dimension:

$$\mathbf{v}'_i = \mathbf{W}_v \mathbf{v}_i + \mathbf{b}_v, \quad \mathbf{v}'_i \in \mathbb{R}^{d_i}, \quad (1)$$

where  $\mathbf{W}_v \in \mathbb{R}^{d_i \times d_v}$  and  $\mathbf{b}_v \in \mathbb{R}^{d_i}$  are trainable parameters. At each time step  $t$ , the previous hidden state of the language LSTM  $\mathbf{h}_{t-1}^2$  is concatenated with the mean-pooled image feature  $\bar{\mathbf{v}} = \frac{1}{n} \sum_i \mathbf{v}'_i$  and the previous word embedding  $\mathbf{e}_{t-1} = \mathbf{W}_e w_{t-1}$ , and fed into the attention LSTM:

$$\mathbf{h}_t^1 = \text{LSTM}_{\text{Att}}([\mathbf{h}_{t-1}^2; \bar{\mathbf{v}}; \mathbf{e}_{t-1}], \mathbf{h}_{t-1}^1), \quad \mathbf{h}_t^1 \in \mathbb{R}^{d_h}, \quad (2)$$

where  $[\cdot]$  denotes the concatenation and  $\mathbf{W}_e \in \mathbb{R}^{d_e \times |\mathcal{V}|}$  is the embedding matrix. The  $w_{t-1}$  stands for the  $|\mathcal{V}|$ -dim one-hot vector where the  $w_{t-1}$ -th element is 1 in practice. The attention about the objects  $\alpha = [\alpha_1, \dots, \alpha_{T_C}] \in \mathbb{R}^{n \times T_C}$  is calculated as:

$$\begin{aligned} z_{i,t} &= \mathbf{w}_a^T \tanh(\mathbf{W}_{va} \mathbf{v}'_i + \mathbf{W}_{ha} \mathbf{h}_t^1), \\ \alpha_t &= \text{softmax}(z_t), \end{aligned} \quad (3)$$

where  $\mathbf{w}_a \in \mathbb{R}^{d_a \times 1}$ ,  $\mathbf{W}_{va} \in \mathbb{R}^{d_a \times d_i}$ , and  $\mathbf{W}_{ha} \in \mathbb{R}^{d_a \times d_h}$  are trainable parameters. Finally, the attended image visual feature  $\hat{\mathbf{v}}_t = \sum_{i=1}^n z_{i,t} \mathbf{v}'_i$  and  $\mathbf{h}_t^1$  are used as the input of the language LSTM, which predicts the conditional distribution over the possible word:

$$\begin{aligned} \mathbf{h}_t^2 &= \text{LSTM}_{\text{Lang}}([\hat{\mathbf{v}}_t; \mathbf{h}_t^1], \mathbf{h}_{t-1}^2), \quad \mathbf{h}_t^2 \in \mathbb{R}^{d_h}, \\ p(w_t | w_{1:t-1}) &= \text{softmax}(\mathbf{W}_o \mathbf{h}_t^2 + \mathbf{b}_o), \end{aligned} \quad (4)$$

with trainable parameters  $\mathbf{W}_o \in \mathbb{R}^{|\mathcal{V}| \times d_h}$  and  $\mathbf{b}_o \in \mathbb{R}^{|\mathcal{V}|}$ .

As for the Transformer model, it consists of an encoder and a decoder, both of which contain a stack of layers. We provide details in the Supp. and especially explain how to extract the attention on objects here. For captioning task, the transformed visual features  $\mathbf{V}'$  are fed into the encoder and we get the output  $\mathbf{V}^* \in \mathbb{R}^{d_{tr} \times n}$ . For each decoder layer in the decoder, it contains a multi-head self-attention layer and a multi-head cross-attention layer. All the word embeddings  $\mathbf{E}$  are fed into the self-attention layer to get the output  $\mathbf{E}^* \in \mathbb{R}^{d_{tr} \times T_C}$ . In each head  $j \in [1, H]$ , the attention weights  $\alpha^j \in \mathbb{R}^{n \times T_C}$  about objects are computed by:

$$\alpha^j = \text{softmax} \left( \frac{\mathbf{V}^{*T} \mathbf{E}^*}{\sqrt{d_{tr}}} \right). \quad (5)$$

We average the  $\alpha^j$  across  $H$  heads and obtain the final  $\alpha$ .

