Sketching Image Gist: Human-Mimetic Hierarchical Scene Graph Generation

Wenbin Wang[1,2[0000−0002−4394−0145]], Ruiping Wang[1,2[0000−0003−1830−2595]], Shiguang Shan[1,2[0000−0002−8348−392X]], and Xilin Chen[1,2[0000−0003−3024−4404]]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
wenbin.wang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

Abstract. Scene graph aims to faithfully reveal humans' perception of image content. When humans analyze a scene, they usually prefer to describe image gist first, namely major objects and key relations in a scene graph. This humans' inherent perceptive habit implies that there exists a hierarchical structure about humans' preference during the scene parsing procedure. Therefore, we argue that a desirable scene graph should be also hierarchically constructed, and introduce a new scheme for modeling scene graph. Concretely, a scene is represented by a human-mimetic **H**ierarchical **E**ntity **T**ree (HET) consisting of a series of image regions. To generate a scene graph based on HET, we parse HET with a Hybrid Long Short-Term Memory (Hybrid-LSTM) which specifically encodes hierarchy and siblings context to capture the structured information embedded in HET. To further prioritize key relations in the scene graph, we devise a Relation Ranking Module (RRM) to dynamically adjust their rankings by learning to capture humans' subjective perceptive habits from objective entity saliency and size. Experiments indicate that our method not only achieves state-of-the-art performances for scene graph generation, but also is expert in mining image-specific relations which play a great role in serving downstream tasks.

Keywords: Image Gist, Key Relation, Hierarchical Entity Tree, Hybrid-LSTM, Relation Ranking Module

1 Introduction

In an effort to thoroughly understand a scene, scene graph generation (SGG) [10,42] in which objects and pairwise relations should be detected, has been on the way to bridge the gap between low-level recognition and high-level cognition, and contributes to tasks like image captioning [40,25,44], VQA [1,36], and visual reasoning [31]. While previous works [42,17,43,16,50,28,36,39,49,53] have pushed this area forward, the generated scene graph may be still far from perfect, e.g., they seldom consider whether the detected relations are what humans want to convey from the image or not. As a symbolic representation of an image, the scene graph is expected to record the image content as complete as possible.

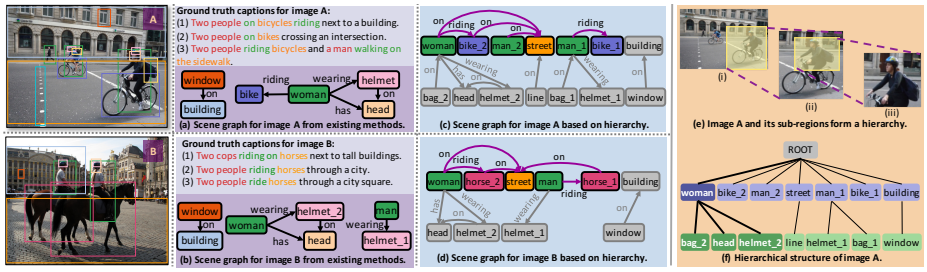


Fig. 1: Scene graphs from existing methods shown in (a) and (b) fail in sketching the image gist. The hierarchical structure about humans' perception preference is shown in (f), where the bottom left highlighted branch stands for the hierarchy in (e). The scene graphs in (c) and (d) based on hierarchical structure better capture the gist. Relations in (a) and (b), and purple arrows in (c) and (d), are top-5 relations, while gray ones in (c) and (d) are secondary.

More importantly, a scene graph is not just for being admired, but for supporting downstream tasks, such as image captioning, where a description is supposed to depict the major event in the image, or the namely **image gist**. This characteristic is also one of the humans' inherent habits when they parse a scene. Therefore, an urgently needed feature of SGG is to assess the relation importance and prioritize the relations which form the major events that humans intend to preferentially convey, i.e., **key relations**. This is seldom considered by existing methods. What's worse, the universal phenomenon of unbalanced distribution of relationship triplets in mainstream datasets exacerbates the problem that the major event cannot be found out. Let's study the quality of top relations predicted by existing state-of-the-art methods (e.g., [49]) and check whether they are "key" or not. In Figure 1(a)(b), two scene graphs shown with top-5 relations for image A and B are mostly the same although major events in A and B are quite different. In other words, existing methods are deficient in mining image-specific relations, but biased towards trivial or self-evident ones (e.g., $\langle$*woman*, *has*, *head*$\rangle$ can be obtained from commonsense without observing the image), which fail in conveying image gist (colored parts in ground truth captions in Figure 1), and barely contribute to downstream tasks.

Any pair of objects in a scene can be considered relevant, at least in terms of their spatial configurations. Faced with such a massive amount of relations, how do humans choose relations to describe the images? Given picture (ii) in Figure 1(e), a zoom-in sub-region of picture (i), humans will describe it with $\langle$*woman*, *riding*, *bike*$\rangle$, since *woman* and *bike* belong to the same perceptive level and their interaction forms the major event in (ii). When it comes to picture (iii), the answers would be $\langle$*woman*, *wearing*, *helmet*$\rangle$ and $\langle$*bag*, *on*, *woman*$\rangle$, where *helmet* and *bag* are finer details of *woman* and belong to an inferior perceptive level. It suggests that there naturally exists a hierarchical structure about humans' perception preference, as shown in Figure 1(f).

