Neurocomputing 566 (2024) 127052

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Survey Paper

Scene Graph Generation: A comprehensive survey

Hongsheng Li [a], Guangming Zhu [a], Liang Zhang [a,*], Youliang Jiang [a], Yixuan Dang [a], Haoran Hou [a], Peiyi Shen [a], Xia Zhao [a], Syed Afaq Ali Shah [b], Mohammed Bennamoun [c]

[a] School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, 710071, China
[b] Centre for AI and Machine Learning, Edith Cowan University, Joondalup, Australia
[c] School of Computer Science and Software Engineering, The University of Western Australia, Perth, Australia

ARTICLE INFO

Communicated by W. Wang

Keywords:
Scene Graph Generation
Visual relationship detection
Object detection
Scene understanding

ABSTRACT

Deep learning techniques have led to remarkable breakthroughs in the field of object detection and have spawned a lot of scene-understanding tasks in recent years. Scene graph has been the focus of research because of its powerful semantic representation and applications to scene understanding. Scene Graph Generation (SGG) refers to the task of automatically mapping an image or a video into a semantic structural scene graph, which requires the correct labeling of detected objects and their relationships. In this paper, a comprehensive survey of recent achievements is provided. This survey attempts to connect and systematize the existing visual relationship detection methods, to summarize, and interpret the mechanisms and the strategies of SGG in a comprehensive way. Deep discussions about current existing problems and future research directions are given at last. This survey will help readers to develop a better understanding of the current researches.

1. Introduction

The ultimate goal of computer vision (CV) is to build intelligent systems, which can extract valuable information from digital images, videos, or other modalities as humans do. In the past decades, machine learning (ML) has significantly contributed to the progress of CV. Inspired by the ability of humans to interpret and understand visual scenes effortlessly, *visual scene understanding* has long been advocated as the holy grail of CV and has already attracted much attention from the research community.

Visual scene understanding includes numerous sub-tasks, which can be generally divided into two parts: recognition and application tasks. These recognition tasks can be described at several semantic levels. The earlier works which mainly concentrated on image classification, only assign a single label to an image, e.g., an image of a cat or a car, and go further in assigning multiple annotations without localizing where in the image each annotation belongs [1]. A large number of neural network models have emerged and even achieved near humanlike performance in image classification tasks [2–5]. Furthermore, several other complex tasks, such as semantic segmentation at the pixel level, object detection and instance segmentation at the instance level, have suggested the decomposition of an image into foreground objects vs background clutter. The pixel-level tasks aim at classifying each pixel of an image (or several) into an instance, where each instance (or category) corresponds to a class [6]. The instance-level tasks focus on the detection and recognition of individual objects in the given scene and delineating an object with a bounding box or a segmentation mask, respectively. A recently proposed approach named Panoptic Segmentation (PS) takes into account both per-pixel class and instance labels [7]. With the advancement of Deep Neural Networks (DNN), we have witnessed important breakthroughs in object-centric tasks and various commercialized applications based on existing state-of-the-art models [8–12]. However, scene understanding goes beyond the localization of objects. The higher-level tasks lay emphasis on exploring the rich semantic relationships between objects, as well as the interaction of objects with their surroundings, such as visual relationship detection (VRD) [13–16] and human–object interaction (HOI) [17–19]. These tasks are equally significant and more challenging. To a certain extent, their development depends on the performance of individual instance recognition techniques. Meanwhile, the deeper semantic understanding of image content can also contribute to visual recognition tasks [20–24]. Divvala et al. [25] investigated various forms of context models, which can improve the accuracy of object-centric recognition tasks. In the last few years, researchers have combined computer vision with natural language processing (NLP) and proposed a number of

* Corresponding author.
E-mail addresses: hsli@stu.xidian.edu.cn (H. Li), gmzhu@xidian.edu.cn (G. Zhu), liangzhang@xidian.edu.cn (L. Zhang), mqjyl2012@163.com (Y. Jiang), dyx4work@gmail.com (Y. Dang), unse3ry@gmail.com (H. Hou), pyshen@xidian.edu.cn (P. Shen), zxmdi@163.com (X. Zhao), afaq.shah@ecu.edu.au (S.A.A. Shah), mohammed.bennamoun@uwa.edu.au (M. Bennamoun).

https://doi.org/10.1016/j.neucom.2023.127052
Received 26 June 2023; Received in revised form 11 November 2023; Accepted 14 November 2023
Available online 20 November 2023