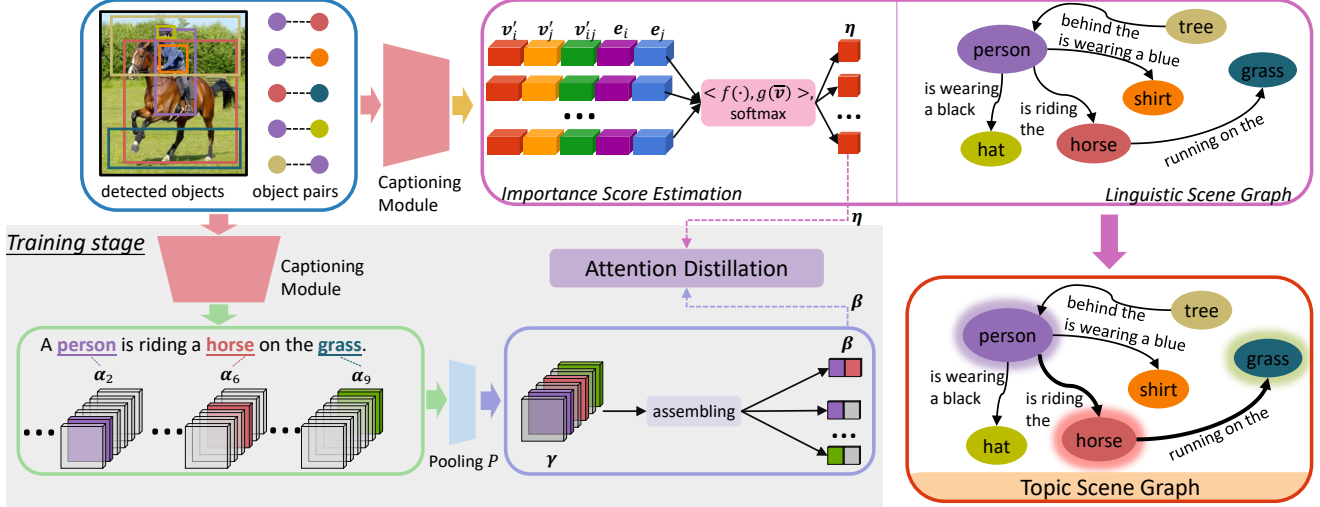


Figure 2. The framework of our method. The detected objects and the object pairs are fed into a shared captioning module to generate the image caption and relational captions which form a linguistic scene graph. During the training stage, the attention information is collected, pooled over multiple time steps, and assembled to produce the second-order attention  $\beta$  about the relationships. Simultaneously the importance scores  $\eta$  of the relationships are also estimated, and then regularized by  $\beta$ . In the testing phase, the estimated importance scores are used for generating the final topic scene graph.

### 3.3. Linguistic Scene Graph

In this part, we share the captioning module to make it applicable to relational captioning so that a linguistic scene graph is realizable. Following the general scene graph generation process, we build combinations of  $N$  detected objects and obtain  $O(N^2)$  object pairs. For the subject  $o_i$  and the object  $o_j$ , we extract their union visual feature  $v_{ij} \in \mathbb{R}^{d_u}$  which contains rich context information by applying ROI pooling [40] with the union box of  $b_i$  and  $b_j$ . Besides, as the relative position between two objects is found as effective prior information, we follow [38] to build the geometry feature:

$$g_{ij} = \left[ \frac{x_j - x_i}{\sqrt{w_i h_i}}, \frac{y_j - y_i}{\sqrt{w_i h_i}}, \sqrt{\frac{w_j h_j}{w_i h_i}}, \frac{w_i}{h_i}, \frac{w_j}{h_j}, \frac{b_i \cap b_j}{b_i \cup b_j} \right], \quad (6)$$

where the  $(x, y)$  is the center position and  $w$  and  $h$  denote width and height of a box. It is further projected to a 64-dim feature and concatenated with  $v_{ij}$  to obtain the final union feature  $v'_{ij} \in \mathbb{R}^{d_i}$ :

$$v'_{ij} = \mathbf{W}_u [v_{ij}; (\mathbf{W}_g g_{ij} + \mathbf{b}_g)] + \mathbf{b}_u, \quad (7)$$

where  $\mathbf{W}_u \in \mathbb{R}^{d_i \times (d_u + 64)}$ ,  $\mathbf{b}_u \in \mathbb{R}^{d_i}$ ,  $\mathbf{W}_g \in \mathbb{R}^{64 \times 6}$ , and  $\mathbf{b}_g \in \mathbb{R}^{64}$  are trainable parameters.

Different from the image captioning which should pay attention to all objects, relational captioning focuses on two designated objects. Specifically, for the Up-Down model, only the  $v'_i$ ,  $v'_j$  and  $v'_{ij}$  are used for decoding. For the Transformer, all object features are fed into the encoder to construct contextual information  $\mathbf{V}^*$ , but only the  $v'_i$ ,  $v'_j$  and  $v'_{ij}$  are fed into the decoder.