Inspired by observations above, we argue that a desirable scene graph should be hierarchically constructed. Specifically, we represent the image with a human-mimetic Hierarchical Entity Tree (HET) where each node is a detected object and each one can be decomposed into a set of finer objects attached to it. To generate the scene graph based on HET, we devise Hybrid Long Short-Term Memory (Hybrid-LSTM) to encode both hierarchy and siblings context [49,36] and capture the structured information embedded in HET, considering that important related pairs are more likely to be seen either inside a certain perceptive level or between two adjacent perceptive levels. We further intend to evaluate the performances of different models on key relation prediction but the annotations of key relations are not directly available from existing datasets. Therefore, we extend Visual Genome (VG) [13] to VG-KR dataset which contains indicative annotations of key relations by drawing support from caption annotations in MSCOCO [21]. We devise a Relation Ranking Module to adjust the rankings of relations. It captures humans' subjective perceptive habits from objective entity saliency and size, and achieves ultimate performances on mining key relations.[1]

2 Related Works

Scene graph generation (SGG) and Visual Relationship Detection (VRD), are the two most common tasks aiming at extracting interaction between two objects. In the field of VRD, various studies [24,3,15,50,47,27,51,46,52] mainly focus on detecting each relation triplet independently rather than describe the structure of the scene. The concept of scene graph is firstly proposed in [10] for image retrieval. Xu et al. [42] define SGG task and creatively devise message passing mechanism for scene graph inference. A series of succeeding works struggle to design various approaches to improve the graph representation. Li et al. [17] induce image captions and object information to jointly address multitasks. [49,36,39,22] draw support from useful context construction. Yang et al. [43] propose Graph-RCNN to embed the structured information. Qi et al. [28] employ a self-attention module to embed a weighted graph representation. Zhang et al. [53] propose contrastive losses to resolve the related pair configuration ambiguity. Zareian et al. [48] creatvely treat the SGG as an edge role assignment problem. Recently, some methods try to borrow advantages from using knowledge [2,5] or causal effect [35] to diversify the predicted relations. Liang et al. [19] prune the dominant and easy-to-predict relations in VG to alleviate the annihilation problem of rare but meaningful relations.

Structured Scene Parsing, has been paid much attention in pursuit of higher-level scene understanding. [33,30,20,6,55,45] construct various hierarchical structures for their specific tasks. Unlike existing SGG studies that indiscriminately detect relations no matter whether they are concerned by humans or not, our work introduces the idea of hierarchical structure into SGG task, and try to give priority to detect key relations, then the trivial ones for completeness.

[1] Source code and dataset are available at http://vipl.ict.ac.cn/resources/codes or https://github.com/Kenneth−Wong/het−eccv20.git.

Saliency vs. Image Gist. An extremely rich set of studies [14,37,23,38,8,54] focus on analyzing where humans gaze and find visually salient objects (high contrast of luminance, hue, and saturation, center position [9,12,41], etc.). It's notable that the visually salient objects are related but not equal to objects involved in image gist. He et al. [7] explore gaze data and find that only 48% of fixated objects are referred in humans' descriptions about the image, while 95% of objects referred in descriptions are fixated. It suggests that objects referred in a description (i.e., objects that humans think important and should form the major events / image gist) are almost visually salient and reveal where humans gaze, but what humans fixate (i.e., visually salient objects) are not always what they want to convey. We provide some examples in supplementary materials to help to understand this finding. Naturally, we need to emphasize that the levels in our HET reflect the perception priority level rather than the object saliency. Besides, this finding supports us to obtain the indicative annotations of key relations with the help of image caption annotations.

3 Proposed Approach

3.1 Overview

The scene graph $\mathcal{G} = \{\mathcal{O}, \mathcal{R}\}$ of an image $\mathcal{I}$ contains a set of entities $\mathcal{O} = \{o_i\}_{i=1}^{N}$ and their pairwise relations $\mathcal{R} = \{r_k\}_{k=1}^{M}$. Each r_k is a triplet $\langle o_i, p_{ij}, o_j \rangle$ where $p_{ij} \in \mathcal{P}$ and $\mathcal{P}$ is the set of all predicates. As illustrated in Figure 2, our approach can be summarized into four steps. (i) We apply Faster R-CNN [29] with VGG16 [32] backbone to detect all the entity proposals and each of them possesses its bounding box $\boldsymbol{b}_i \in \mathbb{R}^4$, 4,096-dimensional visual feature $\boldsymbol{v}_i$, and the class probability vector $\boldsymbol{q}_i$ from the softmax output. (ii) In Section 3.2, HET is constructed by organizing the detected entities according to their perceptive levels. (iii) In Section 3.3, we design the Hybrid-LSTM network to parse HET, which firstly encodes the structured context then decodes it for graph inference. (iv) In Section 3.4, we improve the scene graph generated in (iii) with our devised RRM which further adjusts the rankings of relations and shifts the graph focus to the relations between entities that are close to top perceptive levels of HET.

3.2 HET Construction

We aim to construct a hierarchical structure whose top-down levels are accord with the perceptive levels of humans' inherent scene parsing hierarchy. From a massive number of observations, it can be found that entities with larger sizes are relatively more likely to form the major events in a scene (this will be proved effective through experiments). Therefore, we arrange larger entities as close to the root of HET as possible. Each entity can be decomposed into finer entities that make up the inferior level.

Concretely, HET is a multi-branch tree $\mathcal{T}$ with a virtual root o_0 standing for the whole image. All the entities are sorted in descending order according to

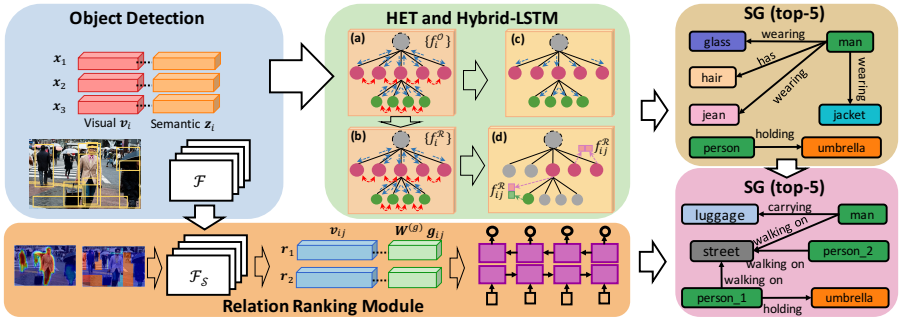


Fig. 2: An overview of our method. An object detector is firstly applied to give support to HET construction. Then Hybrid-LSTM is leveraged to parse HET, and specifically contains 4 processes, (a) entity context encoding, (b) relation context encoding, (c) entity context decoding, and (d) relation context decoding. Finally, RRM predicts a ranking score for each triplet which further prioritizes the key relations in the scene graph.

their sizes and we get an orderly sequence $\{o_{i_1}, o_{i_2}, \ldots, o_{i_N}\}$. For each entity o_{i_n}, we consider entities with larger size, $\{o_{i_m}\}, 1 \leq m < n$, and calculate the ratio

$$P_{nm} = \frac{I\left(o_{i_n}, o_{i_m}\right)}{A(o_{i_n})}, \tag{1}$$

where $A(\cdot)$ denotes the size of the entity and $I(\cdot, \cdot)$ is the intersection area of two entities. If P_{nm} is larger than threshold T, o_{i_m} will be a candidate parent node of o_{i_n} since o_{i_m} contains most part of o_{i_n}. If there is no candidate, the parent node of o_{i_n} is set as o_0. If there are more than one, we further determine the parent with two alternative strategies:

Area-first Strategy (AFS). Considering that entity with a larger size has a higher probability to contain more details or components, the candidate with the largest size is selected to be a parent node.

