A Survey of Scene Graph: Generation and Application

Preprint · April 2020

DOI: 10.13140/RG.2.2.11161.57446

6 authors, including:

Pengfei xu
Xi'an Jiaotong University

Xiaojun Chang
Monash University (Australia)

Po-Yao Huang
Facebook AI Research

Alexander G. Hauptmann
Carnegie Mellon University

A Survey of Scene Graph: Generation and Application

Pengfei Xu, Xiaojun Chang, Ling Guo, Poyao Huang, Xiaojiang Chen, and Alexander G. Hauptmann

***Abstract*—Scene Graph is a data structure, which is mainly used to describe the objects, attributes and object relationships in a scene. Scene Graph is a deep representation of a scene, and is very conducive to many visual tasks, such as image retrieval, image/video captions, VQA, and even to image generation and specific relationship detection. At present, numbers of research works about scene graph are proposed, including the scene graph generation methods and the related applications. These proposed methods based on scene graph have great improvements in relative performances compared with the corresponding traditional methods, which also proves the effectiveness of scene graph in the visual understanding of a scene. Therefore, In this paper, we provide a systematic review of the existing techniques of scene graph generation and application, including not only the state-of-the arts but also those with latest trends. Particularly, we discuss the scene graph generation methods according to the inference models for visual relationship detection, and the applications of scene graph are stated according to the specific visual tasks. Finally, we point out several problems in the current scene graph generation methods, related applications and the future research directions of scene graph.**

***Index Terms*—Scene Graph, Object Detection, Visual feature extraction, Prior Information, Visual Relationship Recognition.**

I. Introduction

A Scene graph is first proposed as a data structure that describes the object instances in a scene and the relationships between the objects [1]. As shown in Fig.1, a complete scene graph can represent the detailed semantics of a dataset of scenes, but not a singe image or a video; and it has powerful representations that encode 2D/3D images [1], [2] and videos [3], [4] into their abstract semantic elements without any restriction on the types and attributes of objects and the relationships between objects. Fig. 1 (a) illustrates a scene graph, and we can see that a scene graph G is a data structure of directed graph, which can be defined as a tuple $G = (O, E)$, where $O = O_1, ..., O_n$ is a set of objects detected in the images. Each object has the form $o_i = (c_i, A_i)$, where c_i and A_i are the category and attributes of the object respectively. While $E \subseteq O \times R \times O$ is a set of directed edges to repress the relationships between objects. At present, a scene graph is commonly associated to an image dataset, but not to only one image; So it can be consider as a visual understanding to relevant images. While, a part of scene graph is grounded to an image by associating the objects to the corresponding regions in an image, as shown in Fig. 1 (b). Scene graph has a powerful representations for semantic features about the scene, and is beneficial for a wide range of visual tasks.

There are some similarities of scene graphs with the commonsense knowledge graph, such as their graphical structures and constituent elements. However, scene graph is a different type of knowledge graph, which is mainly reflected in the following aspects: (a) Each node in scene graph is associated with an image region, and these nodes come in pairs, namely a subject and an object; while each node in knowledge graph is the general concept of its semantic label. (b) In a scene graph, the directed edges represent the relationships between pairs of objects; while each edge in knowledge graph encodes a relational fact involving a pair of concepts [5].

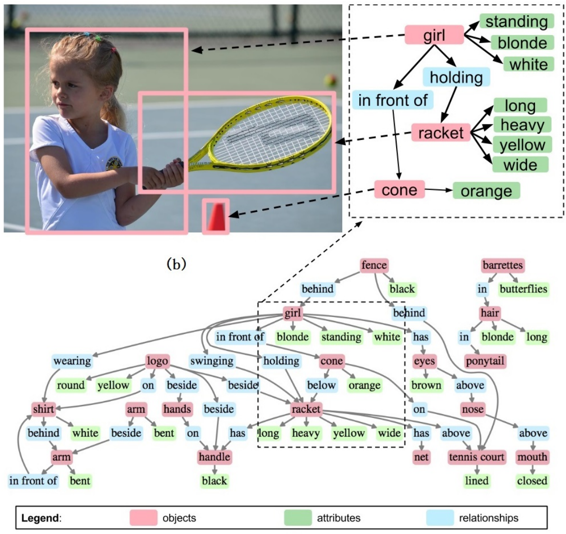
The idea of using the visual features of different objects in the image and the relationships between them have been proposed for achieving the visual tasks of action recognition [6], image captioning [7] and other relevant tasks [8] as early as 2015. Then, Johnson et al. proposed the concpt f scene graph [1], and gave the correspondin g notation representations. In [1], scene graph is generated manually from a dataset of real-world scene graphs, so as to capture the detailed semantics of a scene. Since then, the research on scene graph has received extensive attentions. Subsequently, several scene graph datasets are introduced [9], [10], [11], [12]. Based on these datasets, many scene graph generation (SGG) methods are proposed, and these methods can be divided SGG methods with facts alone as well as introducing prior information. At present, these SGG methods pay more attention to the methods with fact alone, including CRF-based (conditional random field) SGG [1], [13], [14], VTransE-based (visual translation embedding) SGG [15], [16], [17], Faster RCNN-based SGG [18], [19], [20], RNN/LSTM-based SGG [21], [22], [23], GNN [24], [25], [26], and other SGG methods with fact alone [27], [28], [29]. In addition, different types of prior information are introduced for SGG, such as Language Priors [9],visual contextual information [30], [22], Knowledge priors [31], [32], visual cue [33], and so on. Scene graph has the powerful representations for the semantic features of a scene, thus, it has widely applied to related visual tasks, such as image retrieval [1], [34], image generation [35], [36], specific relationship recognition [37], [38], [39], image/video captioning [40], [41], [42], VQA [43], [44], and so on. Therefore, we can see that scene graph has become a hot research topic in computer vision, and it will still receive continuous attention in the future.

Since the concept of Scene graph was proposed in 2015 and first applied to image retrieval, then the relevant researches on

P. Xu, L. Guo and X. Chen are with the School of Information Science & Technology, Northwest University.

X. Chang is with the Faculty of Information Technology, Monash University. Email: cxj273@gmail.com.

P. Huang and A. Hauptmann is with School of Computer Science, Carnegie Mellon University. Email: alex@cs.cmu.edu.



(b)

(a)

Fig. 1: (a)): an example of a scene graph encodes objects, attributes, and relationships in a scene. (b)): a grounding of the scene graph associates the object to the regions in an image.

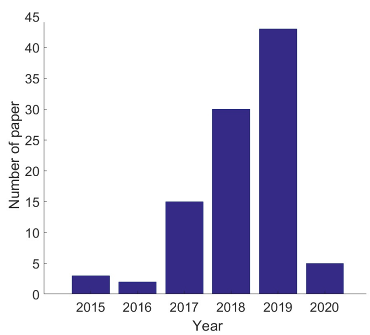
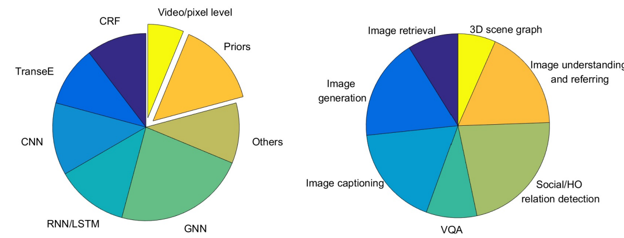


Fig. 2: Classification and statistics of the researches on scene graph from 2015 to 2020.



(a) SGG methods (b) Apllications

Fig. 3: The Classification and statistics of SGG methods and Apllications

Scene graph have increased significantly, especially in 2019 (As shown in figure 2). In these research results, we mainly focus on the scene graph generation (SGG) methods and the applications of scene graph. Fig.3 (a) shows the relevant works on SGG, and it can be seen that more researches are focused on SGG by using GNN models and introducing relevant prior information. While the applications of Scene graph mainly refer to image generation, image/video captioning and image semantic understanding and reasoning, etc. as shown in Fig.3 (b). There also are a few applications on VQA and image retrieval. In addition, several works utilized 3D scene graph for 3D object detection and recognition. With the increasing researches on scene graph, the scene graph databases related to specific tasks are constantly updated and established, which enable reliable data for the further researches on scene graph in the future.

At present, the researches on scene graph mainly try to solve the following three problems:(1) How to generate a more accurate and complete scene graph ;(2) How to simplify the computational complexity of SGG;(3) How to apply scene graph to more tasks in a more appropriate and extensive way. Although there have been many related methods proposed for solving these problems, there still need deep researches on the solution of these problems. Moreover, there are still other problems that need to be further solved. For example, the unbiased scene graph data has always been a problem in scene graph generation, and will be a problem to be solved in the later research. In addition, the descriptions of the relationships between objects in datasets are rough and inaccurate. Therefore, we need to further optimize the annotations in related scene graph datasets.

In this paper, we mainly discusses the generation and application of scene graph relevant to computer vision in this paper. In section 2, we first introduce several existing datasets that are commonly used for scene graph, as well as the performance evaluation of scene graph generation models. Section 3 briefly introduces basic notations of scene graph, and then provide a thorough review of current available scene graph generation techniques, including those work with facts alone, as well as using different types of prior information. Meanwhile, We describe the overall frameworks of models, model training, as well as pros and cons of such techniques. In section 4, we further explores the applications of scene graph to a wide variety of computer vision tasks. Furthermore, Section 4 f Section 5 will discuss the main problems in the generation and application of scene graph at present and the future researches of scene graph. Finally, we present our concluding remarks in Section 6.

II. Datasets for Scene Graphs

A long-standing goal of computer vision is to develop models that can understand the visual information in scenes, and further reason some unseen visual events from the current scenes. While in terms of current AI technologies, the performance of the relevant network models is still largely dependent on the knowledge learned from the existing datasets. If these models are transferred from their original datasets to other datasets with relatively unfamiliar scenes, the performance of the models is likely to decline dramatically or even fail to work. Therefore, large scale visual datasets for specific tasks are critical to the computer vision network models. In this section, We discuss several existing datasets that have been released for scene graph generation and applications of relevant downstream tasks. We briefly state the basic data structure of these main scene graph datasets, and make a further comparative analysis on these data sets.

Real-World Scene Graphs Dataset. In 2015, Johnson proposed the notion of scene graph, as well as Real-World Scene Graphs Dataset (RW-SGD) [1], which may be the first dataset explicitly created for scene graph generation and application (image retrieval). RW-SGD is built by manually selecting 5,000 images from YFCC100m [45] and Microsoft COCO datasets [46], and then Amazon's Mechanical Turk (AMT) is used to produce a human-generated scene graph from these selected images. The Final RW-SGD contains over 93,832 object instances, 110,021 attribute instances, and 112,707 relationship instances.

Visual Relationship Dataset (VRD) [9] is constructed for the task of visual relationship prediction. VRD has 100 object classes detected from 5000 images, and also contains 37,993 relationships. However, the distribution of the visual relationships has the common problem of the long tail of infrequent relationships in scene graph datasets.

Visual Genome Dataset (VGD) [10] is a large scale visual dataset, and consists the components of objects, attributes, relationships, question answer pairs, and so on. At present, VGD has widely applied to scene graph generation and application for its large number of images, objects, relationships, and so on. In addition, another scene graph dataset (Visually-Relevant Relationships Dataset (VrR-VG)) [19] is constructed based on VGD.

UnRel Dataset (UnRel-D) [11] is a new challenging dataset of unusual relations, and contains more than 1000 images, which can be queried with 76 triplet queries.

HCVRD Dataset [12] has 52,855 images with 1,824 object categories and 927 predicates, and also contains 28,323 relationships types. Similar to VRD, HCVRD also has the long-tail distribution of infrequent relationships.

III. Scene Graphs Generation

The concept of scene graph is first proposed by Johnson in [1], and manually established the corresponding scene graph on a real-time World scene Graph dataset. A scene graph is a topological representation of a scene, which mainly encodes object and their relationships. The task of scene graph generation (SGG) is to construct a graph structure that best associates its nodes and edges with the objects and their relationships in a scene. While the key challenge task is to detect/recognize the relationships of the objects.