### 3.4. Topic Scene Graph

With the attention about objects provided by image captioning, we propose to assemble it to obtain attention about the relationships, which is used as the weak supervision to guide the estimation of the importance of the relationships.

Suppose that there are  $m$  relationships. We first estimate the importance score for each relationship consisting of subject  $o_i$  and object  $o_j$ . Specifically, as depicted in the top middle part in Figure 2, we concatenate the  $v'_i$ ,  $v'_j$ ,  $v'_{ij}$  and the semantic embeddings of the subject and object categories,  $e_i, e_j \in \mathbb{R}^{d_{sem}}$ , to form a query  $q_{ij}$ , and compute the key using the global feature  $\bar{v}$ :

$$q_{ij} = f([v'_i; v'_j; v'_{ij}; e_i; e_j]) \in \mathbb{R}^{d_s}, \quad (8)$$

$$k = g(\bar{v}) \in \mathbb{R}^{d_s}, \quad (9)$$

where  $f$  and  $g$  are two learnable linear transformation functions. The estimated importance score  $s_{ij}$  is calculated as the inner product of the query and key and then normalized with softmax function:

$$s_{ij} = \frac{q_{ij}^T k}{\sqrt{d_s}}, \quad (10)$$

$$\eta = \text{softmax}(s) \in \mathbb{R}^m. \quad (11)$$

On the other hand, we have the attention information  $\alpha$  with respect to individual objects, which is used to assemble the attention  $\beta$  with respect to relationships. As shown in the bottom left part (with gray background) in Figure 2, firstly, we gather the attention score for each object over multiple time steps with a pooling function  $P$ , resulting in

$\gamma \in \mathbb{R}^n$ . Then the second order attention for a relationship is assembled as:

$$\delta_{ij} = \gamma_i + \gamma_j, \quad \beta = \text{softmax}(\delta) \in \mathbb{R}^m. \quad (12)$$

Finally, the estimated score  $\eta$  is regularized with the induced second-order attention  $\beta$  via KL-divergence.

### 3.5. Optimization

The optimization process is divided into two stages. In the first stage, given a single ground truth image caption  $(w_t)_{t=1}^{T_C}$  and  $m$  ground truth relational captions  $[(y_{kt})_{t=1}^{T_R}]_{k=1}^m$ , the captioning module is optimized with the traditional cross-entropy loss consisting of the image captioning part and the relational captioning part:

$$L_{CE} = \sum_{t=1}^{T_C} -\log p(w_t|w_{1:t-1}) + \lambda \left( \frac{1}{m} \sum_{k=1}^m \sum_{t=1}^{T_R} -\log p(y_{kt}|y_{k1:k(t-1)}) \right), \quad (13)$$

where  $\lambda$  is the balance parameter. In the second stage, the attention distillation module is optimized with the KL-divergence loss:

$$L_{KL} = KL(\eta||\beta). \quad (14)$$

Although the reinforce algorithm such as SCST [41] is widely used for further optimizing the captioning models, some researches [75] found that SCST actually does harm to the text-to-image grounding because it encourages the n-gram consistency rather than visual semantic alignment. As our framework has a high demand for superior grounding performance, we do not optimize the captioning module with SCST in this research and leave it to future works.

## 4. Experiments

### 4.1. Datasets and Evaluation Metrics

**Datasets.** There is no existing dataset with both the image caption and relational caption. Inspired by [21], we refer to their data construction procedure to collect the relational caption, and further collect the image caption, as well as the important relationship annotation which is only for training the upper bound models and evaluation. Specifically, we use the 51,208 images in both VG and MSCOCO [30] datasets which have both relationship and image caption annotation. Firstly, we cleanse the VG dataset and keep a large scale vocabulary including 3,000 object categories and 800 attributes. The relationships about the objects beyond these categories are filtered. To obtain the important relationship annotation, we apply the Scene Graph Parser [43] to extract the relationships from the image caption, and

align them with the annotated ones by matching their subject and object WordNet [36] synsets. Finally we convert the remaining relationships into sentences. It is worth mentioning that we do not make a category-wise vocabulary for relationships, but keep the relationships in their original free and open form. To further enrich the concepts, the attributes of each object are randomly selected to add to the sentences. We obtain 35,928 images for 29,928/1,000/5,000 splits for train/validation/test sets respectively and 11,437 vocabularies (including 3,000 object categories and 800 attributes).