Intersection-first Strategy (IFS). We compute ratio

$$Q_{nm} = \frac{I\left(o_{i_n}, o_{i_m}\right)}{A(o_{i_m})}. \tag{2}$$

A larger Q_{nm} means that o_{i_n} is relatively more important to o_{i_m} than to other candidates. Therefore, o_{i_m} where $m = \arg\max_k Q_{nk}$ is chosen as parent of o_{i_n}.

3.3 Structured Context Encoding and Scene Graph Generation

The interpretability of HET implies that important relations are more likely to be seen between entities either inside a certain level or from two adjacent levels. Therefore, both hierarchical connection [36] and sibling association [49] are useful for context modeling. Our Hybrid-LSTM encoder is proposed, which consists of

a bidirectional multi-branch TreeLSTM [34] (Bi-TreeLSTM) for encoding the hierarchy context, and a bidirectional chain LSTM [4] (Bi-LSTM) for encoding the siblings context. We use two identical Hybrid-LSTM encoders to encode two types of context for each entity, one is **entity context** which helps predict the information of entity itself, and the other is **relation context** which plays a role in inferring the relation when interacting with other potential relevant entities. For brevity we only provide a detailed introduction of entity context encoding (Figure 2(a)). Specifically, the input feature $\boldsymbol{x}_i$ of each node o_i is concatenation of visual feature $\boldsymbol{v}_i$ and weighted sum of semantic embedding vectors, $\boldsymbol{z}_i = \boldsymbol{W}_e^{(1)}\boldsymbol{q}_i$, where $\boldsymbol{W}_e^{(1)}$ is word embedding matrix initialized from GloVe [26]. For the root node o_0, $\boldsymbol{v}_0$ is obtained with the whole-image bounding box, while $\boldsymbol{z}_0$ is initialized randomly.

The hierarchy context (blue arrows in Figure 2(a)) is encoded as:

$$\boldsymbol{C} = \text{BiTreeLSTM}(\{\boldsymbol{x}_i\}_{i=0}^N), \tag{3}$$

where $\boldsymbol{C} = \{\boldsymbol{c}_i\}_{i=0}^N$ and each $\boldsymbol{c}_i = \left[\overrightarrow{\boldsymbol{h}_i^{\mathcal{T}}}; \overleftarrow{\boldsymbol{h}_i^{\mathcal{T}}}\right]$ is the concatenation of the top-down and bottom-up hidden states of Bi-TreeLSTM:

$$\overrightarrow{\boldsymbol{h}_i^{\mathcal{T}}} = \text{TreeLSTM}\left(\boldsymbol{x}_i, \overrightarrow{\boldsymbol{h}_p^{\mathcal{T}}}\right), \tag{4a}$$

$$\overleftarrow{\boldsymbol{h}_i^{\mathcal{T}}} = \text{TreeLSTM}\left(\boldsymbol{x}_i, \left\{\overleftarrow{\boldsymbol{h}_j^{\mathcal{T}}} \middle| j \in C(i)\right\}\right), \tag{4b}$$

where $C(\cdot)$ denotes the set of children nodes while subscript p denotes the parent of node i.

The siblings context (red arrows in Figure 2(a)) is encoded within each set of children nodes which share the same parent:

$$\boldsymbol{S} = \text{BiLSTM}(\{\boldsymbol{x}_i\}_{i=0}^N), \tag{5}$$

where $\boldsymbol{S} = \{\boldsymbol{s}_i\}_{i=0}^N$ and each $\boldsymbol{s}_i = \left[\overrightarrow{\boldsymbol{h}_i^{\mathcal{L}}}; \overleftarrow{\boldsymbol{h}_i^{\mathcal{L}}}\right]$ is concatenation of forward and backward hidden states of Bi-LSTM:

$$\overrightarrow{\boldsymbol{h}_i^{\mathcal{L}}} = \text{LSTM}\left(\boldsymbol{x}_i, \overrightarrow{\boldsymbol{h}_l^{\mathcal{L}}}\right), \ \overleftarrow{\boldsymbol{h}_i^{\mathcal{L}}} = \text{LSTM}\left(\boldsymbol{x}_i, \overleftarrow{\boldsymbol{h}_r^{\mathcal{L}}}\right), \tag{6}$$

where l and r stand for left and right sibling which share the same parent with i. We further concatenate hierarchy and siblings context to obtain the entity context, $\boldsymbol{f}_i^{\mathcal{O}} = [\boldsymbol{c}_i; \boldsymbol{s}_i]$. Missing branches or siblings are padded with zero vectors.

The relation context is encoded (Figure 2(b)) in the same way as entity context except that the input of each node is replaced by $\{\boldsymbol{f}_i^{\mathcal{O}}\}_{i=0}^N$. Another Hybrid-LSTM encoder is applied to get the relation context $\{\boldsymbol{f}_i^{\mathcal{R}}\}_{i=0}^N$.