Currently, there are two main scene graph generation approaches [25]. The first approach has the two stages, that is object detection and pair-wise relationship recognition [13], [49], [9], [50]. The other approach is to jointly detect and recognize the objects and their relationships [20], [24], [48]. The subsequent SGG methods are proposed to generate a complete scene graph with facts alone or by introducing additional prior information. In this section, we will review the SGG methods using only facts observed in the given images/videos; and further discusses the techniques that incorporate other priors.

A. Scene Graphs Generation with Facts Alone

1) CRF-based SGG: Johnson et al. proposed the concept of scene graph, and give the corresponding formulations. While they used Amazon's Mechanical Turk (AMT) to constructe a scene graph manually on RW-SGD [1]. Furthermore, conditional random field (CRF) is construct for image retrieval using the generated scene graph.However, it takes much cost for generating a scene graph manually, and it has the influence of subjective factors of understanding a scene. subsequently, Schuster et al. [8] proposed a method of scene graph generation automatically using two parsers: a rule-based parser and a classifier-based parser, which map dependency syntax representations to scene graphs. Based on the constructed scene graph, they also achieved the image retrieval task via CRF. These may be the two early methods that involved the construction and applications of scene graph.

Formally, for a given scene graph $G = (O, E)$, there are many possible ways of grounding the scene graph to an image I. The inference tasks and other visual tasks at the high level are to recognize objects, predict the objects' coordinates, and detect/recognize pairwise relationship predicates between objects [51]. Therefore, the first stage of identifying the categories and attributes of the detected objects is achieved mainly using RPN or Faster RCNN [52]. Furthermore, most of the works focus on the key challenge of reasoning the visual relationship. In [13], the CRF model of secne graph also has two unary potentials that associate the objects with their visual features and relationshipss. While, for relational modeling, Deep Relational Network (DR-Net) are explored to detect the relationships.

In [14], SG-CRF is proposed to improve the accuracy of SGG. In this methods, Semantic Compatibility Network (SCN) are used to learn the semantic compatibility of nodes in the scene graph, and approximates scene graph inference by mean-field approximation algorithm, which can be expressed as $Q^t = MeanField(\psi_u, L_e, Q^{t-1})$. Then the pairwise potential ψ_p of each node is calculated according to the label word embeddings of its 1-hop neighbors. Q^T is the final output of last mean-field iteration. Assume that I is the given input image, and SG denotes the final generated scene graph. Then the objective for SG-CRF can be formulated as maximizing the following probability function [14]:

$$P(SG|I) = \prod_{o_i \in O} P(o_i, o_i^{bbox}|I) \prod_{r_{i\to j} \in E} P(r_{i\to j}|I) \qquad (1)$$

Where, the term $P(o_i, o_i^{bbox}|I)$ is a unary potential, which models the agreements of the visual features of the box o_i^{bbox} with the category and attributes of the object o_i. Then, CRFs for SGG can be formulated to find the optimal solution of x^*= arg $max_x P(X)$, which obeys the Gibbs distribution. Where,

$$P(X) = \frac{1}{Z(X)} exp(-\sum_i \psi_u(x_i) - \sum_{j\neq i} \psi_p(x_i, x_j)) \qquad (2)$$

Similar to Eq.1, the unary potential $\psi_u(x_i)$ is the measurement for assigning the node x_i, and the pairwise potential $p(x_i; x_j)$ is the cost of assigning x_i to x_j.

Dataset	images/videos	Obj. instances	Obj. classes	Att. instances	Att. types	Rel. Instances	Rel. types	Pre. per Obj. Category	Pre.
COCO [46]	124,828	886,284	80	-	-	-	-	-	-
YFCC100m [45]	-845735	534,309	200	-	-	-	-	-	-
RW-SGD[1]	5000	93,832	6745	110,021	3743	112,707	1310	3.3	-
VRD [9]	5000	-	100	-	-	37,993	6,672	24.25	-
VGD [10]	100k	33,877	3,843,636	-	-	-	40,480	-	
UnRel [11]	1000	-	-	-	-	76	-	-	-
HCVRD [12]	52,855	-	1824	-	-	256,550	28,323	10.63	927
VrR-VG [19]	58,983	282,460	1600	-	-	203,375	117	-	
Visual Phrase [47]	2,769	3,271	8			2040	13	120	-
VG150[48]	87,670	738,945	150	-	-	413,269	50	-	-

TABLE I: Aggregate statistics for scene graph datasets.

2) TransE-based SGG: There are similarities between scene graph and knowledge graph in terms of object relationship reasoning. Therefore, inspired by the advances of Translation Embedding networks (TransE) in relational representation learning of knowledge bases and object detection networks, relevant models and methods based on TransE are explored for visual relationship detection/recognition to build a scene graph. [15], [16], [17], [51]. These TransE-based SGG methods place the visual relationships between objects in a low-dimensional relation space, where the relations are modeled as simple translation vectors.

VTransE [15] extends TransE networks [53] for modeling visual relations and the predicates. In VTransE, the detected subjects and objects are mapped into a low-dimensional relation space, and their relationships are formulated as translation vectors for scene graph generation. Similar to other SGG methods, object detection needs to carry out first, and VTransE networks can be married to any object detection networks such as Faster-RCNN [52], SSD [54] and YOLO [55], which are used to locate the objects and recognize their categories for the following task of relation recognition.

VTransE represents any valid relation $\langle subject, predicate, object \rangle$ in vectors s, p and o respectively. If the relation holds,then this relation can be represented as a translation: $s + p \approx o$ in the embedding space, otherwise $s + p! \approx o$. Besides VTransE learns the relation translation vector $t_p \in R^r$ as in TransE, and it learns two projection matrices W_s, W_o by $s = W_s x_s$ and $o = W_o x_o$:

$$W_s x_s + t_p \approx W_o x_o \tag{3}$$

Where x_s, x_o are the visual features of subjects and objects, respectively. Furthermore, a prediction loss is proposed to solve the problem of problematic sampling negative triplets due to the incomplete relation annotation:

$$_{rel} = \sum_{(s,p,o) \in R} -logsoftmax(t_p^T (W_o x_o) - W_s x_s)) \tag{4}$$

Finally, the relation detection scores are obtained by summing the scores of subject/object detection and relation predicate prediction.

UVTransE [16] is proposed to improve generalization for the rare or unseen relations based on VTransE. There are lots of obvious object relations in scenes, but also exist many unseen relations. Therefore, the relation detection models also need to recognize the hidden relations. Inspired by VTransE [15], UVtransE introduces the union of subject and object, and a context-augmented translation embedding model is proposed to capture both common and rare relations in scenes. Similar to [15], UVtransE needs to learn three projection matrices W_s, W_o and W_u by minimizing the following multi-class cross-entropy loss function:

$$\begin{aligned} L_{vis} = \sum_{(s,p,o) \in T} -log \frac{exp(p^\top \widehat{p})}{\sum_{q \in P} exp(q^\top \widehat{p})} + C([\|W_s s\|_2^2 - 1]_+ \\ + [\|W_s s\|_2^2 - 1]_+ + [\|W_s s\|_2^2 - 1]_+) \end{aligned} \tag{5}$$

Where, T and P are the set of all relationship triplets and the set of all predicate labels. $\widehat{p} = W_u u - W_s s - W_o o$, $[x]_+ = max(0, x)$. C is a hyper-parameter, which is used to determine the importance of the soft constraints. Eq.(5) is different from VTransE [15] in terms of the introduced contextual union feature. Finally, similarly to [15], the score of the entire triplet is also to use the sum of the subject/object detection score and the predicate recognition score.

MATransE (Multimodal Attentional Translation Embeddings) [17] is proposed to satisfy $s + p = o$ by guiding the features' projection with attention and Deep Supervision. Similar to [15], MATransE needs to learn the projection matrices W_s, W_p, and W_o by employing a Spatio-Linguistic Attention module (SLA-M) [13]. As shown in Eq.(3), a two-branch architecture is designed in MATransE: one branch is to drive the predicate features into scores $t_p = W_{p)x_P}$ (P-branch), and another branch is used to classify the object-subject features $W_o x_o - W_s x_s$ (OS-branch).

Finally, both the scores of P-branch and OS-branches are connected into a single vector, which would be used to train a meta-classifier to predict the categories of the predicates. Thus, with $W = (W_s, W_p, W_o)$, the total loss can be formulated as:

$$L(W) = \lambda_f L_f(W) + \lambda_p L_p(W_p) + \lambda_{os} L_{os}(W_o, W_s) \tag{6}$$

Where λ is used to balance the importance of each term.

RLSV [56] is proposed to solve the problem of the incomplete scene graph, and the formulation of RLSV is to predicte the missing relations between the objects. RLSV is staged by three modules: visual feature extraction, hierarchical projection and train objective module. By combining location and visual information of entities, the visual feature extraction model embeds the inputting image as visual projection vectors v_{p_h}, v_{p_r}, v_{p_t} for head h, relation r and tail t respectively. Based on v_{p_h}, v_{p_r}, v_{p_t}, the hierarchical projection module projects a given visual triple (h, r, t) onto attribute space, relation space and visual space, resulting in a new presentation $(h_\perp, r_\perp, t_\perp)$. Then followed by TransE, the score function can be defined as:

$$E_I(h, r, t) = ||h_\perp + r_\perp - t_\perp||_{L_1/L_2} \tag{7}$$

Finally, a max-margin function with negative sampling is formulated as the training objective:

$$L = \sum_{I \in \mathcal{I}} \sum_{(h,r,t) \in \mathcal{T}_I} \sum_{(h',r',t') \in \mathcal{T}_I'} [E_I(h, r, t) - E_I(h', r', t') + \gamma]_+ \tag{8}$$

where γ is a marginal hyperparameter, $\mathcal{T}_I'$ is the negative sampled visual triple set generated from positive visual triple set $\mathcal{T}_I$.

3) CNN-based SGG: **DR-Net** [13] is a framework which formulates a triplet in the form of $\langle subject, predicate, object \rangle$ as its prediction output, and jointly predicts the category label of the triplet by exploiting the spatial configuration and statistical dependency among the elements of the triplet. The whole framework has three stages: object detection, filtering pairs of objects and joint recognition. In the object detection stage, Fast RCNN is used to detect a set of candidate objects in the images, and each detected object comes with a bounding box and its visual features. The next stage is to filter out some pairs of detected objects by a low-cost neural network. Then the retained pairs of objects are fed to the joint recognition module. Finally, The joint recognition module would produce a triplet as the output by considering the visual features of each object, the spatial configurations between any paired objects and the statistical dependency between the relationship predicates. Specifically, to represent the spatial configurations, dual spatial masks are designed by deriving from the bounding boxes, which may overlap with each other. To exploit the statistical relations, DR-Net is developed to incorporate statistical relational modeling into a deep neural network framework.

SIN (Structure Inference Network) [57] is a detector which is designed to infer the object category labels by improving Faster R-CNN with a graphical model. The visual features of the objects, the relationships between objects in a single image and the scene contextual information are exploited in SIN for improving the performances of object detection. The framework of SIN is as follows. ROIs are derived from the input images, and then each ROI is pooled into a feature map with fixed-size f_i^v, which is considered as a node in graph modeling. Meanwhile, a scene of an image is generated from its global feature f^s in the same way. The scenes and nodes are put into the SIN as Scene GRUs. Afterwards, both the spatial and visual features of the nodes v_i and v_j are jointly combined to form a directed edge $e_{j \to i}$ from v_j to v_i, which represents the influence of v_j on v_i. All edges will be passed into SIN as Edge GRUs. In SIN, the state of each GRU is updated at each iterative, and the final integrated node representations are used to predict the object category and bounding box offsets.

Rel-PN [18]. Relationship Proposal Networks (Rel-PN) first detect all meaningful proposals of object, subject and relationship by running 3-branch RPN in Faster RCNN [52] respectively. Although the object instances and subject instances belong to the same category space, their distribution is inconsistent, so they are extracted separately.