**Evaluation metrics.** We use BLEU, METEOR, CIDEr-D, ROUGE-L, and SPICE for image captioning. For relational captioning, we refer to [19] and [21] and use the following metrics. (1) mean Average Precision (mAP): it uses METEOR score [9] with thresholds  $\{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$  for language and IoU thresholds  $\{0.2, 0.3, 0.4, 0.5, 0.6\}$  for localization. Only the pair whose subject and object have IoUs greater than thresholds is a true positive sample. The mAP is obtained by averaging across all the combinations of language and localization thresholds. (2) image-level recall (Img-Lv.Recall): it ignores the localization and evaluate the recall of the bag of predicted relational captions.

Besides, in order to evaluate whether the important relationships are properly found, we refer to the metrics in traditional scene graph generation [52, 55], *i.e.*, Recall@K where K is set to 20, 50, and 100. Under this metric, only the important relationships are regarded as ground truth and the top K relationships are evaluated, which means that the predicted relationships should be sorted. A relational caption is correct only if the following two conditions are satisfied: (1) both the subject and object have IoUs greater than 0.5, and (2) the METEOR score is greater than the thresholds above. We average the recall on different language thresholds. To evaluate the performance on discovering correct important object pairs, we derive the Recallns@K metric which only requires the above first condition and does not consider the METEOR scores.

### 4.2. Implementation Details

We firstly train the faster-RCNN [40] detector with the ResNeXt-101 [54] backbone on the objects of 3,000 categories of our dataset. During the scene graph training, the parameters of the object detector are frozen. More details are given in the Supp. Source codes are available at <https://vipl.ict.ac.cn/resources/codes> or <http://www.kennethwong.tech/>.

### 4.3. Experiments on Linguistic Scene Graph

In our method, a shared captioning module is trained for image captioning (IC) and relational captioning (RC), which has never been explored before. We start with exploring the effectiveness of this practice. To this end, we

Table 1. Image captioning results. B1, B4, M, R, C, S denote BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr-D, and SPICE respectively. “-ICRC” denotes that the model is trained with image captions and relational captions.

Model	$\lambda$	B1	B4	M	R	C	S
UD [1]	-	69.8	29.6	25.0	52.3	94.1	18.0
	0.1	<b>71.1</b>	<b>30.4</b>	<b>25.1</b>	<b>52.6</b>	<b>95.3</b>	<b>18.3</b>
UD-ICRC	0.3	70.7	30.0	24.9	52.5	94.6	18.1
	0.7	70.5	30.1	25.0	52.4	94.8	18.2
	1.0	71.0	30.0	24.8	52.5	93.5	17.9
Transformer [48]	-	68.8	26.8	23.5	50.4	85.6	17.3
Transformer-ICRC	0.7	<b>70.3</b>	<b>28.6</b>	<b>24.4</b>	<b>51.7</b>	<b>91.5</b>	<b>18.0</b>

adjust the balance parameter  $\lambda$  to control the proportion of the loss of the relational captioning in the final loss function. Results under more value settings are provided in Supp. The evaluation is divided into two parts. On the side of image captioning, the baselines are Up-Down (UD) and Transformer, which are trained only using the image captions. We compare the baselines to the UD-ICRC and Transformer-ICRC trained with the image captions and relational captions. From Table 1, we observe that mixed training actually brings benefit to the image captioning, but as the  $\lambda$  increases, this benefit will slightly drop. It suggests that mixed training is feasible despite the fact that the assembled relational captions will bring some noises.

On the side of relational captioning, we use TriLSTM [21], UD-RC, and Transformer-RC as baselines, which are trained only using the relational captions. The TriLSTM is re-implemented and trained on our dataset. The relational captions are sorted by the product of the probabilities of the generated words, *i.e.*, likelihoods. As shown in Table 2, compared to the TriLSTM, both the UD-RC and UD-ICRC outperform it obviously. Comparing the UD-RC with UD-ICRC, we find that as the  $\lambda$  increases, the UD-ICRC roughly performs better on the image level metrics, and surpasses the UD-RC baseline when  $\lambda$  is greater than 0.7. However, the performance drops on the important relationship recall metrics. We think that it is because the increasing  $\lambda$  makes the model fit the relational captions data better, but the increasing sentence likelihood loses its discrimination and is less suitable for importance estimation. It also suggests that neither the sentence likelihood of the relational caption nor the score product of the traditional triplet are unstable for importance estimation. As for the Transformer, mixed training has little impact on the performances. With a comprehensive consideration on the performances of the two tasks, we set the  $\lambda$  as 0.7 and freeze the UD-ICRC / Transformer-ICRC models for generating the topic scene graph in the following experiments.