To generate a scene graph, we should decode the context to obtain entity and relation information. In HET, a child node strongly depends on its parent, i.e., information of parent node is helpful for prediction of child node. Therefore,

to predict entity information, we decode entity context in a top-down manner following Eq. (4a) as shown in Figure 2(c). For node o_i, the input $\boldsymbol{x}_i$ in Eq. (4a) is replaced with $[\boldsymbol{f}_i^{\mathcal{O}}; \boldsymbol{W}_e^{(2)}\boldsymbol{q}_p]$, where $\boldsymbol{W}_e^{(2)}$ is word embedding matrix and $\boldsymbol{q}_p$ is the predicted class probability vector of the parent of o_i. The output hidden state is fed into a softmax classifier and bounding box regressor to predict entity information of o_i. To predict the predicate p_{ij} between o_i and o_j, we feed $\boldsymbol{f}_{ij}^{\mathcal{R}} = [\boldsymbol{f}_i^{\mathcal{R}}; \boldsymbol{f}_j^{\mathcal{R}}]$ to an MLP classifier (Figure 2(d)). As a result, a scene graph is generated, and for each triplet containing subject o_i, object o_j and predicate p_{ij}, we obtain their scalar scores s_i, s_j, and s_{ij}.

3.4 Relation Ranking Module

So far, we obtain the hierarchical scene graph based on HET. As we collect the key relation annotations (Section 4.1), we intend to further maximize the performance on mining key relations with supervised information, and explore the advantages brought by HET. Consequently, we design a Relation Ranking Module (RRM) to prioritize key relations. As analyzed in **Related Works**, regions of humans' interest can be tracked under the guidance of *visual saliency* although they do not always form the major events that humans want to convey. Besides, the *size*, which guides HET construction, not only is an important reference for estimating the perceptive level of entities, but also is found helpful to rectify some misleadings in humans' subjective assessment on the importance of relations (see supplementary materials). Therefore, we propose to learn to capture humans' subjective assessment on the importance of relations under the guidance of visual saliency and entity size information.

We firstly employ DSS [8] to predict the pixel-wise saliency map (**SM**) $\mathcal{S}$ for each image. To effectively collect entity size information, we propose a pixel-wise area map (**AM**) $\mathcal{A}$. Given the image $\mathcal{I}$ and its detected N entities $\{o_i\}_{i=1}^N$ with bounding boxes $\{\boldsymbol{b}_i\}_{i=1}^N$ (specially o_0 and $\boldsymbol{b}_0$ for the whole image), the value a_{xy} of each position (x, y) on $\mathcal{A}$ is defined as the minimum normalized size of entities which cover (x, y):

$$a_{xy} = \begin{cases} \min\left\{ \dfrac{A(o_i)}{A(o_0)} \middle| i \in \mathcal{X} \right\}, \text{if } \mathcal{X} \neq \emptyset \\ 0, \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{X} = \{i | (x, y) \in \boldsymbol{b}_i, 0 < i \leq N\}$. The sizes of both $\mathcal{S}$ and $\mathcal{A}$ are the same as that of input image $\mathcal{I}$. We apply adaptive average pooling ($\text{AAP}(\cdot)$) to smooth and down-sample these two maps to align with the shape of conv5 feature map $\mathcal{F}$ from Faster-RCNN, and obtain the attention embedded feature map $\mathcal{F}_S$:

$$\mathcal{F}_S = \mathcal{F} \odot (\text{AAP}(\mathcal{S}) + \text{AAP}(\mathcal{A})), \tag{8}$$

where $\odot$ is the Hadamard product.

We predict a score for each triplet to adjust their rankings. The input contains visual representation for a triplet, $\boldsymbol{v}_{ij} \in \mathbb{R}^{4096}$, which is obtained by RoI Pooling

on $\mathcal{F}_S$. Besides, the geometric information is also an auxiliary cue for estimating the importance. For a triplet containing subject box $\boldsymbol{b}_i$ and object box $\boldsymbol{b}_j$, the geometric feature $\boldsymbol{g}_{ij}$ is defined as a 6-dimensional vector following [11]:

$$\boldsymbol{g}_{ij} = \left[\frac{x_j - x_i}{\sqrt{w_i h_i}}, \frac{y_j - y_i}{\sqrt{w_i h_i}}, \sqrt{\frac{w_j h_j}{w_i h_i}}, \frac{w_i}{h_i}, \frac{w_j}{h_j}, \frac{\boldsymbol{b}_i \cap \boldsymbol{b}_j}{\boldsymbol{b}_i \cup \boldsymbol{b}_j}\right], \tag{9}$$

which is projected to a 256-dimensional vector and concatenated with $\boldsymbol{v}_{ij}$, resulting in the final representation for a relation $\boldsymbol{r}_{ij} = [\boldsymbol{v}_{ij}; \boldsymbol{W}^{(g)}\boldsymbol{g}_{ij}]$ where $\boldsymbol{W}^{(g)} \in \mathbb{R}^{256\times 6}$ is projection matrix. Then we use a bi-directional LSTM to encode global context among all the triplets so that ranking score of each triplet can be reasonably adjusted considering scores of other triplets. Concretely, the ranking score t_{ij} for a pair (o_i, o_j) is achieved as:

$$\{\boldsymbol{h}_{ij}^{\mathcal{R}}\} = \mathrm{BiLSTM}\left(\{\boldsymbol{r}_{ij}\}\right), \tag{10}$$

$$t_{ij} = \boldsymbol{W}_2^{(r)}\mathrm{ReLU}(\boldsymbol{W}_1^{(r)}\boldsymbol{h}_{ij}^{\mathcal{R}}). \tag{11}$$

$\boldsymbol{W}_1^{(r)}$ and $\boldsymbol{W}_2^{(r)}$ are weights of two fully connected layers. The ranking score is fused with classification scores so that both the confidences of three components of a triplet and ranking priority are considered, resulting in the final ranking confidence $\phi_{ij} = s_i \cdot s_j \cdot s_{ij} \cdot t_{ij}$, which is used for re-ranking the relations.