The relationship branch is to reduce the number of the pairs of objects, otherwise there would have $object \times subject$ pairs of relations. In [18], 9 kinds of relationship proposals are selected according to several conditions. Then two branches of visual compatibility and spatial compatibility modules are used to output the visual and spatial scores. For visual compatibility module, three visual features are connected to obtain a (5x5x512) vector, and then the module output visual score s_v. Moreover, three groups of spatial difference features are connected to obtain a (64x64) vector, and then output spatial score s_s. Finally, p_v and p_s are integrated into a final score.

$$p = \alpha p_v + (1 - \alpha) p_s \tag{9}$$

where, α is the ratio of visual compatibility.

Based on the model in [18], the model in [58] considers three types of features: visual, spatial and semantic features using three corresponding models, and these features are then fused for the final relationship identification. Different from [18], the model in [58] used an additional semantic module to learn the semantic features, and achieved better performances.

ViP-CNN [20] has the capacity to jointly learn the specific visual features for the interaction and to consider the visual dependency. ViP-CNN has four branches for triplet proposal and phrase recognition. Likely to other SGG models, Faster R-CNN with VGG-Net [59] as backbone is used to detect the objects and locate the corresponding bounding boxes, so as to provide the triplet proposals. For the triplet proposal branches, the extracted CNN features by VGG-Net are used for proposing regions of interest (ROIs), and then triplet proposals are obtained by grouping these ROIs. Furthermore, triplet non-maximum suppression (triplet NMS) is proposed to solve the problem of the sparsity of relationship annotations, so as to reduce the redundancy information. While, the remained triplets are used for the branch of phrase recognition.

BAR-Net[60] uses the standard object detection methods to detect pair-wise relationships, which is achieved by decomposing the relation detection task into two tasks of retentive object detection. In BAR-Net, one detector (Such as faster RCNN) is used to detect all objects in the image, and then the other detector was used to detect the objects, which have interactions with each object. The bounding boxes obtained by the first detector are used as the inputs for the second detector, and the the joint probability can be represent by simpler conditional probabilities:

$$pro(s, p, o|I) = pro(s|I) pro(p, o|s, I) \tag{10}$$

The second probability iterm $pro(p, o|s, I)$ models the probability that an object o in the image is related to the subject S, which is called Box Attention.

LinkNet [51] is proposed to improve scene graph generation by explicitly modeling inter-dependency among all related objects, rather than an object in isolation. Linknet mainly has three modules:1).A relational embedding module is used to classify the objects and their relationships. Given an image, objects' proposals and labels are extracted by a object detection method, such as Faster R-CNN [52].2).A global context encoding module is used to extract global information, which contains as much as possible all proposal information in the image, and is used to assist the classification of object relations. 3).A geometrical layout encoding module is used to assist in the classification of object relations using the spatial information between the object proposals. Finally, the two categories can be used to generate the scene graph, and the loss function of whole network is the weighted sum of the losses for predicting the bounding boxes and categories of the detected objects, and even the relationship categories between the objects.

4) RNN/LSTM-based SGG: **Iterative Message Passing** [48]. As many previous works focued on building a scene graph given an image, surrounding context in it is ignored. However, scene graph prediction based on context information could resolve vagueness since local predictions is isolated. Motivated by this observation, Xu et.al. proposed an iterative message passing based model to fulfill scene graphs generation.

Given an image, their model first generates object proposals B_I by Region Proposal Network (RPN). With these object proposals, three tasks is to be fulfilled: object class label inferring, bounding box offsets computation and predicate prediction. Thus, the scene graph generation problem is formulated as optimizing

$$x^* = argmax_x Pr(x|B_I, I) \tag{11}$$

,where

$$Pr(x|B_I, I) = \prod_{i\in V}\prod_{j\neq i} Pr(x_i^{cls}, x_i^{bbox}, x_{i\to j}|B_I, I), \tag{12}$$

x_i^{cls} is the $i-th$ object proposal, x_i^{bbox} is the $i-th$ proposal's bounding box offsets, $x_{i\to j}$ is the predicate that $i-th$ object proposal applies to $j-th$ object proposal.

To cut down the cost of inference on a dense graph, a general RNN unit, namely GRU, is used to seek out the hidden states in the training framework. Based on GRU, $Pr(x|B_I, I)$ is transformed as follows.

$$\begin{aligned} Pr(x|B_I, I) = &\prod_{i\in V} Q(x_i^{cls}, x_i^{bbox}|h_i)Q(h_i|f_i^v) \\ &\prod_{j\neq i} Q(x_{i\to j}|h_{i\to j})Q(h_{i\to j}|f_{i\to j}^e) \end{aligned} \tag{13}$$

where f_i^v is the $i-th$ feature node and $f_{i\to j}^e$ is the edge feature from $i-th$ node to $j-th$ node. To prompt the infering effeciency, a prime dual message passing scheme between node GRU graph and edge GRU graph is used with the bipartition of a scene graph.

PANet[21]. Many previous works focus on context and scene information for relationship prediction, which ignors internal connection among predicates. Therefore, this paper proposed Predicate Association Network (PANet) to effectively aquiring contexts and relationship between predicates.

PANet is a two-stage network. In the first stage, Faster-RCNN is used to generate object proposals, of which each b_i represents three kinds of object features: category embedding (E_{b_i}i), spatial information (S_{b_i}) and visual feature (F_{b_i}).

$$V_{b_i} = \sigma(W_b(E_{b_i} \circ F_{b_i} \circ S_{b_i}) + b_b) \tag{14}$$

Based on these object proposals, instance-level context and scene-level context are obtained via an RNN. For each object pair $< s_i, o_i >$, their class probability $P(s_i|I)$ and $P(o_i|I)$ are aquired by these contexts.

In the second stage, the connections of predicates are explored via another RNN with matching technique and attention principle. Each predicate p_i is represented by a word embedding E_{p_i} . The combination of each pair of objects $< s, o >$'s instance-level and scene-level contexts is denoted as $< G_s, G_o >$. Feature maps of their union bounding box $F_{s,o}$ are used to represent the state of union region. $F_{s,o}$ is then fed into a fully connected layer to reduce dimension. The converged feature vector $U_{s,o}$ of the two objects is:

$$U_{s,o} = (G_s * G_o) \circ \sigma(W_u F_{s,o} + b_u) \tag{15}$$

where $G_s * G_o$ is used to find out each object pair's contexts.

Alignment feature is obtained by refining predicate label E_{p_i} with $P_{s,o}$. Then these features R_{p_i} are feed into an RNN module to extract predicate sets $\gamma_{pi}^{(2)}$ of the i-th predicate:

$$\gamma_{pi}^{(2)}, h_{pi}^{(2)} = RNN(R_{pi}, h_{pi-1}^{(2)}) \tag{16}$$

where, $h_{pi}^{(2)}$ is the hidden state in step i of the RNN, and $\gamma_{pi}^{(2)}$ is contextual information of predicate p_i. Then the final weighted contexts γ_{att} are computed for each pair $< s, o >$ as:

$$\gamma_{att} = \sum_{i=1}^{m} w_i \gamma_{p_i}^{(2)} s.t. 0 \leq w_i \leq 1 \tag{17}$$

The predicate label is determined as with the highest probability:

$$P(p_i|I, s_j, o_j) = max f(W_r \gamma_{att} + b_r) \tag{18}$$

where W_r and b_r are weights and bias respectively.

CMNs. [61]. Two issues exist in previous works on referential expressions, where one is that referential expressions were treated uniformly, thus failing to uncover the consistance between textual components and visual units in an image, the other is that relationship categories of most previous works are fixed.

To solver these two problem, this paper focus on referential expressions via inter-object relationships that can be represented as a triplet *(subject-relationship-object)* and proposed Compositional Modular Networks (CMNs), an end-to-end architecture that explicitly models the combinational linguistic structure of referential expressions and their groundings and supports interpretation of arbitrary language.

CMNs resolves the referential expression into a subject, relationship and an object with three soft attention maps discriminatively and arranges the extracted textual notations with image regions by a modular neural framework. There are two modules in CMNs, one for localizing specific textual components by generating scores for each component, and the other for setting the relationship between two bounding-box pairs by pairwise scores over region-region pairs. LSTM is used for expression parsing with attention in CMNs.

VCTREE [23]. Prior layout structures, like chain, fully connected graphs, are reliable for visual context encoding. Such prior layout structures are not perfect for the following two reasons. First, linear structures are too simple and might only obtain some spatial information or co-occurrence bias, and complete graphs are not discriminative for hierarchical relations since dense connections could result in saturated message passing during a subsequential context encoding. Second, object layouts should vary as contents or questions change. Therefore, fixed chains and complete graphs are inadequate for dynamic visual contexts.

In this paper, a model named VCTREE, which imposes dynamic "trees" on encoding object-level visual contexts for visual infering tasks, is proposed. VCTREE model has the following four steps. 1) Faster-RCNN is implemented for object proposal detection. The visual feature of proposal i is denoted as x_i, combining a RoIAlign feature $v_i \in R^{2048}$ with a spatial feature $b_i \in R^8$ with 8 variables, the bounding box coordinates (x_1, y_1, x_2, y_2), center $(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$ and size (x_2-x_1, y_2-y_1). Segmented features, like instance segmentations or panoptic segmentations, could also be options since the visual feature x_i is not just bounding box. 2) A matrix is learned to construct VCTREE. Since the VCTREE construction could be separate and the score matrix is non-differentiable from the loss of end-task, a hybrid learning strategy is developed. 3) Bidirectional Tree LSTM (Bi-TreeLSTM) is employed to encode the contextual information through VCTREE. 4) The encoded contexts will be decoded for each specific end-task.

AHRNN. [62] Most existing approaches to generate scene graphs suffer from two limitations that prevent them from generating a sound and effective scene graph. First, object-detection–based approaches will result in the generation of useless object bounding boxes or meaningless relationship pairs. Second, these methods rely on a ranking of probability for outputting relationships, which will result in semantically redundant relationships. Motivated by these two observations, the authors proposed an architecture that satisfied two demands: directly paying attention to and recognizing regions of interest in images without extra object detection; automatically ranking the sequence of relationships to output based on the learned features.

The overall architecture consists of a CNN model for extracting convolutional features, an AHRNN for generating a sequence of relationship pairs, and an algorithm for scene graph construction based on entity localization. Following the mainstream ”encoder-decoder” framework, a CNN model is employed to extract a set of feature vectors that represent a global visual description of an input image. Then, an Attention-based Hierarchical RNN (AHRNN) is responsible for dynamically mapping the feature vectors into the target relationship triplets. The AHRNN is composed of two models, an Attention-based Triplet RNN (ATRNN) to receive the image features and sequentially produce a topic vector by roughly attending to parts of the image features composing each relationship triplet, and an Attention-based Word RNN (AWRNN) to recognize each target word in the *(subject-predicate-object)* triplet under the guidance of the topic vector. Finally, with the predicted relationship triplets, entity localization is performed to determine the final components in the scene graph and an algorithm is designed for automatic scene graph construction.

MOTIFNET.[22] Elements of visual scenes have strong structural regularities. Based on this motivation, the authors examined some structural repetitions in scene graphs (called as motifs), using the Visual Genome dataset, which provides annotated scene graphs for $100k$ images from COCO, consisting of over $1M$ instances of objects and $600k$ relations. Their analysis leads to two key findings. First, there are strong regularities in the local graph structure such that the distribution of the relations is highly skewed once the corresponding object categories are given, but not vice versa. Second, structural patterns exist even in larger subgraphs and over half of images contain previously regularly appearing substructures in scene graphs (called as motifs). Based on the above analysis, a baseline is introduced: given object detections, predict the most frequent relation between object pairs with the given labels. The baseline improves over prior state-of-the-art by 1.4 mean recall points which suggests that an effective scene graph model must capture both the asymmetric dependence between objects and their relations, along with larger contextual patterns.