#### 4.4. Experiments on Topic Scene Graph

As we know this is the first time to study the topic linguistic scene graph generation. We replace some key components to show the effectiveness of our proposed

model (UD-ICRC-attn) and facilitate the ablation study. As we have the important relationship annotation, we train the upper bound models named as UD-ICRC-label and Transformer-ICRC-label under the supervision of the annotated important relationships with binary cross entropy loss. The results are shown in Table 3.

**Pooling function.** The pooling function  $P$  is used for gathering attention information over multiple time steps for each individual object. We compare two functions: max pooling (MAX) and mean pooling (MEAN). Comparing the 4th row and the 5th row in the UD-ICRC-attn section, the max pooling function is much more effective than mean pooling. It is reasonable because we want to maximize the scores of the objects which are mentioned in the image captions, while the mean pooling reduces the attention scores and makes it hard to shed light on the key objects.

**Input features.** We try to use different concatenated features for obtaining the query  $q$  when estimating the importance scores, including the union features (U), subject and object features (SO), subject, object, and union features (SOU), and subject, object, union features together with the semantic embeddings of subject and object categories (SOUS). By comparing the 1st~3rd rows and the 6th row in the UD-ICRC-attn section, and the rows in the UD-ICRC-label section, it is found that the SOU significantly improves the performances compared with U and SO, suggesting that these three types of features cannot be used independently, as the SO provides information about objects and the U provides relative spatial information. The semantic embeddings bring slight improvement, and it is not as obvious as that in the upper bound models.

**Masking non-noun words.** When gathering the attention information, we explore whether all the words of a sentence should be considered or not. Different from considering all words, we try an alternative way that only the attention of noun words are collected since they are probably to be correctly grounded to the regions, and other words are masked. To this end, we apply the NLTK POS tagger [3] to filter out the non-noun words. Comparing the 5th row and 6th row in the UD-ICRC-attn section, it is interesting to find that masking the non-noun words does harm to the performance instead. This phenomenon may imply that the context plays a crucial role and the non-noun words would also contribute to the attention of the center nouns.

Overall, compared with the TriLSTM, UD-ICRC and Transformer-ICRC baselines, the application of our attention alignment module significantly improves the performances, and obviously reduces the gap between the baselines and the upper bound. The best configuration is to use the SOUS input features, max pooling function and collect the attention from all words. It’s noted that our method does not need the complicated collection of important relationship annotation, but can still provide the useful important

Table 2. Results of relational captioning (%). “-RC” denotes that the model is only trained with the relational captions. “-ICRC” denotes that the model is trained with image captions and relational captions. R-ns means Recall-ns. Img-Lv. Recall means the image level recall.

Model	$\lambda$	mAP	METEOR	Img-Lv. Recall	R@20	R-ns@20	R@50	R-ns@50	R@100	R-ns@100
TriLSTM [21]	-	3.80	30.21	72.72	1.31	3.20	3.93	9.58	8.42	20.88
UD-RC [1]	-	<b>5.61</b>	42.40	88.77	3.02	3.71	<b>10.46</b>	12.92	<b>22.97</b>	28.90
UD-ICRC	0.1	4.84	38.31	84.81	<b>3.45</b>	<b>4.43</b>	10.22	<b>13.99</b>	20.77	<b>29.00</b>
	0.3	5.14	40.36	86.93	3.39	4.18	9.87	12.88	21.57	27.99
	0.7	5.43	42.26	89.15	2.75	3.49	9.97	12.40	20.76	26.46
	1.0	5.41	<b>42.75</b>	<b>89.52</b>	2.31	2.90	8.09	10.20	19.97	25.56
Transformer-RC [48]	-	5.26	41.62	88.65	2.11	2.73	6.83	9.12	16.36	21.91
Transformer-ICRC	0.7	5.15	41.63	88.64	2.05	2.70	6.86	9.19	16.21	21.91

Table 3. Results (%) comparison on discovering the important relationships. “Feat.” denotes different input features. “P” denotes the pooling function. “Mask” denotes masking the non-noun words ( $\checkmark$ ) or not ( $\times$ ).