3.5 Loss Function

We adopt the cross-entropy loss for optimizing Hybrid-LSTM networks. Let e' and l' denote the predicted label of entity and predicate respectively, e and l denote the ground truth labels. The loss is defined as:

$$\mathcal{L}_{CE} = \mathcal{L}_{entity} + \mathcal{L}_{relation} = -\frac{1}{Z_1}\sum_i e'_i \log(e_i) - \frac{1}{Z_2}\sum_i\sum_{j\neq i} l'_{ij}\log(l_{ij}). \tag{12}$$

When the RRM is applied, the final loss function is the sum of $\mathcal{L}_{CE}$ and ranking loss $\mathcal{L}(\mathcal{K},\mathcal{N})$, which is used to maximize the margin between the ranking confidences of key relations and those of secondary ones:

$$\mathcal{L}_{Final} = \mathcal{L}_{CE} + \mathcal{L}(\mathcal{K},\mathcal{N}) = \mathcal{L}_{CE} + \frac{1}{Z_3}\sum_{r\in\mathcal{K}, r'\in\mathcal{N}} \max(0, \gamma - \phi_r + \phi_{r'}), \tag{13}$$

where γ denotes margin parameter, $\mathcal{K}$ and $\mathcal{N}$ stand for the set of key and secondary relations, r and r' are relations sampled from $\mathcal{K}$ and $\mathcal{N}$ with ranking confidences ϕ_r and $\phi_{r'}$. Z_1, Z_2, and Z_3 are normalization factors.

4 Experimental Evaluation

4.1 Dataset, Evaluation and Settings

VRD [24], is the benchmarking dataset for visual relationship detection task, which contains 4,000/1,000 training/test images and covers 100 object categories and 70 predicate categories.

Visual Genome (VG), is a large-scale dataset with rich annotations of objects, attributes, dense captions and pairwise relationships, containing 75,651/32,422 training/test images. We adopt the most widely used version of VG, namely VG150 [42], which covers 150 object categories and 50 predicate categories.
VG200 and VG-KR. We intend to collect the indicative annotations of key relations based on VG. Inspired by the finding illustrated in **Related Works**, we associate the relation triplets referred in caption annotations in MSCOCO [21] with those from VG. The details of our processing and more statistics are provided in supplementary materials.
Evaluation, Settings, and Implementation Details. For conventional SGG following triplet-match rule (only if three components of a triplet match the ground truth will it be a correct one), we adopt three universal protocols [42]: PREDCLS, SGCLS, and SGGEN. All protocols use Recall@K (R@K=20, 50, 100) as a metric. When evaluating key relation prediction, there are some variations. First, we only evaluate with PREDCLS and SGCLS protocols to eliminate the interference of errors from object detector, and add a tuple-match rule (only the subject and object are required to match the ground truth) to investigate the ability to find proper pairs. Second, we introduce a new metric, **Key Relation Recall (kR@K)**, which computes recall rate on key relations. As the number of key relations is usually less than 5 (see supplementary materials), the K in kR@K is set to 1 and 5. When evaluating on VRD, we use RELDET and PHRDET [47], and report R@50 and R@100 at 1, 10, and 70 predicates per related pair. The details about the hyperparameters settings and implementation are provided in supplementary materials.

4.2 Ablation Studies

Ablation studies are separated into two sections. The first part is to explore some variants of HET construction. We conduct these experiments on VG150. The complete version of our model is **HetH**, which is configured with IFS and Hybrid-LSTM. The second part is an investigation into the usage of SM and AM in RRM. Experiments are carried out on VG-KR. The complete version is **HetH-RRM**, whose implementation follows Eq. (8).

Ablation study on HET construction. We firstly compare **AFS** and **IFS** for determining the parent node. Then we investigate the effectiveness of the chain LSTM encoder in Hybrid-LSTM. The ablative models mentioned above are shown in Table 1 as **HetH-AFS** (i.e.replace IFS by AFS), and **HetH w/o chain**. We observe that using IFS together with Hybrid-LSTM encoder has the best performances, which indicates that HET would be more reasonable using IFS. It's noteworthy that if the Bi-TreeLSTM encoder is abandoned, the Hybrid-LSTM encoder would almost degenerate to MOTIFS. Therefore, through comparisons between HetH and MOTIFS, HetH and HetH w/o chain, it implies that both hierarchy and siblings context should be encoded in HET.

Ablation study on RRM. In order to explore the effectiveness of saliency and size, we ablate HetH-RRM with the following baselines: (1) **RRM-Base:** $\boldsymbol{v}_{ij}$ is extracted from $\mathcal{F}$ rather than $\mathcal{F}_S$, (2) **RRM-SM:** only $\mathcal{S}$ is used, and (3)

Table 1: Results table (%) on VG150 and VG200. The results of the full version of our method are highlighted.

		SGGEN			SGCLS			PREDCLS		
	R@	20	50	100	20	50	100	20	50	100
VG150	VRD [24]	-	0.3	0.5	-	11.8	14.1	-	27.9	35.0
	IMP [42]	-	3.4	4.2	-	21.7	24.4	-	44.8	53.0
	IMP† [42,49]	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3
	Graph-RCNN [43]	-	11.4	13.7	-	29.6	31.6	-	54.2	59.1
	MemNet [39]	7.7	11.4	13.9	23.3	27.8	29.5	42.1	53.2	57.9
	MOTIFS [49]	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1
	KERN [2]	-	27.1	29.8	-	36.7	37.4	-	65.8	67.6
	VCTree-SL [36]	21.7	27.7	31.1	35.0	37.9	38.6	59.8	66.2	67.9
	HetH-AFS	21.2	27.1	30.5	33.7	36.6	37.3	58.1	64.7	66.6
	HetH w/o chain	21.5	27.4	30.7	32.9	35.9	36.7	57.5	64.5	66.5
	HetH	**21.6**	**27.5**	**30.9**	**33.8**	**36.6**	**37.3**	**59.8**	**66.3**	**68.1**
VG200	MOTIFS [49]	15.2	19.9	22.8	24.5	26.7	27.4	52.5	59.0	61.0
	VCTree-SL [36]	14.7	19.5	22.5	24.2	26.5	27.1	51.9	58.4	60.3
	HetH	**15.7**	**20.4**	**23.4**	**25.0**	**27.2**	**27.8**	**53.6**	**60.1**	**61.8**

Table 2: Results table (%) of key relation prediction on VG-KR.