Thereafter, a neural network architecture called Stacked Motif Network (MOTIFNET) is proposed. The architecture breaks scene graph parsing into stages predicting bounding regions, labels for regions, and then relationships. Between each stage, global context is computed using bidirectional LSTMs and is then used for subsequent stages. In the first stage, a detector proposes bounding regions and then contextual information among bounding regions is computed and propagated (object context). The global context is used to predict labels for bounding boxes. Given bounding boxes and labels, the model constructs a new representation (edge context) that gives global context for edge predictions. Finally, edges are assigned labels by combining contextualized head, tail, and union bounding region information with an outer product. The method can be trained end-to-end. **MSDN** [24] explored the possibility in understanding the image through a single neural network model from three levels together, namely, object detection, scene graph generation and image caption. Since the features for these three tasks are highly correlated and can be the complementary information of each other, the authors proposed an end-to-end Multi-level Scene Description Network (MSDN) to simultaneously detect objects, recognize their relationships and predict captions at salient image regions, which effectively leveraged the rich annotations at three semantic levels and their connections for image understanding.

The entire process of MSDN is summarized as below: 1) Region proposal. To generate ROIs for objects, phases and, region captions. 2) Feature specialization. Given ROIs, to obtain specialized features that will be used for different semantic tasks. 3) Dynamic graph construction. Dynamically construct a graph to model the connections among feature nodes of different branches based on the semantic and spatial relationships of corresponding ROIs. 4) Feature refining. To jointly refine the features for different tasks by passing messages of different semantic levels along the graph. 5) Final prediction. Using the refined features to classify objects, predicates and generate captions. The scene graph is generated from detected objects and their recognized relationships.

The key procedure of MSDN is dynamic graph construction which realizes that regioin features, phrase features and object features are extracted from the orinal features seperately and delivered for image caption, phrase detection and object detection respectively after feature refining. Given the region features, a LSTM-based language model is used to generate natural sentences to describe the region.

5) GNN-based SGG: Currently, there are two polpular frameworks to generate scene graphs. One detects the objects first and then recognizes their pair-wise relationships, the other is to jointly infer the objects and their relationships based on the object region proposals. Both frameworks would generate a quadratic number of objects, which is time-consuming.

To improve the efficiency of scene graph generation, a subgraph-based connection graph is proposed to concisely represent the scene graph during the inference. A bottom-up clustering method is used to factorize the entire graph into subgraphs, where each subgraph contains several objects and a subset of their relationships. By replacing the numerous relationship representations of the scene graph with fewer subgraph and object features, the computation in the intermediate stage is significantly reduced.

Factorizable Net [25]. The bject region proposals are detected by RPN first, and then they are grouped into pairs to build up a fully-connected graph, where every two objects are connected with two directed edges. Thereafter, a more concise connection graph is generated by merging edges which refer to similar union regions into subgraphs. Based on the obtained obejects and subgraphs, the corresponding features (2-D feature maps for subgraph and feature vectors for objects) are generated. These features are refined through Spatial-weighted Message Passing (SMP) structure and the refined features would passed to Spatial-sensitive Relation Inference (SRI) module for predicate recognition. Here, SMP, a GNN approach, is used for better feature representation.

Predicate Prior Model[63]. Simlar to the idea that generating a scene graph by us ing language prior of relationship triples [9], Hwang et.al. generated it by joiontly combining the prior of predicate distribution [63]. The framework of [63] is similar to that of [48]. The framework first extracts visual features of nodes and edges from a set of object proposals. Then, mean field is used to perform approximate inference by using an iterative message passing scheme modeled with Gated Recurrent Units (GRU), which is to classify objects, predict their bounding box offsets, and classify relationship predicates between each pair of objects. The difference between [63] and [48] is that a pre-trained tensor-based relational module was added as a dense relational prior in [63] to refine the relationship estimation during the iterative messgae passing period, which is also a fine-tuning of the learning process of the scene graph module [48]. Here, a iterative message passing scheme with GRUs is uesed as a GNN way to improve the scene graph generation performance with better feature repesentation.

Graph R-CNN [64] is factorized into three logical stages: 1) object node extraction, 2) relationship edge pruning, and 3) graph context integration. In the object node extraction stage, a standard object detection pipeline is utilized to obtain a set of localized object regions. Then two novelties in the rest of the pipeline are used to incorporate the real-world regularities in object relationships. The first is a relation proposal network (RePN) that learns to efficiently compute relatedness scores between object pairs which are used to intelligently prune unlikely scene graph connections. Second, given the resulting sparsely connected scene graph candidate, an attentional graph convolution network (AGCN) is implemented to propagate higher-order context throughout the graph - updating each object and relationship representation based on its neighbors. These two mechanisms can effectively leverage object-relationship regularities to intelligently sparsify and reason over candidate scene graphs for scene graph generation.

PISP (Permutation-Invariant Structured Prediction model) [65]. A scene graph predictor should capture this dependence in order to improve prediction accuracy through uncovering the inter-dependency between objects and relations. Motivated by this observation, this paper denmenstrated that the architechture of a nueral network should stay invariant to a particular type of input permutation. Formally, a framework or a function $\mathcal{F}$ should produce the same result when given the same features, up to a permutation of the input. For example, consider a label space with three variables y_1, y_2, y_3, and assume that $\mathcal{F}$ takes as input $z = (z_1, z_2, z_3, z_{12}, z_{13}, z_{23}) = (f_1, f_2, f_3, f_{12}, f_{13}, f_{23})$, and outputs a label $y = (y_1^*, y_2^*, y_3^*)$. When $\mathcal{F}$ is given an input that is permuted in a consistent way, say, $z' = (f_2, f_1, f_3, f_{21}, f_{23}, f_{13})$, the output should still be $y = (y_1^*, y_2^*, y_3^*)$.

The authors proved this property according to the fact that such architechture or framework can aggregate information from the entire graph in a permutation invariant manner. Based this property, they suggested several common architectural structures like attention nueral networks and RNNs, which was used in their scene graph module.

Attention Graph model [66]. An Attention Graph mechanism is proposed to produce a scene graph structure that can be lifted directly from the top layer of a pre-trained Transformer model. This Transformer model with additional layers enables us to obtain graph node connectivity and class information directly.

The OpenAI Transformer Language Model was used as the foundation of the phrase parsing model, and the Language Model's final layer outputs were then fed in to a customised "Attention Graph" layer. The Attention Graph mechanism is trained using the sum of two cross-entropy loss terms against the respective target node types and parent node indices, weighted by a factor chosen to approximately equalise the contributions to the total loss of the classification and Attention Graph losses. The overall structure allows our graph elements to be created 'holistically', since the nodes are output in a parallel fashion, rather than through stepwise transition-based parsing.

Few-Shot Scene Graph Prediction [67]. The long-tailed distribution of relationships can be an obstacle for traditional approaches since they can only be trained on a small set of predicates that carry sufficient labels. Based on this observation, the authors introduce a scene graph prediction model that supports few-shot learning of predicates, enabling scene graph approaches to generalize to a set of new predicates.

The pipeline of Few-Shot Learning is a s follows. 1) Fully train Graph Convolution model and spatial and semantic shift functions on relationships with abundant data. 2) Define shift functions for new rare relationships with few examples using fully trained shift functions. 3) Fine-tune new shift functions with few training examples

The novelty of their module is that predicates are defined as functions, resulting in a scene graph model where object representations can be used for few-shot predicate prediction. Instead of using the object representations to predict predicates, this paper instead treats predicates as two individual functions: a forward function that transforms the subject representation into the object, and an inverse function that transforms the object representation back into the subject.

ARN [26] is manily propsoed to address the following two problems. One is that most of the existing works neglect the semantic relationship between the visual features and linguistic knowledge, and the intra-triplet connections. The other is that most revious works follow a stepby-step manner to capture the representation of nodes and edges, leading to neglect the global structure and information in whole image.

The proposed Attentive Relational Network (ARN) mainly consists of four parts: (1) Object Detection Module: capturing the visual feature and the location of each entity bounding box with their pair-wise relation bounding boxes. Then a softmax function is employed to obtain initial classification scores for each entity and relation; (2) Semantic Transformation Module: producing the semantic embedded representations by transforming label word embeddings and visual features into a common semantic space; (3) Graph Self-Attention Module: leveraging a self-attention mechanism to embed entities via constructing an adjacency matrix based on the space position of nodes; (4) Relation Inference Module: creating the joint global graph representation and predicting entity and relationship labels as the final scene graph result.

RelDN [68]. Since a subject or object is related to one of many instances of the same class, most models fail to distinguish between the target instance and the others. Besides, the model fails to identify the correct pairing as the image contains multiple subject-object pairs interacting in the same way. These two obstacles result in two types of errors respectively, Entity Instance Confusion and Proximal Relationship Ambiguity.

In this paper a set of Graphical Contrastive Losses are proposed to tackle these issues. The losses use the form of the margin-based triplet loss, but are specifically designed to address the two aforementioned errors. It adds additional supervision in the form of hard negatives specific to Entity Instance Confusion and Proximal Relationship Ambiguity

The proposed Relationship Detection Network (RelDN), which has two stages. The first stage of the RelDN exhaustively returns bounding box regions containing every pair. In the second stage, it computes three types of features for each relationship proposal: semantic, visual, and spatial. Each feature is used to output a set of class logits, which we combine via element wise addition, and apply softmax normalization to attain a probability distribution over predicate classes.

CMAT [69]. The coherency of the visual context is not captured effectively by existing SGG methods due to the main reason: the cross-entropy (XE) based training objective is not graph-coherent which means the detected objects and relationships should be contextually consistent but not independent, and the training objective of SGG should be local-sensitive which implies the training objective is sensitive to the change of a single node.

This paper proposes a novel training paradigm: Counterfactual critic Multi-Agent Training (CMAT), to simultaneously meet the graph-coherent and local-sensitive requirements. Its framework is as follows. Given an image, the model uses RPN to propose object regions. Then, each object (agent) communicates with others to encode visual context. After agent communication, the model predicts class confidence for all objects. Based on the confidence, it selects (random or greedily sampling) object labels and infers visual relationship of object pairs. Finally, it generates the scene graph. In the training stage, a counterfactual critic is used to calculate the individual contribution.

Objects are viewed as cooperative agents to maximize the quality of the generated scene graph in the communicative multi-agent model . The action of each agent is to predict its object class labels, and each agent can communicate with others using pairwise visual features. The communication retains the rich visual context in SGG. After several rounds of agent communication, a visual relationship model triggers the overall graph-level reward by comparing the generated scene graph with the ground-truth.

DSG [70]. Given an image and a triplet query ⟨*subject, relation, object*⟩, this paper attempts to find the bounding boxes of the subject and object that participate in the relation. This work designs Differentiable Scene-Graphs (DSG) to address the above challenges of which the architecture is as follows. The input consists of an image and a relationship query triplet subject, relation, object. A detector produces a set of bounding box proposals. An RoiAlign layer extracts object features from the backbone using the boxes. In parallel, every pair of box proposals is used for computing a union box, and pairwise features extracted in the same way as object features. These features are used as inputs to a Differentiable Scene-Graph Generator Module which outputs the Differential Scene Graph, a new and improved set of node and edge features. The DSG is used for both refining the original box proposals, as well as a Referring Relationships Classifier, which classifies each bounding box proposal as either Subject, Object, Other or Background. The ground-truth label of a proposal box will be Other if this proposal is involved in another query relationship over this image. Otherwise the ground truth label will be

Background.

DSGs are an intermediate representation trained end-to-end from the supervision for a downstream reasoning task. The key idea is to relax the discrete properties of scene graphs such that each entity and relation is described with a dense differentiable descriptor.

Triplet-Aware Scene Graph Embeddings [71]. This paper attempts to solve the layout prediction problem which predict scene layout masks and object localization (bounding boxes), based upon the structure of the scene graph. A layout prediction network is proposed as follows. A GCNN first processes an input scene graph to produce embeddings corresponding to object nodes in the graph. Singleton object embeddings are passed to the next stage of the layout prediction network to form a set of triplet embeddings where each is composed of a ⟨*subject, predicate, object* ⟩ embedding. These are passed via a triplet mask prediction network. Rather than just learn individual class labels, the network learns to label objects as either subject or object, enforcing both an ordering and relationship between objects. Triplet embeddings are also passed through a triplet superbox regression network, where the network is trained to do joint localization over subject and object bounding boxes. Ultimately, all of the outputs of the second stage of the layout prediction model are used to compose a scene layout mask with object localization.