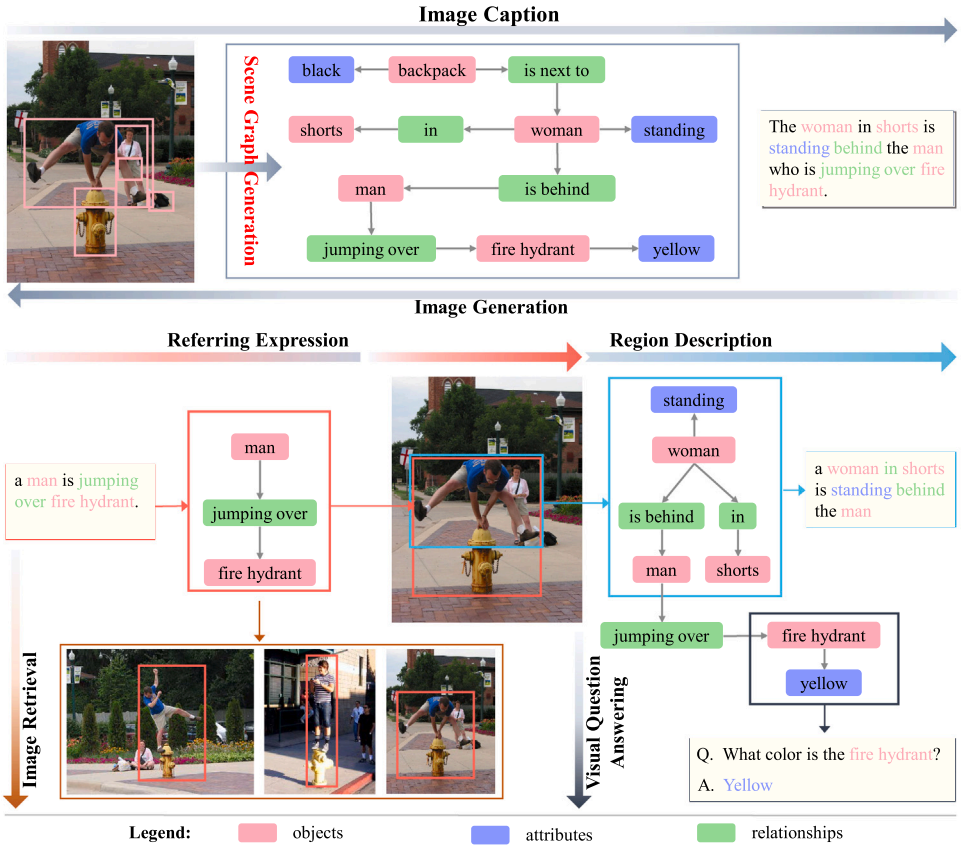


Fig. 1. A visual illustration of a scene graph and some applications. **Scene Graph Generation** takes an image as an input and generate a visually-grounded scene graph. **Image Caption** can be generated from a scene graph directly. In contrast, **Image Generation** inverts the process by generating realistic images from a given sentence or scene graph. The **Referring Expression** marks a region of the input image corresponding to the given expression, while the region and expression map the same subgraph of the scene graph. Scene graph-based **Image Retrieval** takes a query as an input, and regards the retrieval as a scene graph matching problem. For the **Visual Question Answering** task, the answer can sometimes be found directly on the scene graph, even for the more complex visual reasoning, the scene graph is also helpful.

advanced research directions, such as image captioning, visual question answering (VQA), visual dialog and so on. These vision-and-language topics require a rich understanding of our visual world and offer various application scenarios of intelligent systems.

Although rapid advances have been achieved in the scene understanding at all levels, there is still a long way to go. Overall perception and effective representation of information are still bottlenecks. As indicated by a series of previous works [26–28], building an efficient structured representation that captures comprehensive semantic knowledge is a crucial step towards a deeper understanding of visual scenes. Such representation can not only offer contextual cues for fundamental recognition challenges, but also provide a promising alternative to high-level intelligence vision tasks. *Scene graph*, proposed by Johnson et al. [26], is a visually-grounded graph over the object instances in a specific scene, where the nodes correspond to object bounding boxes with their object categories, and the edges represent their pair-wise relationships.

Because of the structured abstraction and greater semantic representation capacity compared to image features, scene graph has the instinctive potential to tackle and improve other vision tasks. As shown in Fig. 1, a scene graph parses the image to a simple and meaningful structure and acts as a bridge between the visual scene and textual description. Many tasks that combine vision and language can be handled with scene graphs, including image captioning [29–31], visual question answering [32,33], content-based image retrieval [26,34], image generation [35,36] and referring expression comprehension [37]. Some tasks take an image as an input and parse it into a scene graph, and then generate a reasonable text as output. Other tasks invert the process by extracting scene graphs from the text description and then generate realistic images or retrieve the corresponding visual scene.

Chang et al. [38] have produced a thorough survey on scene graph generation, which analyzes the SGG methods based on five typical models (CRF, TransE, CNN, RNN/LSTM and GNN) and also includes a discussion of important contributions by prior knowledge. Moreover, a detailed investigation of the main applications of scene graphs was also provided.

The current survey focuses on visual relationship detection of SGG, and our survey's organization is based on feature representation and refinement. Specifically, we first provide a comprehensive and systematic review of 2D SGG. In addition to multimodal features, prior information and commonsense knowledge to help overcome the long-tailed distribution and the large intra-class diversity problems, is also provided. To refine the local features and fuse the contextual information for high-quality relationship prediction, we analyze some mechanisms, such as message passing, attention, and visual translation embedding. In addition to 2D SGG, spatio-temporal and 3D SGG are also examined. Further, a detailed discussion of the most common datasets is provided together with performance evaluation measures. Finally, a comprehensive and systematic review of the most recent research on the generation of scene graphs is presented.

We provide a survey of 138 papers on SGG,[1] which have appeared since 2016 in the leading computer vision, pattern recognition, and machine learning conferences and journals. Our goal is to help the reader study and understand this research topic, which has gained a significant momentum in the past few years. The main contributions of this article are as follows:

[1] We provide a curated list of scene graph generation methods, publicly accessible at https://github.com/mqjyl/awesome-scene-graph.

1. A comprehensive review of 138 papers on scene graph generation is presented, covering nearly all of the current literature on this topic.
2. A systematic analysis of 2D scene graph generation is presented, focusing on feature representation and refinement. The long-tail distribution problem and the large intra-class diversity problem are addressed from the perspectives of fusing prior information and commonsense knowledge, as well as refining features through message passing, attention, and visual translation embedding.
3. A review of typical datasets for 2D, spatio-temporal and 3D scene graph generation is presented, along with an analysis of the performance evaluation of the corresponding methods on these datasets.

The rest of this paper is organized as follows; Section 2 gives the definition of a scene graph, thoroughly analyzes the characteristics of visual relationships and the structure of a scene graph. Section 3 surveys scene graph generation methods. Section 4 summarizes almost all currently published datasets. Section 5 compares and discusses the performance of some key methods on the most commonly used datasets. Finally, Section 6 summarizes open problems in the current research and discusses potential future research directions. Section 7 concludes the paper.