	Triplet Match				Tuple Match			
	SGCLS		PREDCLS		SGCLS		PREDCLS	
kR@	1	5	1	5	1	5	1	5
VCTree-SL	5.7	14.2	11.4	30.2	8.4	22.2	16.1	46.4
MOTIFS	5.9	14.5	11.3	30.0	8.5	21.8	16.0	46.2
HetH	**6.1**	**15.1**	**11.6**	**30.4**	**8.6**	**22.7**	**16.4**	**47.1**
MOTIFS-RRM	8.6	16.4	16.7	33.8	13.8	26.3	27.9	57.1
HetH-RRM	**9.2**	**17.1**	**17.5**	**35.0**	**14.6**	**27.3**	**28.9**	**59.1**
RRM-Base	8.4	16.8	16.2	33.7	13.4	26.8	26.6	57.2
RRM-SM	9.0	16.9	17.2	34.5	14.3	27.1	28.6	58.7
RRM-AM	8.9	16.9	16.9	34.4	14.1	27.0	28.1	58.2

Table 3: Results table (%) on VRD.

	RELDET						PHRDET					
	k=1		k=10		k=70		k=1		k=10		k=70	
R@	50	100	50	100	50	100	50	100	50	100	50	100
ViP [15]	17.32	20.01	-	-	-	-	22.78	27.91	-	-	-	-
VRL [18]	18.19	20.79	-	-	-	-	21.37	22.60	-	-	-	-
KL-Dist [47]	19.17	21.34	22.56	29.89	22.68	31.89	23.14	24.03	26.47	29.76	26.32	29.43
Zoom-Net [46]	18.92	21.41	-	-	21.37	27.30	24.82	28.09	-	-	29.05	37.34
RelDN-L_0 [53]	24.30	27.91	26.67	32.55	26.67	32.55	31.09	36.42	33.29	41.25	33.29	41.25
RelDN [53]	25.29	28.62	28.15	33.91	28.15	33.91	31.34	36.42	34.45	42.12	34.45	42.12
HetH	22.42	24.88	26.88	31.69	26.88	31.81	30.69	35.59	**35.47**	**42.94**	**35.47**	**43.05**

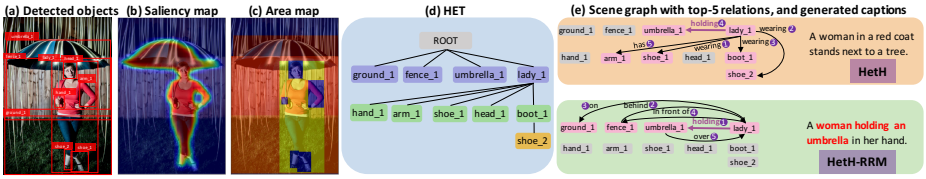


Fig. 3: Qualitative Results of HetH and HetH-RRM. In (e), the pink entities are involved in top-5 relations, and the purple arrows are key relations matched with ground truth. The purple numeric tags next to the relations are the rankings, and "1" means that the relation gets the highest score.

RRM-AM: only $\mathcal{A}$ is used. Results in Table 2 suggest that both saliency and size information indeed contributes to discovering key relations, and the effect of saliency is slightly better than that of the size. The hybrid version achieves the highest performances. From the following qualitative analysis, we can see that with the guidance of saliency and rectification effect of size, RRM further shifts the model's attention to key relations significantly.

4.3 Comparisons with State-of-the-Arts

For scene graph generation, we compare our **HetH** with the following state-of-the-art methods: **VRD** [24] and **KERN** [2] use knowledge from language or statistical correlations. **IMP** [42], **Graph-RCNN** [43], **MemNet** [39], **MOTIFS** [49] and **VCTree-SL** [36] mainly devise various message passing methods for improving graph representations. For key relation prediction, we mainly evaluate two latest works, MOTIFS and VCTree-SL on VG-KR. Besides, we further incorporate RRM to MOTIFS, namely **MOTIFS-RRM**, to explore the transferability of RRM. Results are shown in Table 1 and 2. We give statistical significance of the results in the supplementary materials.

Quantitative Analysis. In Table 1, when evaluated on **VG150**, HetH dominantly surpasses most methods. Compared to MOTIFS and VCTree-SL, HetH using multi-branch tree structure outperforms MOTIFS and yields comparable recall rate with VCTree-SL which uses a binary tree structure. It indicates that

hierarchical structure is superior to plain one in terms of modeling context. We observe that HetH achieves better performances compared to VCTree-SL under PREDCLS protocol, while there exists a slight gap under SGCLS and SGGEN protocols. This is mainly because our tree structure is generated with artificial rules and some incorrect subtrees inevitably emerge due to occlusion in 2D images, while VCTree-SL dynamically adjusts its structure in pursuit of higher performances. Under SGCLS and SGGEN protocols in which object information is fragmentary, it is difficult for HetH to rectify the context encoded from wrong structures. However, we argue that our interpretable and natural multi-branch tree structure is also adaptive to the situation when there is an increment of object and relation categories but fewer data. It can be seen from evaluation results on **VG200** that the HetH outperforms MOTIFS by 0.67 mean points and VCTree-SL by 1.1 mean points. On the contrary, in this case, the data are insufficient for dynamic structure optimization.

As SGG task is highly related to VRD task, we apply HetH on **VRD** and the comparison results are shown in Table 3. Both the HetH and RelDN [53] use pretrained weights on MSCOCO, while only [46] states that they use ImageNet pretrained weights and others remain unknown. It's shown that our method yields competitive results and even surpasses state-of-the-arts under some metrics.

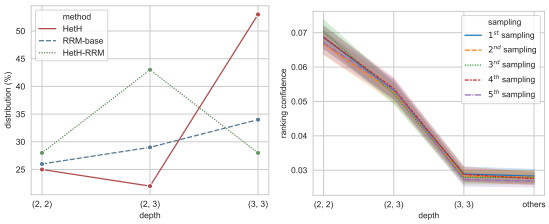
When it comes to key relation prediction, we directly evaluate HetH, MOTIFS, and VCTree-SL on **VG-KR**. As shown in Table 2, HetH substantially performs better than other two competitors, suggesting that the structure of HET provides hints for judging the importance of relations, and parsing the structured information in HET indeed capture humans' perceptive habits.