In their work, several new supervisory signals that are conditioned upon triplet embeddings are introduced to train scene layout prediction models. Besides, data augmentation is applied by using heuristic-based relationships to maximize the number of triplets during training. The goal is to learn a triplet-aware scene graph embedding with the hypothesis additional supervision and data augmentation will enriched the embedding representation.

6) Other SGG methods with facts alone: **SG-GAN** [27]. Most of previous SGG methods use detectors to detect all the objects, and then generate the whole scene graph. Therefore, these methods have limitations of bounding boxes being available and without using the objects' attributes. The method first generates small sub-graphs, which can describe a specific region of the input image about a scene. Then, all of the generated sub-graphs are used to construct the complete scene graph. In this method, the images and noise information are first fed to a generator, then a CNN is used to extract the image features, and a dynamic image representation and attention vector are obtained using an attention mechanism. Finally, the image representations are used to produces triples by LSTM. Inspired by GAN, the triple generator is trained adversarially. While the trained triple Generator would resolve all the triples into a graph.

VRL [29] may be the first SGG method by using reinforcement learning [72]. This method is to gradually generate the scene graph, and the relationships between subjects and objects are generated in each step, so that the final complete scene graph will be gradually formed like a tree. For the whole model framework, the input states of reinforcement learning is parts of state features, including image features, subject features, object features and history phrase information. Then there are three branches of output actions, which are to determine the properties of the subjects, the relationships between the current subjects and objects and the categories of the next objects. Variation-structured reinforcement learning actually refers to that the action space of the model varies according to the state in each step, so as to reduce the action selecting space and improve the accuracy. To this end, Directed Semantic Action Graph is constructed by the training set, which is actually the statistical information of relations and attributes in the data set relative to the object categories. Finally, three reward functions is defined to reflect the detection accuracy of taking action in a specific state.

CMAT [69]. To improve the quality of scene graph, the most important thing is to improve the performances of relationship recognition. Therefore, CMAT combines objects recognition and relation recognition to effectively improve the quality of scene graph, and each object in the images is regarded as an agent. The existing algorithms use the cross entropy as the loss function of object detection and recognition, but there is a problem that the importance of each object is different. To this end, graph-level metrics (such as Recall @k [9] and SPICE [73]) are used to evaluate the detection results, and used as a supervisory signal for model training. Then, The final multi-agent policy-gradient is used to maximize the graph-level metrics.

Analogies Transfer [74]. During generating the scene graph, there are many unseen relationships of the individual entities in the dataset.In order to generate a complete scene graph, Peyre et al. proposed to use analogy transformations to detect the unseen relationships that involve similar objects for the model. The whole network model has two stages. In the first stage, all the subjects and objects are detected, and the module of visual phrase embedding is used to learn the features of subjects, objects, predicates and visual phrases by optimizing the joint loss $L_{joint} = L_s + L_o + L_p + L_{vp}$. Then if we need to identify a unseen triplet, the model can utilize analogy transformation to compute the similarity between the unseen triplet and its similar triplets to estimate this unseen relationship.

GB-NET [5]. Due to a unified formulation of the two constructs of Knowledge Graph and Scene Graph, Graph Bridging Network (GB-NET) is proposed to incorporate the combination of the rich visual and commonsense information. In GB-NET, the scene graph and entity bridges are first initialized using Faster R-CNN. Then a variant of GGNN [75] is used to propagate the messages throughout the graph to update node representations, which establishes the bridge between the instance-level visual knowledge and commonsense knowledge, and a scene graph would be generated.

In addition, A simple and effective SSG method was proposed in [76] by jointly embedding the images and scene graphs. This method try to generate a scene graph from images by investigating several existing methods based on bag-of-words, sub-path representations and GNNs.

B. SGG by introducing additional information

To generate a scene graph faster and more accurately, scene graph generation models pay more attention to introducing multiple types of prior information, such as language priors, visual priors, knowledge priors, contexts, and so on. In this section, we discuss the related works of SGG by introducing additional information.

Phrase Cues [33]. Plummer et al. proposed a model framework for localizing or grounding the phrases in the images by using a large collection of linguistic and visual cues, which is obtained from the captions. Then the single phrase cues (SPCs) and the phrase pair cues (PPCs) are used to combine with Canonical Correlation Analysis (CCA) [77] to detect visual relationships. Therefore, in [33], the introduced priors are a list of the cues with corresponding phrases from the sentence, and these cues are extracted from the captions.

Language Prior Model [9]. Given the relationship annotations between the objects, which are detected in a set of fully supervised images. Lu et.al. proposed to train a visual appearance module and a language module individually, and later to combine them together through a objective function, so as to improve the final performance to generate a scene graph.

Compared to the method of Visual Phrase, in which a separate detector is designed for every single relationship, Language Prior Model uses the visual appearance module to learn the individual visual features of its comprising objects and predicates. In computational complexity, for the N objects and K predicates, Visual Phrases would need to train $O(N^2K)$ unique detectors, while only $O(N+K)$ detectors need through visual appearance module in Language Prior Model. In addition, the language module in Language Prior Model is very novel to project relationships into a word embedding space for SGG. In language module, the similar relationships are optimized to be close together based on the semantic priors of relationships. In this way, the rare relationships can also be predicted so as to solve the problem of the long tail of infrequent relationships to a certain extent.

LK Distillation [50]. In most previous SGG methods, the visual relationship between two entities are generated. While in [50], Yu et al. try to model the three entities in a scene jointly, which can more accurately reflect these entities' relationships compared to modeling them independently. However, to reduce the complexities of model learning, the prior knowledge of linguistic statistics is used to regularize the visual feature learning. The useful linguistic knowledge can be extracted from the training relation annotations (internal knowledge) and other publicly information, such as Wikipedia. While in the teacher-student knowledge distillation framework [78], the distilled linguistic knowledge is used to predict the predicates using the visual features.

CDDN [30]. Cui et al. proposed a context-dependent diffusion network (CDDN) framework to identify the visual relationships. Before carrying out CDDN, object detectors are used to obtain the locations, labels and confidence scores of all the detected objects, which would be used as the input for CDDN. Then two types of global context information (semantic priors and spatial scenes) are used for visual relationship detection. Semantic priors are learned by a word semantic graph from language priors, and spatial scenes are obtained by a visual scene graph to extract the visual features. Then these two types of global context information are adaptively aggregated by a diffusion network to estimated the predicates.

CISC [79] is another SGG method by introducing the context information. Besides significative visual pattern is als be explored for SGG. In Relationship Context-InterSeCtion Region (CISC) method, the context for relationships is constructed to benefit the relationship recognition from their association, and the proposed intersection region are used to discover the effective visual pattern for relationship recognition.

Knowledge-embedded routing network. In the real world, the distribution of the relationships is unbalanced, which leads to the poor performance of the existing methods in recognizing the relationships with the low frequency. To solve this problem, the SGG model based on knowledge-embedded routing network is proposed by Chen et al. in [31]. In this method, a series of object regions are generated using Faster RCNN. Then, a graph network is used to propagate the features of nodes on the graph to learn the more contextualized features, so as to predict the labels in each object pair. Moreover, another graph is used to correlate the pairs of objects, and their relationships are predicted by a GNN model. The same process is repeated for all the pairs of objects to recognize their relationships, and the final scene graph is generated. Therefore, the statistical correlations between pairs of objects and their given relationships are used as the introduced priors for SGG in Knowledge-embedded routing network.

KB-GAN [32]. Since the existing scene graph datasets have the problem of the long tail in the distribution of object and relationship labels. Commonsense knowledge extracted from the external knowledge bases (KB) is used to refine object and phrase features for SGG, and an auxiliary image reconstruction path based on GAN is introduced to regularize the whole SGG network (KB-GAN). Therefore, in fact KB-GAN is also an application of scene graph on image generation.

C. videos and pixels-level for SGG

SGFB. In [80], a new data structure: Action Genome is introduced as a representation of spatio-temporal scene graphs. To generate the spatio-temporal scene graphs, Scene Graph Feature Banks (SGFB) is proposed, and the spatio-temporal scene graphs are further incorporated into a sequence of scene graph features as the final representation $F_{SG} = [f_1, f_2, ..., f_T]$, which is used to predict action labels by 3D CNNs. With Action Genome, the action recognition task has achieved better performance on the Charades dataset.

Ontology graph is proposed in [3] to describe objects, parts, actions and attributes in a scene. Ontology graph has several similarities with scene graph, for example, these two types of graph structures have objects, attributes and relationships, and both of them also have their sub-graphs. In [3], ontology graph is used for scene-centric joint-parsing of cross-view videos, and the tasks of object detection, multi-object tracking, action recognition and human attributes recognition are used to evaluate the proposed scene-centric joint-parsing framework.

Pixels2Graph [81]. the existing relationship detection methods usually have two steps: object detection and relationship recognition, while Pixels2Graph is to directly get objects and relationships from the pixels in the original images. In the method of Pixels2Graph, The elements of the scene graph, including nodes and edges, are detected first, actually that is the objects and their bounding boxes of the relations on the graph are detected. Then these elements are combined with associative embedding to form the relationships of the objects.

IV. Applications of Scene Graph

Scene graph can describe the objects in a scene and the relationships between the objects, which provides better visual representations for relevant visual tasks, and can greatly improve the model performance of these visual tasks. In this section, we stated the applications of scene graph to different types of visual tasks.

A. Image Retrieval

Image retrieval is a classic visual task in computer vision. For retrieving the target images, the query could be the content of an image or the text describing the image. Content-based image retrieval methods typically use low-level visual features. Recently, there has more and more interest in the models of jointly reasoning about the visual and textual features. However, these models have their limitations in terms of expressiveness. While text-based image retrieval methods have the problem of the inherent referential uncertainty of text-based representations. Scene graphs are a structured representation of visual scenes, and a scene graph can explicitly represents and reasons about the objects, attributes and relationships in images. Therefore, scene graph-based image retrieval has broad development prospects.

In 2015, J.Johnson et al.[1] proposed the concept of scene graph, and design a conditional random field model for image retrieval by utilizing the scene graph, which is constructed manually. In [82], a new text-based image retrieval framework is proposed based on binary representations and semantic graphs. This framework mainly consists of four parts: cross-modal binary representation, semantic graph across different modalities, the joint objective function and the online update method. In 2019, Ramnath et al. [83] proposed a neural-symbolic approach for a one-shot retrieval of images from a large scale catalog, given the caption description. To facilitate this, they represent the catalog and caption as scene-graphs and model the retrieval task as a learnable graph matching problem, trained end-to-end with a reinforce algorithm. Wang et al. [84] propose to represent image and text with two kinds of scene graphs: visual scene graph (VSG) and textual scene graph(TSG),and the image-text retrieval task is then naturally formulated as cross-modal scene graph matching. Given a query in one modality (a sentence query or an image query), the goal of the image-text cross-modal retrieval task is to find the most similar sample from the database in another modality. Therefore, their Scene Graph Matching (SGM) model aims to evaluate the similarity of the image-text pairs by dissecting the input image and text sentence into scene graphs. The framework of SGM is illustrated in Figure, which consists of two branches of networks. In the visual branch, the input image is represented into a visual scene graph (VSG) and then encoded into the visual feature graph (VFG). Simultaneously, the sentence is parsed into a textual scene graph (TSG) and then encoded into the textual feature graph (TFG) in the textual branch. Finally, the model collects object features and relationship features from the VFG and TFG and calculates the similarity score at the object-level and relationship-level, respectively.

B. Image Generation

Image generation for the complex scenes with multiple objects and desired layouts is a hot topic in computer vision research. Despite image generation based on computer vision technology has significant recent progress, it is still a difficult problem to generate the images with multi-objects and complex scenes.