2. Scene graph

A scene graph is a structural representation, which can capture detailed semantics by explicitly modeling objects (“man”, “fire hydrant”, “shorts”), attributes of objects (“fire hydrant is yellow”), and relations between paired objects (“man jumping over fire hydrant”), as shown in Fig. 1. The fundamental elements of a scene graph are objects, attributes and relations. Subjects/objects are the core building blocks of an image and they can be located with bounding boxes. Each object can have zero or more attributes, such as color (e.g., yellow), state (e.g., standing), material (e.g., wooden), etc. Relations can be actions (e.g., “jump over”), spatial (e.g., “is behind”), descriptive verbs (e.g., wear), prepositions (e.g. “with”), comparatives (e.g., “taller than”), prepositional phrases (e.g., “drive on”), etc [39–41]. In short, a scene graph is a set of *visual relationship triplets* in the form of $\langle subject, relation, object\rangle$ or $\langle object, is, attribute\rangle$. The latter is also considered as a relationship triplet (using the “is” relation for uniformity [42,43]).

From the point of view of graph theory, a scene graph is a directed graph with three types of nodes: object, attribute, and relation. However, for the convenience of semantic expression, a node of a scene graph is seen as an object with all its attributes, while the relation is called an edge. A subgraph can be formed with an object, which is made up of all the related visual triplets of the object. Therefore, the subgraph contains all the adjacent nodes of the object, and these adjacent nodes directly reflect the context information of the object. From the top-down view, a scene graph can be broken down into several subgraphs, a subgraph can be split into several triplets, and a triplet can be split into individual objects with their attributes and relations. Accordingly, we can find a region in the scene corresponding to the substructure that is a subgraph, a triplet, or an object.

The well-known *knowledge graph* is represented as multi-relational data with enormous fact triplets in the form of (*head entity type*, *relation*, *tail entity type*) [44,45]. Here, we have to emphasize that the visual relationships in a scene graph are different from those in social networks and knowledge bases. In the case of vision, images and visual relationships are incidental and are not intentionally constructed. Especially, visual relationships are usually image-specific because they only depend on the content of the particular image in which they appear. Clearly, a perfectly-generated scene graph corresponding to a given scene should be structurally unique. The process of generating a scene graph should be objective and should only be dependent on the scene. Scene graphs should serve as an objective semantic representation of the state of the scene. The SGG process should not be affected by who labeled the data, on how it was assigned objects and predicate categories, or on the performance of the SGG model used. Although, in reality, not all annotators who label the data produce the exact same visual relationship for each triplet, and the methods that generate scene graphs do not always predict the correct relationships. The uniqueness supports the argument that the use of a scene graph as a replacement for a visual scene at the language level is reasonable.

A two-dimensional (2D) image is a projection of a three-dimensional (3D) world from a particular perspective. Because of the dimension reduction caused by the projection of 3D to 2D, 2D images may have incomplete or ambiguous information about the 3D scene, leading to an imperfect representation of 2D scene graphs. As opposed to a 2D scene graph, a 3D scene graph prevents spatial relationship ambiguities between object pairs caused by different viewpoints. The relationships described above are static and instantaneous because the information is grounded in an image or a 3D mesh that can only capture a specific moment or a certain scene. With videos, a visual relationship is not instantaneous, but varies with time. A digital video consists of a series of images called frames, which means relations span over multiple frames and have different durations. Visual relationships in a video can construct a *Spatio-Temporal Scene Graph*, which includes entity nodes of the neighborhood in the time and space dimensions. Our survey goes beyond 2D scene graphs to include 3D and spatiotemporal scene graphs as well.

3. Scene graph generation

The goal of scene graph generation is to parse an image or a sequence of images in order to generate a structured representation, to bridge the gap between visual and semantic perception, and ultimately to achieve a complete understanding of visual scenes. However, it is difficult to generate an accurate and complete scene graph. Generating a scene graph is generally a bottom-up process in which entities are grouped into triplets and these triplets are connected to form the entire scene graph. Evidently, the essence of the task is to detect the visual relationships, i.e. $\langle subject, relation, object\rangle$ triplets, abbreviated as $\langle s, r, o\rangle$.

Visual Relationship Detection has attracted the attention of the research community since the pioneering work by Lu et al. [41], and the release of the ground-breaking large-scale scene graph dataset Visual Genome (VG) by Krishna et al. [40]. The generation of scene graphs can be currently divided into two types: one is a two-stage method that first detects objects, followed by pairswise relationship recognition; the other is a one-stage method that simultaneously detects and recognizes objects and relations. In two-stage methods, given a visual scene S and its scene graph $\mathcal{T}_S$ [46,47], and when attributes detection and relationships prediction are considered as two independent processes, we can decompose the probability distribution of the scene graph $p(\mathcal{T}_S|S)$ into four components similar to [46]:

$$\begin{aligned} p(\mathcal{T}_S|S) = & p(\mathcal{B}_S|S)p(\mathcal{O}_S|\mathcal{B}_S,S) \\ & (p(\mathcal{A}_S|\mathcal{O}_S,\mathcal{B}_S,S)p(\mathcal{R}_S|\mathcal{O}_S,\mathcal{B}_S,S)) \end{aligned} \tag{1}$$

In Eq. (1), the bounding box component $p(\mathcal{B}_S|S)$ generates a set of candidate regions that cover most of the crucial objects directly from the input image. The object component $p(\mathcal{O}_S|\mathcal{B}_S,S)$ predicts the class label of the object in the bounding box. Both steps are identical to those used in two-stage target detection methods, and can be implemented by the widely used Faster RCNN detector [10]. Conditioned on the predicted labels, the attribute component $p(\mathcal{A}_S|\mathcal{O}_S,\mathcal{B}_S,S)$ infers all possible attributes of each object, while the relationship component $p(\mathcal{R}_S|\mathcal{O}_S,\mathcal{B}_S,S)$ infers the relationship of each object pair [46]. When all visual triplets are collected, a scene graph can then be constructed.