	Feat.	P	Mask	R@20	R-ns@20	R@50	R-ns@50	R@100	R-ns@100	mean
TriLSTM	-	-	-	1.31	3.20	3.93	9.58	8.42	20.88	7.89
UD-ICRC	-	-	-	2.75	3.49	9.97	12.40	20.76	26.46	12.64
UD-ICRC-attn	U	MAX	$\times$	7.27	10.53	17.12	24.10	30.44	42.22	21.95
	SO	MAX	$\times$	7.49	10.88	20.61	28.79	37.06	51.07	25.98
	SOU	MAX	$\times$	<b>15.71</b>	21.80	28.85	39.39	41.09	<b>55.73</b>	33.76
	SOUS	MEAN	$\checkmark$	2.74	4.53	8.76	13.71	19.27	28.05	12.84
	SOUS	MAX	$\checkmark$	10.72	15.43	21.59	30.26	34.43	47.34	26.63
	SOUS	MAX	$\times$	15.46	<b>21.81</b>	<b>29.55</b>	<b>40.72</b>	<b>41.14</b>	55.68	<b>34.06</b>
UD-ICRC-label	U	-	-	13.04	17.35	25.25	33.28	36.72	49.22	29.14
	SO	-	-	30.14	38.86	41.45	53.95	51.55	67.70	47.28
	SOU	-	-	32.17	41.38	43.57	56.68	53.65	70.81	49.71
	SOUS	-	-	<b>34.39</b>	<b>45.13</b>	<b>46.03</b>	<b>60.97</b>	<b>54.60</b>	<b>72.44</b>	<b>52.26</b>
Transformer-ICRC	-	-	-	2.05	2.70	6.86	9.19	16.21	21.91	9.82
Transformer-ICRC-attn	SOUS	MAX	$\times$	<b>17.52</b>	<b>24.96</b>	<b>31.88</b>	<b>44.46</b>	<b>43.71</b>	<b>61.10</b>	<b>37.27</b>
Transformer-ICRC-label	SOUS	-	-	<b>25.79</b>	<b>34.68</b>	<b>39.06</b>	<b>53.02</b>	<b>48.76</b>	<b>66.43</b>	<b>44.62</b>

relationships. We observe that the attention alignment module is more effective for the Transformer, which may imply that the attention in Transformer is more precise.

#### 4.5. Qualitative Results

In Figure 3 (b), we visualize the attention about the objects of each word (the purple heat map) during captioning and the pooled attention over all words (the reddish brown heat map). It can be observed that although the caption may be not so precise, the objects are still correctly attended (the *15\_woman* and *3\_bus* in the first sample, the *1\_court*, *6\_person*, and the *16\_tennis racket* in the second sample). The max pooling function highlights the mentioned objects in the caption. It’s also found that an object can be activated by several words, which gives an explanation for the performance drop when masking the non-noun words. In Figure 3 (c), we draw the scores of relationships for sorting. All the relationships are firstly sorted according to the assembling attention scores  $\beta$  induced from image caption, and then their  $\eta$  scores and the sentence likelihoods are drawn, which are used for sorting by UD-ICRC-attn and

UD-ICRC respectively. The line charts show that the predicted  $\eta$  scores have a similar trend with the  $\beta$  and therefore it can correctly rank the relationships according to their importance. However, the sentence likelihoods do not show this trend, suggesting that these scores (including the product scores used in traditional scene graph) are irrelevant to the importance of the relationships. In Figure 3 (d-e), we also compare the topic linguistic scene graph and the traditional scene graph from the motifs [70]. The topic linguistic scene graph focuses on the relationships of humans interest which are more important in the images. Besides, the scene graph of linguistic style allows the relationships to be expressed in a natural way with more suitable words, despite that the given detected object categories may be not so appropriate in the language context, e.g., in the first example, the *7\_photograph* is expressed as *station* in the relationships.

#### 4.6. Topic Scene Graph for Retrieval

As the topic scene graph provides relationships relevant to the major events in an image, it can be utilized for image