Johnson et al. [35] attempted to generate a realistic image given the corresponding scene graph with object labels and their relationships by Image Generation Network (IG-Net). This problem is a rebuilding work which meets the following three challenges: how to process the graph-structured input, how to guarantee the uniformity between the generated images and their corresponding scene graphs, and how to ensure the authenticity of the synthesized images. These challenges are settled as follows. A scene graph specifying objects and relationships is used as the input in IG-Net, and is processed by a graph convolution network. IG-Net passes the information along edges to compute the embedding vectors for all detected objects, and these vectors thereafter are used to predict the objects' bounding boxes and segmentation masks. Then, these bounding boxes and segmentation masks are jointly combined to form a scene layout, which is then used to generate a rough image $\hat{I}$ using the cascaded refinement network (CRN). Subsequently, the authenticity of $\hat{I}$ is solved by adversarially training IG-Net against a pair of discriminator networks D_{img} and D_{obj}. In this processure, $\hat{I}$ is encouraged to appear realistic, and also to contain the realistic and recognizable objects. Finally, the generated images can be obtained.

To generate images, Zhao et al. proposed an end-to-end approach (Layout2Im) [85] to generate images from the layouts. In Layout2Im, the representation of each object is broken down into specified and unspecified parts. For the specified parts, The category is coded using word embedding, and for unspecified parts, the visual features are extracted as a low-dimensional vector. Then, individual object representations are grouped together using convolution LSTM to obtain the encoding of the complete layout, which is then decoded into an image. During this process, several loss terms are introduced to encourage accurate and diverse image generation.

Since the previous image generation methods cannot introduce new additional information to the existing description, and are limited to generating images at one time. Therefore, Mittal et al. [86] proposed a recursive network architecture that preserves the image content generated in previous steps and modifies the accumulated images based on newly provided scene information. This method allows to preserve the context in sequentially generated images by subjecting certain information to subsequent image generation conditions.

To solve the problem that it needs to ensure whether the generated image conforms to the scene graph, Tripathi et al. [87] propose an image generation method by harnessing scene graph context to improve image generation. In this method, a scene graph context network is introduced to pool the context features, which are generated by a GCN network. Then these pooled context features are passed to a fully-connected layer, where embeddings are generated for both the generator and the discriminator networks during training. The scene context network encourages the generated images to appear realistic, but also to respect the scene graph relationships.

In [88], a semi-parametric method (PasteGAN) is proposed by Yikang et al. for image generation based on the scene graph and the object crops. The scene graph defines the spatial arrangements of the objects and their relationships, while the given object crops are used to determine the object appearances. Then, two branches of networks are trained simultaneously: One branch focuses on diverse image generation with the object crops, which are retrieved from the external memory; While in the other branch, the original crops are used to reconstruct the ground-truth images.

To improve the quality of generated images, several previous methods are proposed for mapping Scene Graph to images, which is invariant to a set of logical equivalences. Tripathi et al. [89] proposed a new image generation method based scene graph. In this method, the scene graph representations are first enhanced with heuristic-based relations, which increases the minimal storage overhead. Then, the extreme points representations are used to supervise the scene composition network learning.

It is a challenging task of generating the realistic images with complex visual scenes, especially when we want to control the layouts of the generated images. To this end, Herzig et al. [36] present a novel model to inherently learn the canonical graph representations. In the proposed model, similar predictions can be semantically obtained from similar scene graphs, and the model networks can capture the representation independently of the objects.

To generate a narrative collage for images, Fang et al. [90] introduced a layer graph and a scene graph to represent the relative depth order and semantic relationship between the objects, so as to generate the main theme or storyline behind an image.

C. Visual-textual transformer

Since the scene graphs contain the structured semantic information in a visual scene, and the semantic information is mainly reflected in the representations of the objects, attributes and pairwise relationships in the images. Thus, the scene graph can provide beneficial priors for the vision tasks of Image/video Captioning and Visual Question Answering (VQA).

1) Image/video Captioning: Different from the traditional image captioning methods, a method with scene-graph based semantic representation for image captioning is proposed in [91]. To embed scene graph as an intermediate state, the task of image captioning is divided into two phases: concept cognition and sentence construction respectively. In this method, a CNN-RNN-SVM framework is proposed to generate the scene-graph-based sequence, which is then transformed into a bit vector, as the input of RNN in the next phase for generating the captions.

Accurately grounding text descriptions to the visual relations is critical to most language-and-vision tasks. In [92], Neural Scene Graph Generators are proposed for tackling two language-and-vision tasks of image-text matching and image captioning. Since the Scene Graph Generators can learn the visual relation features more effectively, which facilitates to ground language to visual relations. Subsequently, these two language-and-vision applications have better performance.

Since the graphical representations with conceptual positional binding can improve image captioning, a novel approach, which is derived from regional visual features of the images, is proposed for generating the image captions, and this approach is called Tensor Product Scene-Graph-Triplet Representation (TPsgtR) in [93]. In TPsgtR, the technique of neuro-symbolic embedding is introduced to embed the identified relationships among different image regions into concrete forms. These neural symbolic representations are beneficial to better definition of neural symbolic space for neuro-symbolic attention, and can be transformed to better captions for the images.

The language inductive bias is exploited as language priors in [40], and Scene Graph Auto-Encoder (SGAE) is proposed to incorporate these inductive bias into the encoder-decoder models for image captioning, which is expected to help this encoder-decoder model have less overfitting to the dataset bias. Specifically, in the textual domain, SGAE is used to learn a dictionary (D) that helps to reconstruct sentences. While in the vision-language domain, D can be shared to guide the encoder-decoder models for image captioning. Thanks to the scene graph representation of the complex structural layouts of images and sentences as well as the shared dictionary D, then the language inductive bias can be transferred across the textual and visual domains.

In [42], a new scene graph-based framework is present for unpaired image captioning, and this whole framework comprises two generators, which are an image scene graph generator and a sentence scene graph generator, and a pair of encoder and decoder: a scene graph encoder and a sentence decoder. Specifically, the scene graph encoder and the sentence decoder are trained on the text modality. Moreover, an unsupervised feature extraction method is proposed to learn the scene graph features by mapping from the visual features of the images to the textual features of the sentences.

The image captioning framework based on scene graphs is proposed in [41] to solve the problem that most previous methods treat the entities in images individually, thus lacking the structured information for generating the sentences. Therefore, the scene graphs are structured by leveraging both visual features and semantic knowledge. The visual features are extracted from the object entities by CNN models, and the semantic relationship features are learned from triples. Based on these obtained features, a hierarchical-attention-based module is designed to learn the discriminative features for word generation.

In [94], the Scene Graph Captioner (SGC) framework is proposed for image captioning. In this framework, the comprehensive structural semantic features are extracted by explicitly modeling the objects, attributes and relationships in the visual scenes, and the LSTM-based models translate these semantic features into the final text descriptions.

Storytelling from an image stream. In [4], the scene graph is used to generate the story from an image stream. The proposed SGVST models visual relations in one image and cross-images, which is conducive to image description. Experimental results show that this method can significantly improve the quality of story generation. The Scene Graph Parser converts an image into a Scene Graph G. Then, the scene Graph is input multi-modal Graph ConvNet, and the nodes in the scene graph are enhanced by Graph convolutional neural network (GCN). In order to model the interaction between images, the temporal convolutional neural network (TCN) is used to further optimize the visual representations of images. Finally, the features of relation aware, which is a set of internal relation and cross-image relation, are obtained and input to Hierarchical Decoder to generate stories.

2) Visual Question Answering: In [95], inspired by conventional QA systems that operate on knowledge graphs, an alternative approach is investigated. Specifically, the scene graphs derived from images is investigated for Visual QA: an image is abstractly represented by a graph with nodes corresponding to objects in the images and edges to object relationships. Then, the graph network (GN) is adapted to encode the scene graph and perform structured reasoning according to the input question. Since scene graphs can already capture essential information of images and graph networks, the QA method based scene graph have the potential to outperform state-of-the-art Visual QA algorithms.

A VQA method with visual attention is proposed based scene graphs [44]. In this method, natural language explanations comprising of evidences are generated for answering the questions, which are asked to images using two sources of information: the entity annotations generated from the scene graphs and the attention map generated by a VQA model.

For achieving the tasks of visual question answering and visual relationship detection, a new multimodal fusion model *BLOCK* is proposed based on the block-superdiagonal tensor decomposition [43] to represent the fine interactions between multi-modalities, while the powerful mono-modal representations are also maintained. Moreover, the end-to-end learnable architectures are designed for representing the relevant interactions between modalities.

In [96], a Scene Graph Convolutional Network (Scene GCN) is designed to jointly reason the object properties and relational semantics for VQA task. In this method, to effectively represent visual relational semantics, a visual relationship encoder is built to yield discriminative and type-aware visual relationship embeddings constrained by both the visual context and language priors. Moreover, SceneGCN is proposed to reason about the visual clues for the correct answer under the guidance of the question.

D. Visual social and human-object relationship detection

In this section, we will discuss the methods for visual social relationship recognition and visual human-object relationship recognition by using scene graph.

Social relationships are the foundation of human social structure. Developing computational models to understand social relationships from visual data is critical to building intelligent machines that can better interact with humans in social environments. In [37], a Dual-Glance model is proposed for social relationship recognition. In this method, the person of interest is detected first, and then attention mechanisms are used to exploit contextual cues. Furthermore, Li et al. proposed an Adaptive Focal Loss to leverage the ambiguous annotations for more effective learning to solve the problem that visually identifying social relationship bears certain degree of uncertainty.

The pose-guided Person-Object Graph and Person-Pose Graph [97] are proposed to model the actions from persons to object and the interactions between paired persons, respectively. Based on the graphs, social relation reasoning is performed by graph convolutional networks. One branch is designed to learn global features from the whole image. A deep CNN, i.e., ResNet is used to learn knowledge about the scenes for social relation recognition. The other branch is focused on regional cues and fine interactions among persons and contextual objects for social relation reasoning, and contains three main procedures. Social relation reasoning is performed on the two graphs by graph convolutional networks. The social relation between a pair of persons is predicted by integrating the global feature from CNN and the reasoning feature from the GCNs.

Adversarial adaptation of scene graph model is present for understanding civic issues in [39]. In this model, Faster R-CNN provides the object labels and their bounding regions. Object context generates a contextualized representation for each object. Edge context generates a contextualized representation for each edge using the representation of the object pairs. During adversarial training, information regarding the edge contexts passed on to the Discriminator, which learns to distinguish between the seen and unseen object pairs. The training objective of the Discriminator results in gradients flowing into the Discriminator as well as the edge context layer. The loss for the model decreases as the model learns to fool the Discriminator by adapting a uniform representation for seen and unseen classes.

Visual relationship recognition aims at interpreting rich interactions between a pair of localized objects. Zoom-Net in [38] is proposed to mining deep feature interactions for visual relationship recognition, and the method of Spatiality-Context-Appearance Module (SCA-M) the core of Zoom-Net, and attempts to capture contextual information by directly fusing pairwise features. The proposed SCA-M integrates the local and global contextual information in a spatiality-aware manner, and three classifiers with intra-hierarchy structures are applied to the features obtained from each branch for visual relationship recognition.

To solve diverse interactions problem, Plesse et al. [98]

proposed guided proposal framework, Semantic knowledge distillation and Internal knowledge distillation. Object detection is only the first step towards image understanding, as images are more than the sum of their parts and can not be fully understood without the relationships between these objects. such tasks have been enabled by the releases of large scale datasets providing bounding box annotations paired with natural language descriptions, or triplet annotations. Predicates are semantically similar when they appear in similar contexts. The purpose of this method is to restrict the outputs to a subset of predicates that are the most probable for a given pair of objects.

Recognizing human object interactions (HOI) is an important part of distinguishing the rich variety of human action in the visual world. A novel method for Human-Object Interactions (HOI) recognition is proposed in [99]. In this method, HO-RCNN detects HOIs in two steps. First, proposals of human-object region pairs are generated by using state-of the-art human and object detectors. Then, each human object proposal is passed into a ConvNet to generate HOI classification scores. The whole network adopts a multi-stream architecture to extract features on the detected humans, objects, and human-object spatial relations. Given a human-object proposal, HO-RCNN classifies its HOIs using a multi-stream network, where different streams extract features from differents sources.