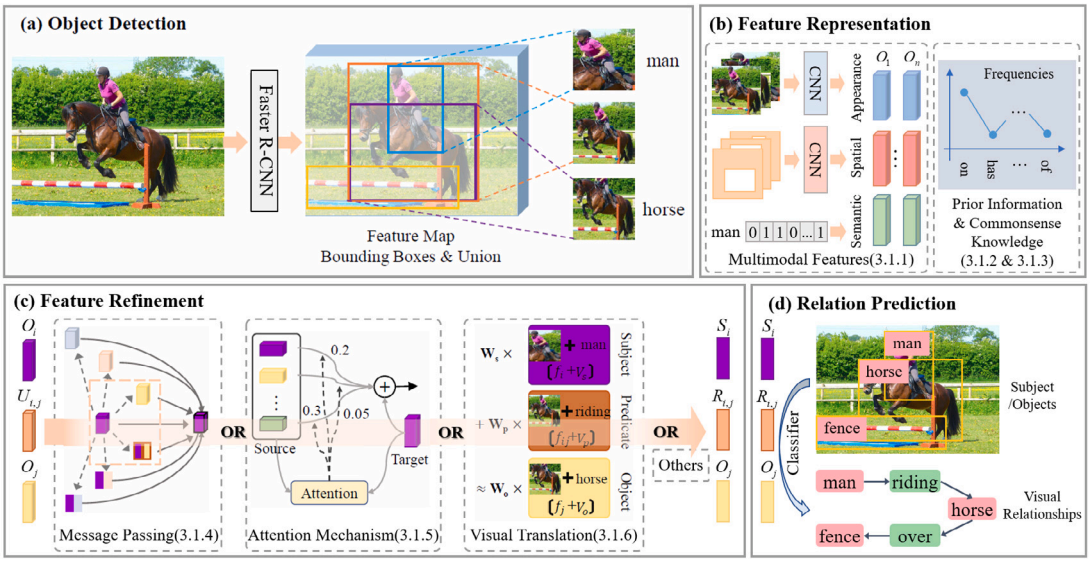


Fig. 2. An overview of 2D general scene graph generation framework. Firstly, off-the-shelf object detectors are used to detect subjects, objects and predicate ROIs. Then, different kinds of methods are used in the stages of **(b) Feature Representation** and **(c) Feature Refinement** to improve the final **(d) Relation Prediction** for high-quality visual relationship detection. This survey focuses on the methods of feature representation and refinement.

Since attribute detection is generally regarded as an independent research topic, visual relationship detection and scene graph generation are often regarded as the same task. Then, the probability of a scene graph $\mathcal{T}_S$ can be decomposed into three factors:

$$p(\mathcal{T}_S|S) = p(\mathcal{B}_S|S)p(\mathcal{O}_S|\mathcal{B}_S, S)p(\mathcal{R}_S|\mathcal{O}_S, \mathcal{B}_S, S) \tag{2}$$

While one-stage methods [48–51] detect and recognize objects and relations at the same time, so their formulas are slightly different:

$$p(\mathcal{T}_S|S) = p(\mathcal{B}_S|S)p(\mathcal{O}_S, \mathcal{R}_S|\mathcal{B}_S, S) \tag{3}$$

where $p(\mathcal{O}_S, \mathcal{R}_S|\mathcal{B}_S, S)$ represents the joint inference model of objects and their relationships based on the object region proposals.

The following section provides a detailed review of more than a hundred deep learning-based methods proposed until 2023 on visual relationship detection and scene graph generation. In view of the fact that 2D SGG has been published much more than 3D or spatio-temporal SGG, a comprehensive overview of the methods for 2D SGG is first provided. This is followed by a review of the 3D and spatiotemporal SGG methods in order to ensure completeness and breadth of the survey.

Note: We use "relationship" or a "triplet" to refer to the tuple of $\langle subject, relation, object\rangle$ in this paper, and "relation" or a "predicate" to refer to a *relation* element.

3.1. 2D scene graph generation

Scene graphs can be generated in two different ways [52]. The mainstream approach uses a two-step pipeline that detects objects first and then solves a classification task to determine the relationship between each pair of objects. The other approach involves jointly inferring the objects and their relationships based on the object region proposals. The former approaches need to first detect all existing objects or proposed objects in the image, and group them into pairs and use the features of their union area (called relation features), as the basic representation for the predicate inference. In the one-stage methods, entities and predicates are often extracted separately from the image in order to reduce the size of relation proposal set.

In this section, we mainly focus on the two-step approach, and Fig. 2 illustrates the general framework for creating 2D scene graphs. Given an image, a scene graph generation method first generates subject/object and union proposals with Region Proposal Network (RPN), which are sometimes derived from the ground-truth human annotations of the image. Each union proposal is made up of a subject, an object and a predicate ROI. The predicate ROI is the box that tightly covers both the subject and the object. We can then obtain appearance, spatial information, label, depth, and mask for each object proposal using the feature representation, and for each predicate proposal we can obtain appearance, spatial, depth, and mask. These multimodal features are vectorized, combined, and refined in the third step of the Feature Refinement module using message passing mechanisms, attention mechanisms and visual translation embedding approaches. Finally, the classifiers are used to predict the categories of the predicates, and the scene graph is generated. In addition, we evaluate a number of recently proposed one-stage scene graph generation methods. These methods directly jointly predict object and their relationships using anchor-free object detection models. When compared to two-stage scene graph generation approaches, they are simple, fast, and easy to train.

In this section, SGG methods for 2D inputs will be reviewed and analyzed according to the following strategies.

(1) Off-the-shelf object detectors can be used to detect subjects, objects and predicate ROIs. The first point to consider is how to utilize the multimodal features of the detected proposals. As a result, Section 3.1.1 reviews and analyzes the use of **multimodal features**, including appearance, spatial, depth, mask, and label.