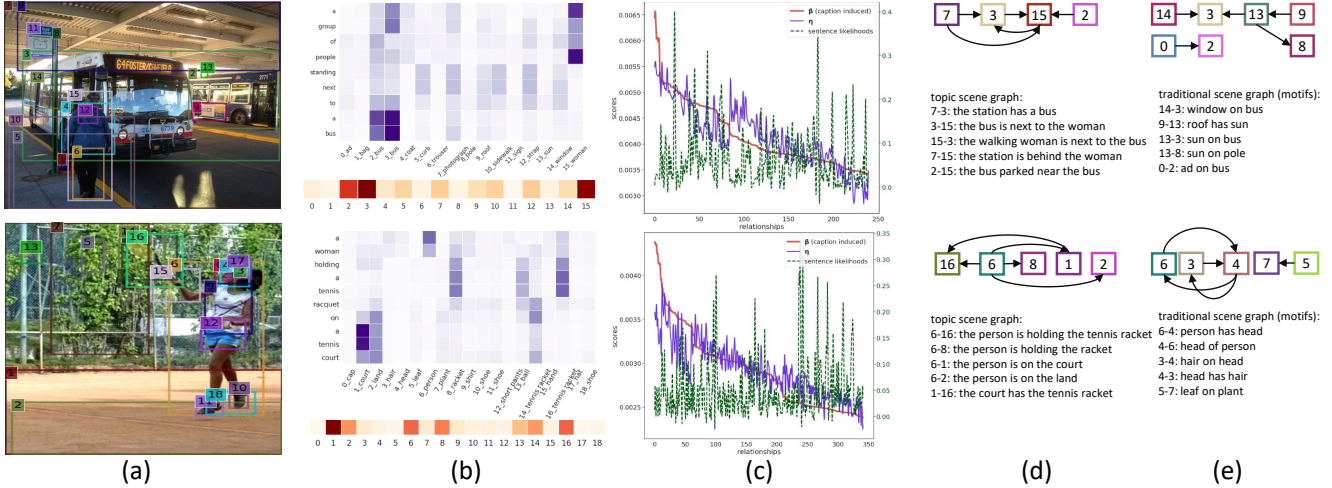


Figure 3. The qualitative results. (a) The objects with their bounding boxes and ids. (b) The attention about the objects during caption generation and the pooled attention are visualized. Darker colors indicate larger weights. (c) Importance scores of the relationships are drawn. Along the X-axis, the relationships are sorted by the  $\beta$  scores in a descending order. All the lines are smoothed. (d-e) The scene graph from our method and motifs [70] consisting of top 5 relationships are shown.

Table 4. The image retrieval results using top 1 relationships. We use the recall at K ( $R@K$ , higher is better) and the median rank of the target image (Med, lower is better).

Model	$R@1$	$R@5$	$R@10$	Med
TriLSTM	1.73	7.47	12.83	135.33
UD-ICRC	5.67	20.40	31.73	27.33
UD-ICRC-attn	<b>9.73</b>	<b>31.67</b>	<b>46.13</b>	<b>12.33</b>
UD-ICRC-label	17.77	49.17	67.37	5.67

retrieval [23, 50]. We adopt the classic image-text matching model SCAN [23]. 1,000 images are randomly chosen from the test set, and their top 1 or 5 relationships are collected as query for retrieving correct target images. The recall ( $R@K$ ,  $K$  is 1, 5, 10) and the median rank of the correctly retrieved images [21] are used as the metrics. We run through this process 3 times and report average results. Significant improvement brought by attention alignment is observed in Table 4. In addition, some major events can be decomposed into multiple relationships, *e.g.*, the major events of the query image in Figure 4 (left column) can be expressed with two relationships which are the top two given by our topic scene graph. If one directly uses the original image or traditional scene graph to retrieve similar images, the results may be not the desired ones. The proposed topic scene graph provides fine-grained descriptions of major events and makes it possible to designate the target content to be retrieved, *e.g.*, to retrieve *woman talking on telephone* or *woman sitting on bench*.

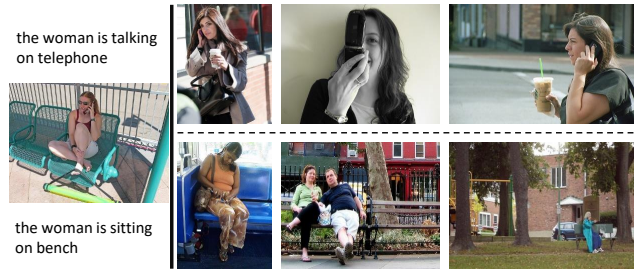


Figure 4. Two important relationships given by topic scene graph of the left image are used to retrieve similar images respectively. The results are shown on the right column.

## 5. Conclusion

In this work, we propose to generate the scene graph jointly with the image caption so that it can not only understand the image comprehensively, but also balance the important and trivial contents. The attention information from the image caption provides guideline to emphasize the important relationships. In addition, we generate the scene graph together with the image caption using a shared captioning module, making it express in a more natural style. Experimental results show the advantages of the proposed method in both performance and its feasibility in mining the important relationships without strong supervision. Besides, the topic scene graph has shown its practicality for controllable and fine-grained retrieval.

**Acknowledgements.** This work is partially supported by National Key R&D Program of China (2020AAA0105200), Natural Science Foundation of China under contracts Nos. U19B2036, 61922080, 61772500, and CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009.