Furthermore, a multi-task approach based on Zero-Shot Learning is proposed in [100] to scale all combinations of human-object interactions. This approach address the challenge of scaling human object interaction recognition by introducing an approach for zero-shot learning that reasons on the decomposition of HOIs as verbs and objects. Specifically, a factorized model consisting of both shared neural network layers as well as independent verb and object networks is introduced. The entire model is trained jointly in a multi-task fashion. For test time, the scores are calculated for all combinations of verb-object prediction pairs to produce the final HOI prediction where the verb and object are tightly localized.

In [101], Transferable interactiveness Prior, which indicates whether human and object interact with each other or not, is explored for human-object interaction detection. The interactiveness prior can be learned across HOI datasets, regardless of HOI category settings. Therefore, the core idea is to exploit an Interactiveness Network to learn the general interactiveness prior from multiple HOI datasets and perform Non-Interaction Suppression before HOI classification in inference.

InteractNet is proposed in [102] to detecting and recognizing Human-Object interactions. This network model is driven by a human-centric approach, and would be used to address the task of detecting $\langle human, verb, object\rangle$ triplets in challenging everyday photos. There is a hypothesis that the appearance of a person is a powerful cue for localizing the objects they are interacting with. To exploit this cue, the model learns to predict an action-specific density over target object locations based on the appearance of a detected person. Moreover, the proposed model also jointly learns to detect people and objects, and by fusing these predictions it efficiently infers interaction triplets in a clean, jointly trained end-to-end system.

Graph Parsing Neural Network [103] is proposed by Qi et al. for addressing the task of detecting and recognizing human-object interactions (HOI) in images and videos. GPNN is a framework that incorporates structural knowledge while being differentiable end-to-end. For a given scene, GPNN infers a parse graph that includes the HOI graph structure represented by an adjacency matrix and the node labels. Within a message passing inference framework, GPNN iteratively computes the adjacency matrices and node labels.

E. image understanding and referring

Scene graphs allows us to reason about the objects and their relationships as compared to an unstructured text description. Possible layouts of the images are then inferred from the scene graph representation. $\langle subject, relation, object\rangle$ is the key to image understanding and reasoning [104], [105], [106]. To understand an image, it needs to recognize different components (objects, actions, scenes) and infer higher-level events, activities, and background context. In addition, to detect and infer such information needs a combination of vision modules, reasoning modules, and background knowledge.

Wang et al.[107] proposed a deep convolutional neural network to increase segmentation accuracy by learning from an Image Descriptions in the Wild (IDW-CNN), which has three important parts, including a ResNet-101 network for feature extraction, a network stream predicts its segmentation label-map, and another stream estimates its object interactions. IDW-CNN jointly trains IDW and existing image segmentation dataset, and fully explores the knowledge from different datasets, thus improves the performance of both datasets. As only weak labels are used, so IDW-CNN can also be used Semi- and Weakly-supervised Image Segmentation.

Aditya et al.[104] present an intermediate knowledge structure called Scene Description Graph (SDG), which uses a deep learning-based perception system to obtain the objects, scenes and constituents with probabilistic weights from an input image. A common-sense knowledge base is built from image annotations along with a Bayesian Network of commonly occurring objects and scene constituents (the concepts that can not be seen, but can be understood from the scene) are inferred to predict how the objects interact in the scene.

Zhang et al.[105] made a research on relationship recognition at an unprecedented scale, where the total number of visual entities is more than 80,000. An image is input to the visual module, and three visual embeddings x_s, x_p, and x_o for subject, relation, and object can be obtained. To this end, a continuous output space is used for objects and relations instead of discrete labels, and a new relationship detection model is developed tp embed objects and relations into two vector spaces, and learns a visual and a semantic module to map the features from the two modalities into a shared space.

Shi et al. [106] advanced NMN towards X visual reasoning by using the proposed explainable and eXplicit Neural Modules (XNMs) reasoning over scene graphs. The scene graph can insulate the “low-level” visual perception from the modules, and thus can prevent reasoning shortcut of both language and vision counterpart. A scene graph is the knowledge representation of a visual input, where the nodes are the entities and the edges are the relationships between entities. Given an input image and a question, first parse the image into a scene graph and parse the question into a module program, and then execute the program over the scene graph. A set of generic base modules are proposed, and this modules can conduct reasoning over scene graphs— explainable and eXplicit Neural Modules (XNMs) —as the reasoning building blocks. Besides, XNMs are totally attention-based, making all the intermediate reasoning steps transparent.

Generating semantic layout from scene graph is a crucial intermediate task of connecting text to image. To Learn the relation from semantic description to its visual incarnation leads to important applications, such as text-to-image synthesis and semantic image retrieval. The underlying taskS of inferring semantic layout from scene graph and connecting text to image are achieved in [108], [109].

Li et al.[108] proposed a conceptually simple, flexible and general framework using sequence to sequence (seq-to-seq) learning to infer semantic layout from scene graph called Seq-SG2SL, which derives sequence proxies for the two modality, and a Transformer-based seq-to-seq model learns to transduce one into the other. A scene graph is decomposed into a sequence of semantic fragments (SF), one for each relationship. A semantic layout is the consequence from a series of brick-action code segments (BACS), dictating the position and scale of each object bounding box in the layout. Viewing the two building blocks, SF and BACS, as corresponding terms in two different vocabularies, a seq-to-seq model is fittingly used to translate. Seq-SG2SL is an intuitive framework that learns BACS to drag-and-drop and scale-adjust the two bounding boxes of subject and object in a relationship to the layout supervised by its SF counterpart.

Advancements on text-to-image synthesis generate remarkable images from textual descriptions. However, these methods are designed to generate only one object with varying attributes. Talavera et al. [109] proposed a method that infers object layouts from scene graphs has been proposed as a solution to this problem, and an object encoding module is designed to capture object features and use it as additional information to the image generation network. The goal is to generate an image that matches the descriptions provided in an input scene graph.

The task of eferring Expression Grounding (REF) is to localize a region in an image, where the region is described by a natural language expression. To achieve this task fundamentally, it should first find out the contextual objects and then exploit them to disambiguate the referent from other similar objects by using the attributes and relationships. Liu et al.[110] present a novel REF framework called Marginalized Scene Graph Likelihood (MSGL), which jointly models all the objects mentioned in the referring expression, and hence allows the visual reasoning with the referent and its contexts. Compared with the other discriminative models which neglect the rich linguistic structure and focus on holistic grounding score calculation, MSGL exploit the full linguistic structure. MSGL first constructs a CRF model based on scene graphs, parsed from the sentences, and then marginalizes out the unlabeled contexts by belief propagation.

F. 3D scene graph

3-D scene graph is defined in [111] by Kim et al. to represent the physical environments in a sparse and semantic way, and a 3-D scene graph construction framework is also proposed. Similar to 2D scene graph generated from 2D images, 3-D scene graph describes the environments compactly by abstracting the environments as graphs, where nodes depict the objects and edges characterize the relations between the pairs of objects. As the proposed 3-D scene graph illustrates the environments in a sparse manner, the graph can cover up an extensive range of physical spaces, which guarantees the scalability. Furthermore, the applicability of the 3-D scene graph is verified by demonstrating two major applications: visual question and answering (VQA) and task planning, and achieved better performance than the traditional methods.

3D Scene Graph can provides numerically accurate quantification to relationships, thus 3-D scene graph as an environment model and the 3-D scene graph construction framework has got excellent scores. In [2], a 3-D scene graph construction method is proposed. The input to this method is the typical output of 3D scanners and consists of 3D mesh models, registered RGB panoramas and the corresponding camera parameters. Each panorama is densely sampled for rectilinear images. Mask R-CNN detection on them are aggregated back on the panoramas with a weighted majority voting scheme. The output is the 3D Scene Graph of the scanned space, which formulates as a four layered graph. Each layer has a set of nodes, each node has a set of attributes, and there are edges between nodes which represent their relationships. Single panorama projections are then aggregated on the 3D mesh. Finally, These detections become the nodes of 3D Scene Graph. A subsequent automated step calculates the remaining attributes and relationships.

In [112], Yang et al.proposed a novel method for inferring precise support relations, and introduced a framework for constructing semantic scene graphs and assessing the quality. In this method, a Convolutional Neural Network is used to detect objects in the given images. Then, the precise support relations between objects are inferred by taking two important auxiliary information in the indoor environments. Finally, a semantic scene graph describing the contextual relations within a cluttered indoor scene is constructed. Compared with the previous methods for extracting support relations, this proposed approach provides more accurate results.

V. Conclusion

It is always the goal of computer vision to have a deep understanding of a scene, and then be able to reason about relevant events, even some unseen events. Since scene graph, a new content for scene description, is proposed in 2015, subsequently, a wave of research works on scene graph generation and application has been set off. Scene graph is a type of data structure that describes the objects, attributes and the relationship between objects in a scene, and has powerful expression for the scene. While the first scene graph is established manually. Subsequently, many scene graph generation methods are proposed to build a more complete scene graph by a variety of network models, feature extraction methods, and even by introducing the prior knowledge. Meanwhile, some relevant models and methods are designed to reduce the computational complexity of scene graph generation. Furthermore, there are also many research works on applying scene graph to different types of visual tasks, such as image retrieval, image generation, image/video caption and so on. Due to the scene graph's powerful ability of scene representation and the introduction of relevant knowledge information, the performances of these visual tasks are greatly improved. Therefore, this paper gives a systematic overview of the current researches on scene graph generation and application. For scene graph generation, the model types of object relation recognition are classified; while we categorize scene graph applications according to the visual tasks. The review of scene graph generation and application is to summarize the latest scene graph research, point out the problems that still need to be solved in future scene graph research. We expect this review can provide an overall technical reference for scene graph research.

Acknowledgment

This work is supported in part by NSFC grant 61702415, Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) under grant no. DE190100626, Air Force Research Laboratory and DARPA under agreement number FA8750- 19-2-0501.

References

[1] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3668–3678.

[2] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5664–5673.

[3] H. Qi, Y. Xu, T. Yuan, T. Wu, and S.-C. Zhu, "Scene-centric joint parsing of cross-view videos," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[4] R. Wang, Z. Wei, P. Li, Q. Zhang, and X. Huang, "Storytelling from an image stream using scene graphs."

[5] A. Zareian, S. Karaman, and S.-F. Chang, "Bridging knowledge graphs to generate scene graphs," arXiv preprint arXiv:2001.02314, 2020.

[6] E. E. Aksoy, A. Abramov, F. Wörgötter, and B. Dellen, "Categorizing object-action relations from semantic scene graphs," in 2010 IEEE International Conference on Robotics and Automation. IEEE, 2010, pp. 398–405.

[7] S. Aditya, Y. Yang, C. Baral, C. Fermuller, and Y. Aloimonos, "From images to sentences through scene description graphs using commonsense reasoning and knowledge," arXiv preprint arXiv:1511.03292, 2015.

[8] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in Proceedings of the fourth workshop on vision and language, 2015, pp. 70–80.

[9] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in European conference on computer vision. Springer, 2016, pp. 852–869.

[10] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision, vol. 123, no. 1, pp. 32–73, 2017.

[11] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5179–5188.

[12] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. van den Hengel, "Hcvrd: a benchmark for large-scale human-centered visual relationship detection," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[13] L. D. Dai Bo, Zhang Yuqi, "Detecting visual relationships with deep relational networks," in Proceedings of the IEEE conference on computer vision and Pattern recognition, 2017, pp. 3076–3086.

[14] W. Cong, W. Wang, and W.-C. Lee, "Scene graph generation via conditional random fields," arXiv preprint arXiv:1811.08075, 2018.

[15] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5532–5540.

[16] Z.-S. Hung, A. Mallya, and S. Lazebnik, "Union visual translation embedding for visual relationship detection and scene graph generation," arXiv preprint arXiv:1905.11624, 2019.

[17] N. Gkanatsios, V. Pitsikalis, P. Koutras, A. Zlatintsi, and P. Maragos, "Deeply supervised multimodal attentional translation embeddings for visual relationship detection," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 1840–1844.

[18] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, "Relationship proposal networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5678–5686.

[19] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, and T. Mei, "Vrr-vg: Refocusing visually-relevant relationships," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10 403–10 412.