(2) A scene graph's compositionality is its most important characteristic, and can be seen as an elevation of its semantic expression from independent objects to visual phrases. A deeper meaning can, however, be derived from two aspects: the frequency of visual phrases and the common-sense constraints on relationship prediction. For example, when "man", "horse" and "hat" are detected individually in an image, the most likely visual triplets are $\langle man, ride, horse\rangle$, $\langle man, wearing, hat\rangle$, etc. $\langle hat, on, horse\rangle$ is possible, though not common. But $\langle horse, wearing, hat\rangle$ is normally unreasonable. Thus, how to integrate **Prior Information** about visual phrases and **Commonsense Knowledge** will be the analyzed in Sections 3.1.2 and 3.1.3, respectively.

(3) A scene graph is a representation of visual relationships between objects, and it includes contextual information about those relationships. To achieve high-quality predictions, information must be fused between the individual objects or relationships. In a scene graph, message passing can be used to refine local features and integrate contextual information, while attention mechanisms can be used to allow the

models to focus on the most important parts of the scene. Considering the large intra-class divergence and long-tailed distribution problems, visual translation embedding methods have been proposed to model relationships by interpreting them as translations operating on the low-dimensional embeddings of the entities. Therefore, we categorize the related methods into **Message Passing**, **Attention Mechanism**, and **Visual Translation Embedding**, which will be deeply analyzed in Sections 3.1.4–3.1.6, respectively.

(4) Inspired by one-stage object detection methods, a number of one-stage scene graph generation methods have emerged in recent years. The one-stage approaches directly identify the relationship from the visual features using the CNN-Transformer or fully convolutional network architecture. Due to their small set of proposals, these one-stage frameworks frequently operate effectively. We analyze several **One-Stage Scene Graph Generation Methods** in Section 3.1.7 and compare their results.

3.1.1. Multimodal features

The appearance features of the subject, object, and predicate ROIs make up the input of SGG methods, and affect SGG significantly. The rapid development of deep learning based object detection and classification has led to the use of many types of classical CNNs to extract appearance features from ROIs cropped from a whole image by bounding boxes or masks. Some CNNs even outperform humans when it comes to detecting/classifying objects based on appearance features. Nevertheless, only the appearance features of a subject, an object, and their union region are insufficient to accurately recognize the relationship of a subject–object pair. In addition to appearance features, semantic features of object categories or relations, spatial features of object candidates, and even contextual features, can also be crucial to understand a scene and can be used to improve the visual relationship detection performance. In this subsection, some integrated utilization methods of *Appearance*, *Semantic*, *Spatial* and *Context* features will be reviewed and analyzed.

Appearance-Semantic Features: A straightforward way to fuse semantic features is to concatenate the semantic word embeddings of object labels to the corresponding appearance features. In [41], there is another approach that utilizes language priors from semantic word embeddings to finetune the likelihood of a predicted relationship, dealing with the fact that objects and predicates independently occur frequently, even if relationship triplets are infrequent. Moreover, taking into account that the appearance of objects may profoundly change when they are involved in different visual relations, it is also possible to directly learn an appearance model to recognize richer-level visual composites, i.e., visual phrases [53], as a whole, rather than detecting the basic atoms and then modeling their interactions.

Appearance-Semantic-Spatial Features: The spatial distribution of objects is not only a reflection of their position, but also a representation of their structural information. A spatial distribution of objects is described by the properties of regions, which include positional relations, size relations, distance relations, and shape relations. In this context, Zhu et al. [54] investigated how the spatial distribution of objects can aid in visual relation detection. Sharifzadeh et al. [55] explored the effect of using different object features with a focus on depth maps.

The subject and object come from different distributions, Zhang et al. [56] proposed a 3-branch Relationship Proposal Networks (RelPN) to produce a set of candidate boxes that represent subject, relationship, and object proposals. In another work [43], the authors proposed a new mode that efficiently combines visual, spatial and semantic features and show explicitly what each feature contributes to the final prediction. Liang et al. [14] also considered three types of features and proposed to cascade the multi-cue based CNN with a structural ranking loss function.

Appearance-Semantic-Spatial-Context Features: Previous studies typically extract features from a restricted object–object pair region and focus on local interaction modeling to infer the objects and pairwise relation. For example, by fusing pairwise features, VIP-CNN [13] captures contextual information directly. However, the global visual context beyond these pairwise regions is ignored, it may result in the loss of the chance to shrink the possible semantic space using the rich context. Xu et al. [57] proposed a multi-scale context modeling method that can simultaneously discover and integrate the object-centric and region-centric contexts for inference of scene graphs in order to overcome the problem of large object/relation spaces. Yin et al. [58] proposed a Spatiality-Context-Appearance module to learn the spatiality-aware contextual feature representation.

In summary, appearance, semantics, spatial and contextual features all contribute to visual relationship detection from different perspectives. The integration of these multimodal features precisely corresponds to the human's multi-scale, multi-cue cognitive model. Using well-designed features, visual relationships will be detected more accurately so scene graphs can be constructed more accurately.

3.1.2. Prior information

The scene graph is a semantically structured description of a visual world. Intuitively, the SGG task can be regarded as a two-stage semantic tag retrieval process. Therefore, the determination of the relation category often depends on the labels of the participating subject and object. Although visual relationships are scene-specific, there are strong semantic dependencies between the relationship predicate r and the object categories s and o in a relationship triplet (s, p, o).