In pursuit of ultimate performances on mining key relations, we jointly optimize the HetH with RRM under the supervision of key relation annotations in VG-KR. From Table 2, both HetH-RRM and MOTIFS-RRM achieve significant gains, and HetH-RRM is better than MOTIFS-RRM, which proves the superiority of HetH again, and shows excellent transferability of RRM.

Qualitative Analysis. We visualize intermediate results in Figure 3(a-d). HET is well constructed and close to human's analyzing process. In the area map, regions of *arm*, *hand*, and *foot* get small weights because of their small sizes. Actually, relations like $\langle$*lady*, *has*, *arm*$\rangle$ are indeed trivial. As a result, RRM suppresses these relations. More cases are provided in supplementary materials.

4.4 Analyses about HET

We conduct additional experiments to validate whether HET has a potential to reveal humans' perceptive habits. As shown in Figure 4(a), we compare the **depth distribution** of top-5 predicted relations (represented by tuple (d_{o_i}, d_{o_j}) consisting of the depths of two entities, and the depth of root is defined as 1.) of HetH, RRM-base and HetH-RRM. After applying RRM, there is a significant increment on the ratio of depth tuples (2, 2) and (2, 3), and a drop on (3, 3). This phenomenon is also observed in Figure 3(e). Previous experiments have proved that RRM obviously upgrades the rankings of key relations. In other words, relations which are closer to the root of HET are regarded as key ones by RRM.



(a) (b)

Fig. 4: (a) Depth distribution of top-5 predicted relations. (b) The ranking confidence of relations from different depths obtained from RRM-base. Sampling is repeated five times.

Strtg.	Metric	HetH-RRM
EP	kR@1	**17.5**
	kR@5	**35.0**
	speed	0.22
SP	kR@1	15.8
	kR@5	31.2
	speed	**0.18**

Fig. 5: Comparison between EP and SP. The inference speed (seconds/image) is evaluated with a single TITAN Xp GPU).

We also analyze the ranking confidence (ϕ) of relations from different depths with the RRM-base model (to eliminate the confounding effect caused by AAP information). We sample 10,000 predicted relation triplets from each depth five times. In Figure 4(b), the ranking confidence decreases as the depth increases. Therefore, different levels of HET indeed indicate different perceptive importance of relations. This characteristic makes it possible to reasonably adjust the scale of a scene graph. If we want to limit the scale of a scene graph but keep its ability to sketch image gist as far as possible, it is feasible for our hierarchical scene graph since we just need to cut off some secondary branches of HET, but is difficult to realize in an ordinary scene graph. We give an example in the supplementary materials.

Besides, different from traditional **Exhausted Prediction** (**EP**, predict relation for every pair of entities) during inference stage, we adopt a novel **Structured Prediction** (**SP**) strategy, in which we only predict relations between parent and children nodes, and any two sibling nodes that share the same parent. In Figure 5, we compare the performances and inference speed between EP and SP for HetH-RRM. Despite the slight gap in terms of performances, the interpretability of connections in HET makes SP feasible to take a further step towards efficient inference, getting rid of the $O(N^2)$ complexity [16] of EP. Further researches need to be conducted to balance performance and efficiency.

5 Experiments on Image Captioning

Do key relations really make sense? We conduct experiments on one of the downstream tasks of SGG, i.e., image captioning, to verify it. [2]

Experiments are conducted on VG-KR since it has caption annotations from MSCOCO. To generate captions, we select different numbers of predicted top

[2] We briefly introduce here and details are provided in supplementary materials.

Table 4: Results of image captioning on VG-KR.

Num.	Model	B@1	B@2	B@3	B@4	ROUGE-L	CIDEr	SPICE	Avg. Growth
all	GCN-LSTM	72.0	54.7	40.5	30.0	52.9	91.1	18.1	
20	HetH-Freq	73.1	55.7	41.0	30.1	53.5	94.0	18.8	0.06
	HetH	74.9	**58.4**	**43.9**	**32.8**	54.9	101.7	19.8	
	HetH-RRM	**75.0**	58.2	43.7	32.7	**55.1**	**102.2**	**19.9**	
5	HetH-Freq	70.7	53.2	38.6	28.0	51.7	84.4	17.2	1.57
	HetH	72.5	55.4	41.2	30.5	53.1	92.6	18.5	
	HetH-RRM	**73.7**	**56.7**	**42.3**	**31.5**	**54.0**	**97.5**	**19.1**	
2	HetH-Freq	68.1	50.8	36.8	26.5	50.2	76.5	15.5	2.10
	HetH	70.8	53.4	39.2	28.7	51.8	86.4	17.6	
	HetH-RRM	**72.3**	**55.2**	**41.0**	**30.4**	**53.1**	**92.2**	**18.4**	

relations and feed them into the LSTM backend following [44]. We reimplement the complete **GCN-LSTM** [44] model and evaluate it on VG-KR since it's one of the state-of-the-art methods and is most related to us. As shown in Table 4, our simple frequency baseline, **HetH-Freq** (the rankings of relations are accord with their frequency in training data), with 20 top relations input, outperforms GCN-LSTM because GCN-LSTM conducts graph convolution using relations as edges, which is not as effective as our method in terms of making full use of relation information. After applying RRM, there is consistent performance improvement on overall metrics. This improvement is more and more significant as the number of input top relations reduces. It's reasonable since the impact of RRM centers at top relations. It suggests that our model provides more essential content with as few relations as possible, which contributes to efficiency improvement. The captions presented in Figure 3(e) shows that key relations are more helpful for generating a description that highly fits the major events in an image.