[20] Y. Li, W. Ouyang, X. Wang, and X. Tang, "Vip-cnn: Visual phrase guided convolutional neural network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1347–1356.

[21] Y. Chen, Y. Wang, Y. Zhang, and Y. Guo, "Panet: A context based predicate association network for scene graph generation," in 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019, pp. 508–513.

[22] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5831–5840.

[23] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6619–6628.

[24] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1261–1270.

[25] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable net: an efficient subgraph-based framework for scene graph generation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 335–351.

[26] M. Qi, W. Li, Z. Yang, Y. Wang, and J. Luo, "Attentive relational networks for mapping images to scene graphs," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3957–3966.

[27] M. Klawonn and E. Heim, "Generating triples with adversarial networks for scene graph construction," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[28] Y.-S. Wang, C. Liu, X. Zeng, and A. Yuille, "Scene graph parsing as dependency parsing," arXiv preprint arXiv:1803.09189, 2018.

[29] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 848–857.

[30] Z. Cui, C. Xu, W. Zheng, and J. Yang, "Context-dependent diffusion network for visual relationship detection," in Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 1475–1482.

[31] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, 2019, pp. 6163–6171.
[32] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1969–1978.
[33] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, "Phrase localization and visual relationship detection with comprehensive image-language cues," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1928–1937.
[34] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," arXiv preprint arXiv:1910.05134, 2019.
[35] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1219–1228.
[36] R. Herzig, A. Bar, H. Xu, G. Chechik, T. Darrell, and A. Globerson, "Learning canonical representations for scene graph to image generation," arXiv preprint arXiv:1912.07414, 2019.
[37] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Visual social relationship recognition," arXiv preprint arXiv:1812.05917, 2018.
[38] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. Change Loy, "Zoom-net: Mining deep feature interactions for visual relationship recognition," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 322–338.
[39] S. Kumar, S. Atreja, A. Singh, and M. Jain, "Adversarial adaptation of scene graph models for understanding civic issues," in The World Wide Web Conference, 2019, pp. 2943–2949.
[40] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 685–10 694.
[41] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," IEEE Transactions on Multimedia, vol. 21, no. 8, pp. 2117–2130, 2019.
[42] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10 323–10 332.
[43] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8102–8109.
[44] S. Ghosh, G. Burachas, A. Ray, and A. Ziskind, "Generating natural language explanations for visual question answering using scene graphs and visual attention," arXiv preprint arXiv:1902.05715, 2019.
[45] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "The new data and new challenges in multimedia research," arXiv preprint arXiv:1503.01817, vol. 1, no. 8, 2015.
[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.
[47] F. A. Sadeghi Mohammad Amin, "Recognition using visual phrases," in Computer Vision and Pattern Recognition (CVPR), 2011.
[48] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5410–5419.
[49] B. R. M. Y. Y. Wentong Liao, Lin Shuai, "Natural language guided visual relationship detection," in arXiv preprint arXiv:1711.06032, 2017, pp. 1–12.
[50] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 1974–1982.
[51] S. Woo, D. Kim, D. Cho, and I. S. Kweon, "Linknet: Relational embedding for scene graph," in Advances in Neural Information Processing Systems, 2018, pp. 560–570.
[52] G. R. S. J. Ren Shaoqing, He Kaiming, "Faster r-cnn:towards real-time object detection with region proposal networks," IEEE Transactions on Pattern and Analysis Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2015.
[53] A. G.-D. Antoine Bordes, Nicolas Usunier, "Translating embeddings for modeling multi-relational data," in NIPS, 2013.
[54] D. E.-S. R. C.-Y. F. A. C. B. Wei Liu, Dragomir Anguelov, "Ssd: Single shot multibox detector," in ECCV, 2016, pp. 21–37.
[55] D. S. G. R. . F. A. Redmon, J., "You only look once: Unified, real-time object detection," in CVPR, 2015.
[56] H. Wan, Y. Luo, B. Peng, and W.-S. Zheng, "Representation learning for scene graph completion via jointly structural and visual embedding." in IJCAI, 2018, pp. 949–956.
[57] Y. Liu, R. Wang, S. Shan, and X. Chen, "Structure inference net: Object detection using scene-level context and instance-level relationships," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6985–6994.
[58] J. Zhang, K. Shih, A. Tao, B. Catanzaro, and A. Elgammal, "An interpretable model for scene graph generation," arXiv preprint arXiv:1811.09543, 2018.
[59] Z. A. Simonyan Karen, "Very deep convolutional networks for large-scale image recognition," 2014.
[60] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari, "Detecting visual relationships using box attention," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
[61] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1115–1124.
[62] W. Gao, Y. Zhu, W. Zhang, K. Zhang, and H. Gao, "A hierarchical recurrent approach to predict scene graphs from a visual-attention-oriented perspective," Computational Intelligence, vol. 35, no. 3, pp. 496–516, 2019.
[63] S. Jae Hwang, S. N. Ravi, Z. Tao, H. J. Kim, M. D. Collins, and V. Singh, "Tensorize, factorize and regularize: Robust visual relationship learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1014–1023.
[64] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 670–685.
[65] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson, "Mapping images to scene graphs with permutation-invariant structured prediction," in Advances in Neural Information Processing Systems, 2018, pp. 7211–7221.
[66] M. Andrews, Y. K. Chia, and S. Witteveen, "Scene graph parsing by attention graph," NIPS, 2018.
[67] A. Dornadula, A. Narcomey, R. Krishna, M. Bernstein, and F.-F. Li, "Visual relationships as functions: Enabling few-shot scene graph prediction," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
[68] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph generation," arXiv preprint arXiv:1903.02728, 2019.
[69] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang, "Counterfactual critic multi-agent training for scene graph generation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4613–4623.
[70] M. Raboh, R. Herzig, J. Berant, G. Chechik, and A. Globerson, "Differentiable scene graphs," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
[71] B. Schroeder, S. Tripathi, and H. Tang, "Triplet-aware scene graph embeddings," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
[72] D. S. A. A. R.-J. V. M. G. B. A. G. M. R. A. K. F. G. O. V. Mnih, K. Kavukcuoglu, "Human-level control through deep reinforcement learning," vol. 518, no. 7540, pp. 529–533, 2015.
[73] M. J. Peter Anderson, Basura Fernando and S. Gould., "Spice: Semantic propositional image caption evaluation," ECCV, 2016.
[74] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1981–1990.
[75] B. M. Z. R. Li Yujia, Tarlow Daniel, "Gated graph sequence neural networks," Computer Science, 2015.
[76] E. Belilovsky, M. Blaschko, J. Kiros, R. Urtasun, and R. Zemel, "Joint embeddings of scene graphs and images," 2017.
[77] I. M. L. S. Gong Yunchao, Ke Qifa, "A multi-view embedding space for modeling internet images, tags, and their semantics," International Journal of Computer Vision, vol. 106, no. 2, 2012.
[78] L. Z. H. E. X. E. Hu Zhiting, Ma Xuezhe, "Harnessing deep neural networks with logic rules," pp. 2410–2420, 2016.
[79] W. Wang, R. Wang, S. Shan, and X. Chen, "Exploring context and visual pattern of relationship for scene graph generation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8188–8197.

[80] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, "Action genome: Actions as composition of spatio-temporal scene graphs," arXiv preprint arXiv:1912.06992, 2019.

[81] A. Newell and J. Deng, "Pixels to graphs by associative embedding," in Advances in neural information processing systems, 2017, pp. 2171–2180.

[82] M. Qi, Y. Wang, and A. Li, "Online cross-modal scene retrieval by binary representation and semantic graph," in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 744–752.

[83] S. Ramnath, A. Saha, S. Chakrabarti, and M. M. Khapra, "Scene graph based image retrieval–a case study on the clevr dataset," arXiv preprint arXiv:1911.00850, 2019.

[84] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 1508–1517.

[85] B. Zhao, L. Meng, W. Yin, and L. Sigal, "Image generation from layout," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8584–8593.

[86] A. A. M. S. M. T. Mittal Gaurav, Agrawal Shubham, "Interactive image generation using scene graphs," 2019.

[87] A. B. H. T. Subarna Tripathi, Anahita Bhiwandiwalla, "Using scene graph context to improve image generation," 2019.

[88] L. Yikang, T. Ma, Y. Bai, N. Duan, S. Wei, and X. Wang, "Pastegan: A semi-parametric method to generate image from scene graph," in Advances in Neural Information Processing Systems, 2019, pp. 3950–3960.

[89] S. Tripathi, S. Nittur Sridhar, S. Sundaresan, and H. Tang, "Compact scene graphs for layout composition and patch retrieval," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[90] H. F. S. H. C. X. Fei Fang, Miao Yi, "Narrative collage of image collections by scene graph recombination," IEEE Trans. Vis. Comput. Graph., vol. 24, no. 9, pp. 2559–2572, 2018.

[91] L. Gao, B. Wang, and W. Wang, "Image captioning with scene-graph based semantic concepts," in Proceedings of the 2018 10th International Conference on Machine Learning and Computing, 2018, pp. 225–229.

[92] K.-H. Lee, H. Palangi, X. Chen, H. Hu, and J. Gao, "Learning visual relation priors for image-text matching and image captioning with neural scene graph generators," arXiv preprint arXiv:1909.09953, 2019.

[93] C. Sur, "Tpsgtr: Neural-symbolic tensor product scene-graph-triplet representation for image captioning," arXiv preprint arXiv:1911.10115, 2019.

[94] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su, "Scene graph captioner: Image captioning based on structural visual representation," Journal of Visual Communication and Image Representation, vol. 58, pp. 477–485, 2019.

[95] C. Zhang, W.-L. Chao, and D. Xuan, "An empirical study on leveraging scene graphs for visual question answering," arXiv preprint arXiv:1907.12133, 2019.

[96] Z. Yang, Z. Qin, J. Yu, and Y. Hu, "Scene graph reasoning with prior visual relationship for visual question answering," arXiv preprint arXiv:1812.09681, 2018.

[97] M. Zhang, X. Liu, W. Liu, A. Zhou, H. Ma, and T. Mei, "Multi-granularity reasoning for social relation recognition from images," in 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019, pp. 1618–1623.

[98] F. Plesse, A. Ginsca, B. Delezoide, and F. Prêteux, "Visual relationship detection based on guided proposals and semantic knowledge distillation," in 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018, pp. 1–6.

[99] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[100] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018, pp. 1568–1576.

[101] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y.-F. Wang, and C. Lu, "Transferable interactiveness prior for human-object interaction detection," arXiv preprint arXiv:1811.08264, 2018.

[102] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359–8367.

[103] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 401–417.

[104] S. Aditya, Y. Yang, C. Baral, Y. Aloimonos, and C. Fermüller, "Image understanding using vision and reasoning through scene description graph," Computer Vision and Image Understanding, vol. 173, pp. 33–45, 2018.

[105] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, "Large-scale visual relationship understanding," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 9185–9194.

[106] J. Shi, H. Zhang, and J. Li, "Explainable and explicit visual reasoning over scene graphs," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8376–8384.

[107] G. Wang, P. Luo, L. Lin, and X. Wang, "Learning object interactions and descriptions for semantic image segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5859–5867.

[108] B. Li, B. Zhuang, M. Li, and J. Gu, "Seq-sg2sl: Inferring semantic layout from scene graph through sequence to sequence learning," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7435–7443.

[109] A. Talavera, D. S. Tan, A. Azcarraga, and K.-L. Hua, "Layout and context understanding for image synthesis with scene graphs," in 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 1905–1909.

[110] D. Liu, H. Zhang, Z.-J. Zha, and F. Wang, "Referring expression grounding by marginalizing scene graph likelihood," arXiv preprint arXiv:1906.03561, 2019.

[111] U.-H. Kim, J.-M. Park, T.-J. Song, and J.-H. Kim, "3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents," IEEE transactions on cybernetics, 2019.

[112] M. Y. Yang, W. Liao, H. Ackermann, and B. Rosenhahn, "On support relations and semantic scene graphs," ISPRS journal of photogrammetry and remote sensing, vol. 131, pp. 15–25, 2017.