Because of the long-tailed distribution of relationships between objects, collecting enough training images for all relationships is time-consuming and too expensive [15,59–61]. Scene graphs should serve as an objective semantic representation of a scene. We cannot arbitrarily assign the relationship of $\langle man, feeding, horse\rangle$ to the scene in Fig. 3(b) just because $\langle man, feeding, horse\rangle$ occurs more frequently than $\langle man, riding, horse\rangle$ in some datasets. However, in fact, weighting the probability output of relationship detection networks by statistical co-occurrences may improve the visual relationship detection performance on some datasets. We cannot deny the fact that human beings sometimes think about the world based on their experiences. As such, prior information, including **Statistical Priors** and **Language Priors**, can be regarded as a type of experience that allows neural networks to "correctly understand" a scene more frequently. Prior information has already been widely used to improve performance of SGG networks.

Statistical Priors: The simplest way to use prior knowledge is to think that an event should happen this time since it always does. This is called statistical prior. Baier et al. [62] demonstrated how a visual statistical model could improve visual relationship detection using absolute frequencies. Dai et al. [63] designed a deep relational network that exploited both spatial configuration and statistical dependency to resolve ambiguities during relationship recognition. Zellers et al. [64] analyzed the statistical co-occurrences between relationships and object pairs on the Visual Genome dataset and concluded that these statistical co-occurrences provided strong regularization for relationship prediction. Amodeo et al. [65] proposed a Ontology-Guided Scene Graph Generation (OG-SGG), which can enhance the performance of an existing machine learning-based scene graph generator by using prior knowledge supplied in the form of an ontology in telepresence robotics scenario.

Furthermore, Chen et al. [46] formally represented this information and explicitly incorporated it into graph propagation networks to aid in scene graph generation.

These methods, however, are data-dependent because their statistical co-occurrence probability is derived from training data. We believe that in the semantic space, language priors will be more useful.

Language Priors: Human communication is primarily based on the use of words in a structured and conventional manner. Similarly, visual relationships are represented as triplets of words. Given the polysemy of words across different contexts, one cannot simply encode

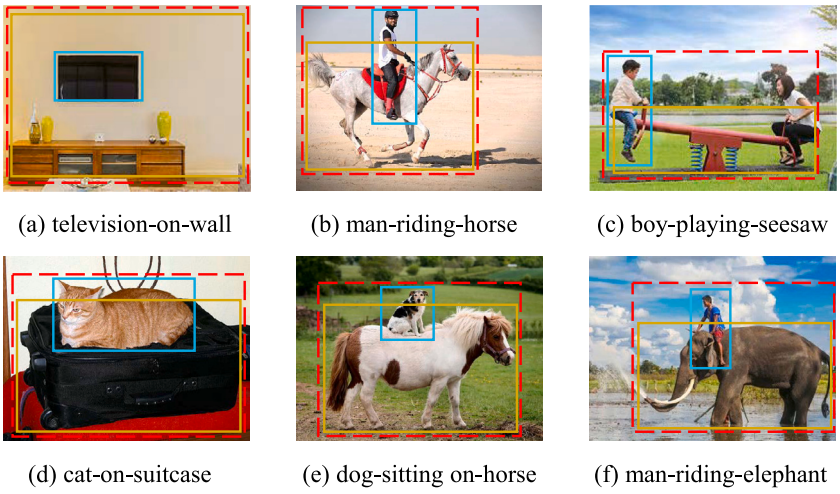


Fig. 3. Examples of the wide variety of visual relationships. Solid bounding boxes indicate individual objects and dash red bounding boxes denote visual relationships.

objects and predicates as indexes or bitmasks. The semantics of object and predicate categories should be used to deal with the polysemy in words. In particular, the following observations can be made. **First**, the visual appearance of the relationships (e.g., the relationships in Fig. 3a and d) which has the same predicate but different agents varies greatly [13]. **Second**, the type of relations between two objects is not only determined by their relative spatial information but also through their categories (such as the relationships in Fig. 3b and e). **Third**, relationships are semantically similar when they appear in similar contexts. That is, in a given context, i.e., an object pair, the probabilities of different predicates to describe this pair are related to their semantic similarity. For example, "person-ride-horse" (Fig. 3b) is similar to "person-ride-elephant" (Fig. 3f), since "horse" and "elephant" belong to the same animal category [41]. It is therefore necessary to explore methods for utilizing language priors in the semantic space.

Lu et al. [41] proposed the first visual relationship detection pipeline, which leverages the language priors (LP) to finetune the prediction. Based on this LP model, Jung et al. [59] further summarized some major difficulties for visual relationship detection and performed a lot of experiments on all possible models with variant modules. Liao et al. [66] assumed that an inherent semantic relationship connects the two words in the triplet rather than a mathematical distance in the embedding space. They proposed to use a generic bi-directional RNN to predict the semantic connection between the participating objects in a relationship from the aspect of natural language. Zhang et al. [15] used semantic associations to compensate for infrequent classes on a large and imbalanced benchmark with an extremely skewed class distribution. Their approach was to learn a visual and a semantic module that maps features from the two modalities into a shared space and then to employ the modified triplet loss to learn the joint visual and semantic embedding. Furthermore, Abdelkarim et al. [67] highlighted the long-tail recognition problem and adopted a weighted version of the softmax triplet loss above.

Some recent work [68–70] try to address one of the challenges of SGG, which is how to handle open-vocabulary. Zhang et al. [68] leverages a powerful pre-trained visual-semantic space (VSS) to achieve language-supervised and open-vocabulary SGG. It contributes a novel method of parsing image language descriptions into semantic graphs, performing region-word alignment in the VSS, and using text prompts for object and relation recognition. He et al. [69] proposes a two-step approach that first pre-trains on a large amount of region-caption data, and then fine-tunes the pre-trained model without updating its parameters using prompt-based techniques. It contributes a new way of supporting inference on completely unseen object classes, which is beyond the capability of existing methods. Zhong et al. [70] presents one of the first methods to learn SGG from image-sentence pairs. It contributes an effective way of leveraging an off-the-shelf object detector to create "pseudo" labels for learning scene graphs, and designing a Transformer-based model that predicts these "pseudo" labels through a masked token prediction task.