6 Conclusion

We propose a new scene graph modeling formulation and make an attempt to push the study of SGG towards the target of practicability and rationalization. We generate a human-mimetic hierarchical scene graph inspired by humans' scene parsing procedure, and further prioritize the key relations as far as possible. Based on HET, a hierarchcal scene graph is generated with the assistance of our Hybrid-LSTM. Moreover, RRM is devised to recall more key relations. Experiments show outstanding performances of our method on traditional scene graph generation and key relation prediction tasks. Besides, experiments on image captioning prove that key relations are not just for appreciating, but indeed play a crucial role in higher-level downstream tasks.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contracts Nos. 61922080, U19B2036, 61772500, CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009, and Beijing Academy of Artificial Intelligence No. BAAI2020ZJ0201.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2425–2433 (2015) 1
2. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6163–6171 (2019) 3, 10, 11
3. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3298–3308 (2017) 3
4. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005) 6
5. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1969–1978 (2019) 3
6. Han, F., Zhu, S.C.: Bottom-up/top-down image parsing with attribute grammar. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **31**(1), 59–73 (2008) 3
7. He, S., Tavakoli, H.R., Borji, A., Pugeault, N.: Human attention in image captioning: Dataset and analysis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 8529–8538 (2019) 4
8. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3203–3212 (2017) 4, 7
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (11), 1254–1259 (1998) 4
10. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 3668–3678 (2015) 1, 3
11. Kim, D.J., Choi, J., Oh, T.H., Kweon, I.S.: Dense relational captioning: Triple-stream networks for relationship-based captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6271–6280 (2019) 8
12. Klein, D.A., Frintrop, S.: Center-surround divergence of feature statistics for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2214–2219 (2011) 4
13. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) **123**(1), 32–73 (2017) 3
14. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5455–5463 (2015) 4
15. Li, Y., Ouyang, W., Wang, X., Tang, X.: Vip-cnn: Visual phrase guided convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7244–7253 (2017) 3, 11

16. Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., Wang, X.: Factorizable net: an efficient subgraph-based framework for scene graph generation. In: Proceedings of European Conference on Computer Vision (ECCV). vol. 11205, pp. 346–363. Springer (2018) 1, 13
17. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1261–1270 (2017) 1, 3
18. Liang, X., Lee, L., Xing, E.P.: Deep variation-structured reinforcement learning for visual relationship and attribute detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4408–4417 (2017) 11
19. Liang, Y., Bai, Y., Zhang, W., Qian, X., Zhu, L., Mei, T.: Vrr-vg: Refocusing visually-relevant relationships. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 10403–10412 (2019) 3
20. Lin, L., Wang, G., Zhang, R., Zhang, R., Liang, X., Zuo, W.: Deep structured scene parsing by learning with image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2276–2284 (2016) 3
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of European Conference on Computer Vision (ECCV). vol. 8693, pp. 740–755. Springer (2014) 3, 9
22. Lin, X., Ding, C., Zeng, J., Tao, D.: Gps-net: Graph property sensing network for scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3746–3755 (2020) 3
23. Liu, N., Han, J., Yang, M.H.: Picanet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3089–3098 (2018) 4
24. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Proceedings of European Conference on Computer Vision (ECCV). vol. 9905, pp. 852–869. Springer (2016) 3, 8, 10, 11
25. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7219–7228 (2018) 1
26. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543 (2014) 6
27. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Weakly-supervised learning of visual relations. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5179–5188 (2017) 3
28. Qi, M., Li, W., Yang, Z., Wang, Y., Luo, J.: Attentive relational networks for mapping images to scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3957–3966 (2019) 1, 3
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 91–99 (2015) 4
30. Sharma, A., Tuzel, O., Jacobs, D.W.: Deep hierarchical parsing for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 530–538 (2015) 3
31. Shi, J., Zhang, H., Li, J.: Explainable and explicit visual reasoning over scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8376–8384 (2019) 1

32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 4
33. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 129–136 (2011) 3
34. Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1556–1566 (2015) 6
35. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3716–3725 (2020) 3
36. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6619–6628 (2019) 1, 3, 5, 10, 11
37. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3183–3192 (2015) 4
38. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4019–4028 (2017) 4
39. Wang, W., Wang, R., Shan, S., Chen, X.: Exploring context and visual pattern of relationship for scene graph generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8188–8197 (2019) 1, 3, 10, 11
40. Wu, Q., Shen, C., Wang, P., Dick, A., van den Hengel, A.: Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(6), 1367–1381 (2018) 1
41. Xie, Y., Lu, H., Yang, M.H.: Bayesian saliency via low and mid level cues. IEEE Transactions on Image Processing (TIP) **22**(5), 1689–1698 (2012) 4
42. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5419 (2017) 1, 3, 9, 10, 11
43. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of European Conference on Computer Vision (ECCV). vol. 11205, pp. 690–706. Springer (2018) 1, 3, 10, 11
44. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of European Conference on Computer Vision (ECCV). vol. 11218, pp. 711–727. Springer (2018) 1, 14
45. Yao, T., Pan, Y., Li, Y., Mei, T.: Hierarchy parsing for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2621–2629 (2019) 3
46. Yin, G., Sheng, L., Liu, B., Yu, N., Wang, X., Shao, J., Loy, C.C.: Zoom-net: Mining deep feature interactions for visual relationship recognition. In: Proceedings of European Conference on Computer Vision (ECCV). vol. 11207, pp. 330–347. Springer (2018) 3, 11, 12

47. Yu, R., Li, A., Morariu, V.I., Davis, L.S.: Visual relationship detection with internal and external linguistic knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1974–1982 (2017) 3, 9, 11
48. Zareian, A., Karaman, S., Chang, S.F.: Weakly supervised visual semantic parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3736–3745 (2020) 3
49. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5831–5840 (2018) 1, 2, 3, 5, 10, 11
50. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5532–5540 (2017) 1, 3
51. Zhang, H., Kyaw, Z., Yu, J., Chang, S.F.: Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4233–4241 (2017) 3
52. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 33, pp. 9185–9194 (2019) 3
53. Zhang, J., Shih, K.J., Elgammal, A., Tao, A., Catanzaro, B.: Graphical contrastive losses for scene graph parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11535–11543 (2019) 1, 3, 11, 12
54. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6024–6033 (2019) 4
55. Zhu, L., Chen, Y., Lin, Y., Lin, C., Yuille, A.: Recursive segmentation and recognition templates for image parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **34**(2), 359–371 (2011) 3