## References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018. 3, 6, 7
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1
- [3] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72, 2006. 6
- [4] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5659–5667, 2017. 3
- [5] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9962–9971, 2020. 3
- [6] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6163–6171, 2019. 2
- [7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10578–10587, 2020. 3
- [8] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3298–3308, 2017. 2
- [9] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 376–380, 2014. 5
- [10] Santosh K Divvala, Ali Farhadi, and Carlos Guestrin. Learning everything about anything: Webly-supervised visual concept learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3270–3277, 2014. 2
- [11] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the Workshop on Automated Knowledge Base Construction*, pages 25–30, 2013. 2
- [12] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10323–10332, 2019. 3
- [13] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969–1978, 2019. 2
- [14] Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8529–8538, 2019. 2, 3
- [15] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11137–11147, 2019. 3
- [16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 3
- [17] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4634–4643, 2019. 3
- [18] Sho Inayoshi, Keita Otani, Antonio Tejero-de Pablos, and Tatsuya Harada. Bounding-box channels for visual relationship detection. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 12350, pages 682–697. Springer, 2020. 2
- [19] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, 2016. 3, 5
- [20] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678, 2015. 1, 2
- [21] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6271–6280, 2019. 2, 5, 6, 7, 8
- [22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1):32–73, 2017. 2
- [23] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 11208, pages 201–216. Springer, 2018. 8
- [24] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 8928–8937, 2019. 3
- [25] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia (TMM)*, 21(8):2117–2130, 2019. 1, 3

- [26] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7244–7253, 2017. 2
- [27] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 11205, pages 346–363. Springer, 2018. 2
- [28] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1261–1270, 2017. 2
- [29] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10403–10412, 2019. 2
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755. Springer, 2014. 5
- [31] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3746–3755, 2020. 2
- [32] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 9905, pages 852–869. Springer, 2016. 2
- [33] Jianming Lv, Qinzhe Xiao, and Jiajie Zhong. Avr: Attention based salient visual relationship detection. *arXiv preprint arXiv:2003.07012*, 2020. 2
- [34] Li Mi and Zhenzhong Chen. Hierarchical graph attention network for visual relationship detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13886–13895, 2020. 2
- [35] Victor Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? *arXiv preprint arXiv:2009.12313*, 2020. 1, 3
- [36] George A. Miller. Wordnet: A lexical database for english. *Communication of the ACM*, 38(11):39–41, 1992. 5
- [37] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10971–10980, 2020. 3
- [38] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5179–5188, 2017. 2, 4
- [39] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3957–3966, 2019. 2
- [40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 91–99, 2015. 4, 5
- [41] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7008–7024, 2017. 3, 5
- [42] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1745–1752, 2011. 2
- [43] 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*, pages 70–80, 2015. 5
- [44] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8376–8384, 2019. 1
- [45] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 966–973. IEEE, 2010. 3
- [46] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3716–3725, 2020. 2
- [47] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6619–6628, 2019. 1, 2
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017. 3, 6, 7
- [49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, 2015. 3
- [50] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1508–1517, 2020. 8
- [51] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation. In *Proceedings of the IEEE*

- Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8188–8197, 2019. [2](#)
- [52] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 12358, pages 222–239. Springer, 2020. [1](#), [2](#), [5](#)
- [53] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6629–6638, 2019. [1](#)
- [54] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. [5](#)
- [55] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5410–5419, 2017. [2](#), [5](#)
- [56] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015. [3](#)
- [57] Ning Xu, An-An Liu, Jing Liu, Weizhi Nie, and Yuting Su. Scene graph captioner: Image captioning based on structural visual representation. *Journal of Visual Communication and Image Representation*, 58:477–485, 2019. [3](#)
- [58] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 11205, pages 690–706. Springer, 2018. [2](#)
- [59] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10685–10694, 2019. [3](#)
- [60] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29, 2016. [3](#)
- [61] Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010. [3](#)
- [62] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 11218, pages 711–727. Springer, 2018. [3](#)
- [63] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2621–2629, 2019. [3](#)
- [64] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 11207, pages 330–347. Springer, 2018. [2](#)
- [65] Fan Yu, Haonan Wang, Tongwei Ren, Jinhui Tang, and Gangshan Wu. Visual relation of interest detection. In *ACM International Conference on Multimedia (MM’20)*, pages 1386–1394, 2020. [1](#), [2](#)
- [66] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1974–1982, 2017. [2](#)
- [67] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 12368, pages 606–623. Springer, 2020. [2](#)
- [68] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3736–3745, 2020. [2](#)
- [69] Alireza Zareian, Haoxuan You, Zhecan Wang, and Shih-Fu Chang. Learning visual commonsense for robust scene graph generation. In *Proceedings of European Conference on Computer Vision (ECCV)*, volume 12368, pages 642–657. Springer, 2020. [2](#)
- [70] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840, 2018. [2](#), [7](#), [8](#)
- [71] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5532–5540, 2017. [2](#)
- [72] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4233–4241, 2017. [2](#)
- [73] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. Large-scale visual relationship understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 9185–9194, 2019. [2](#)
- [74] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11535–11543, 2019. [2](#)
- [75] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4777–4786, 2020. [5](#)