From the perspective of collective learning on multi-relational data, Hwang et al. [71] designed an efficient multi-relational tensor factorization algorithm that yields highly informative priors. Analogously, Dupty et al. [60] learned conditional triplet joint distributions in the form of their normalized low rank non-negative tensor decompositions. In addition, some other papers have also tried to mine the value of language prior knowledge for relationship prediction, such as the Logic Tensor Networks (LTNs) [72], the Hierarchy Guided Feature Learning (HGFL) strategy [73], the Variation-structured Reinforcement Learning (VRL) framework [39] and the Rich and Fair semantic extraction network (RiFa) [74].

In summary, statistical and language priors are effective in providing some regularizations, for visual relationship detection, derived from statistical and semantic spaces. However, additional knowledge outside of the scope of object and predicate categories, is not included. Human mind is capable of reasoning over visual elements of an image based on common sense. Thus, incorporating commonsense knowledge into SGG tasks will be valuable to explore.

3.1.3. Commonsense knowledge

As previously stated, there are a number of models which emphasize on the importance of language priors. However, due to the long tail distribution of relationships, it is costly to collect enough training data for all relationships [61]. We should therefore use knowledge beyond the training data to help generate scene graphs [75]. Commonsense knowledge includes information about events that occur in time, about the effects of actions, about physical objects and how they are perceived, and about their properties and relationships with one another. Researchers have proposed to extract commonsense knowledge to refine object and phrase features to improve generalizability of scene graph generation. In this section, we analyze three fundamental sub-issues of commonsense knowledge applied to SGG, i.e., the **Source**, **Formulation** and **Usage**, as illustrated in Fig. 4.

Source: Commonsense knowledge can be directly extracted from the local training samples. For example, the co-occurrence probability can be calculated as the prior statistical knowledge to assist reasoning [76]. However, considering the tremendous valuable information from the large-scale external bases, e.g., Wikipedia and ConceptNet, increasing efforts have been devoted to distill knowledge from these resources.

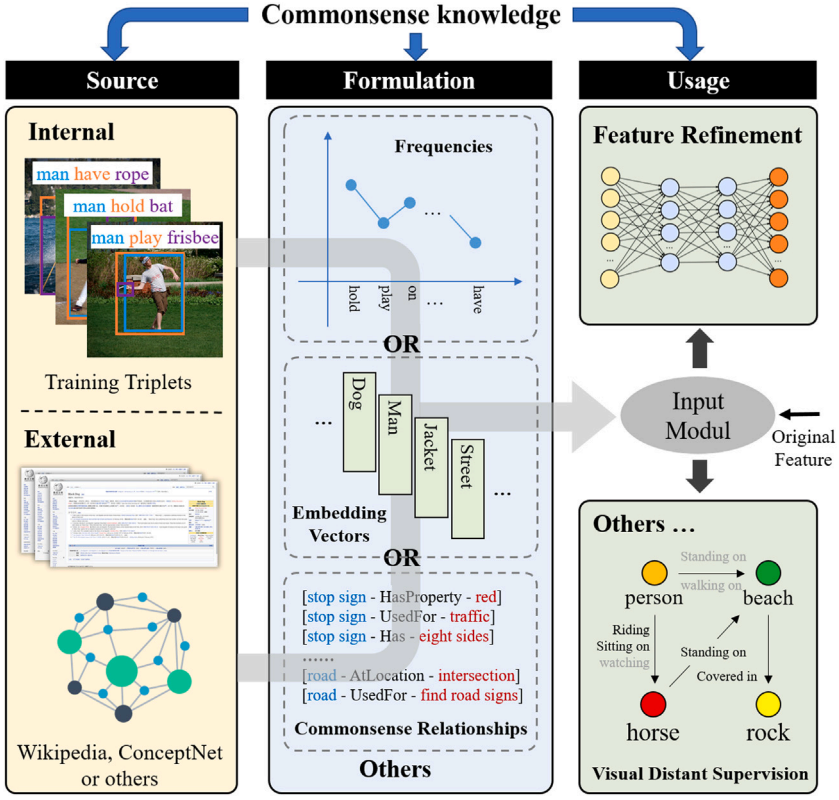


Fig. 4. Three basic sub-issues (source, formulation and usage) of commonsense knowledge applied to scene graph generation.

Gu et al. [77] proposed a knowledge-based module, which improves the feature refinement procedure by reasoning over a basket of commonsense knowledge retrieved from ConceptNet. Yu et al. [61] introduced a Linguistic Knowledge Distillation Framework that obtains linguistic knowledge by mining from both training annotations (internal knowledge) and publicly available text, e.g., Wikipedia (external knowledge). Zhan et al. [78] proposed a novel multi-modal feature based undetermined relationship learning network, in which the linguistic module provides two kinds of features: external linguistic features and internal linguistic features. The former is the semantic representations of subject and object generated by the pretrained word2vec model of Wikipedia 2014.

Formulation: Except from the actual sources of knowledge, it is also important to consider the formulation and how to incorporate the knowledge in a more efficient and comprehensive manner. As shown in several previous studies [46,76], the statistical correlation has been the most common formulation of knowledge. They employ the co-occurrence matrices on both the object pairs and the relationships in an explicit way. Similarly, Linguistic knowledge from [78] is modeled by a conditional probability that encodes the strong correlation between the object pair $\langle subj, obj \rangle$ and the predicate. However, Lin et al. [79] pointed out that they were generally composable, complex and image-specific, and may lead to a poor learning improvement.

Usage: In general, the commonsense knowledge is used as guidance on the original feature refinement for most cases [76,77,79], but there are a lot of attempts to implement it and it contributes from a different aspect on the scene graph model. Yao et al. [80] demonstrated a framework which can train the scene graph models in an unsupervised manner, based on the knowledge bases extracted from the triplets of Web-scale image captions. The relationships from the knowledge base are regarded as the potential relation candidates of corresponding pairs. Inspired by a hierarchical reasoning from the human's prefrontal cortex, Yu et al. [81] built a Cognition Tree (CogTree) for all the relationship categories in a coarse-to-fine manner.

Recently, Zareian et al. [82] proposed a Graph Bridging Network (GB-NET). This model is based on an assumption that a scene graph can be seen as an image-conditioned instantiation of a commonsense knowledge graph. Another work by Zareian et al. [83] points out two specific issues of current researches on knowledge-based SGG methods: **(1)** external source of commonsense tends to be incomplete and inaccurate; **(2)** statistics information such as co-occurrence frequency is limited to reveal the complex, structured patterns of commonsense.

As the main characteristics of commonsense knowledge, external large-scale knowledge bases and specially-designed formulations of statistical correlations have drawn considerable attention in recent years. However, [80,81] have demonstrated that, except from feature refinement, commonsense knowledge can also be useful in different ways. Due to its graph-based structure and enriched information, commonsense knowledge may boost the reasoning process directly. Its graph-based structure makes it very important to guide the message passing on GNN- and GCN-based scene graph generation methods.

3.1.4. Message passing

A scene graph consists not only of individual objects and their relations, but also of contextual information surrounding and forming those visual relationships. From an intuitive perspective, individual predictions of objects and relationships are influenced by their surrounding context. Context can be understood on three levels. **First**, for a triplet, the predictions of different phrase components depend on each other. This is the compositionality of a scene graph. **Second**, the triplets are not isolated. Objects which have relationships are semantically dependent, and relationships which partially share object(s) are also semantically related to one another. **Third**, visual relationships are scene-specific, so learning feature representations from a global view is helpful when predicting relationships. Therefore, message passing between individual objects or triplets are valuable for visual relationship detection.

Constructing a high-quality scene graph relies on a prior layout structure of proposals (objects and unions). There are four forms of

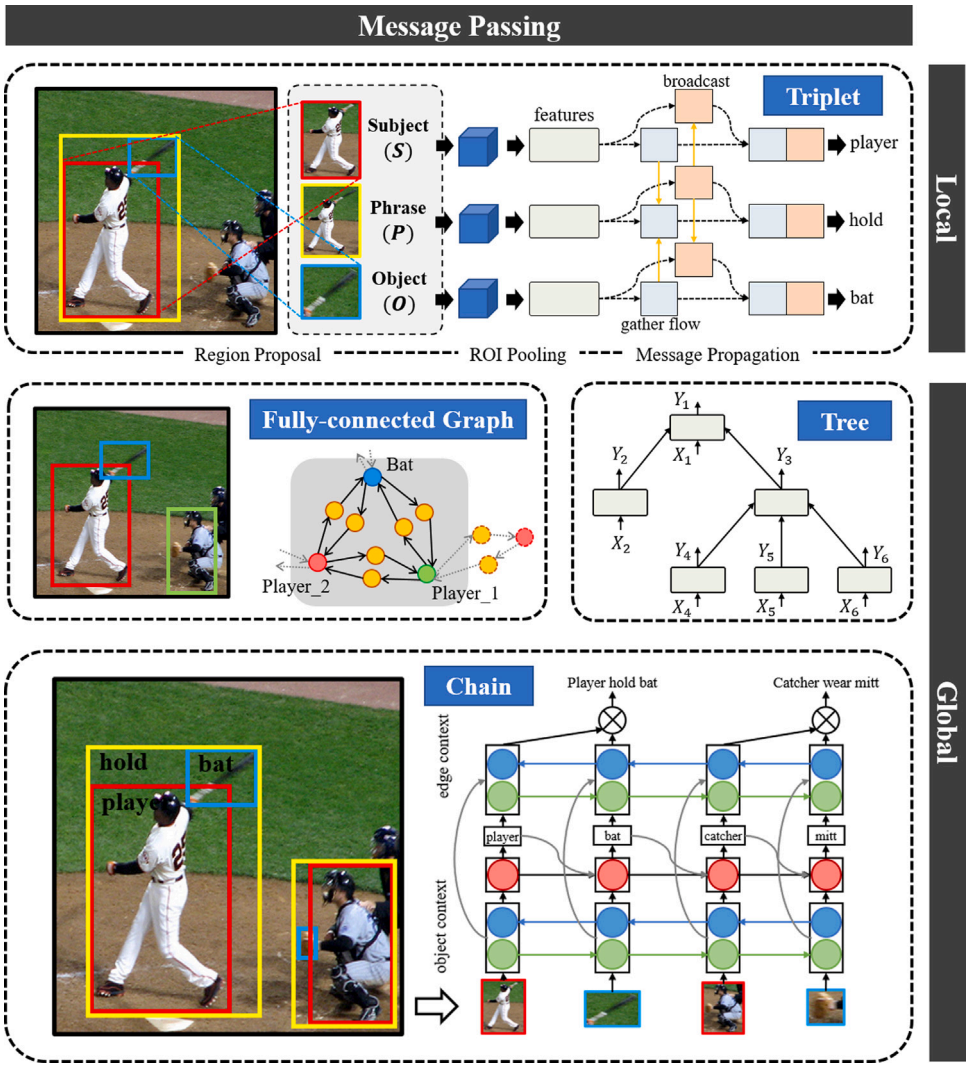


Fig. 5. Examples of two different types of message passing methods, i.e., local propagation within triplet items and global propagation across all the elements.

layout structures: triplet set, chain, tree and fully-connected graph. Accordingly, RNN and its variants (LSTM, GRU) as sequential models are used to encode context for chains while TreeLSTM [84] for trees and GNN (or CRF) [85–87] for fully-connected graphs.

Basically, features and messages are passed between elements of a scene graph, including objects and relationships. To refine object features and extract phrase features, several models rely on a variety of message passing techniques. Our discussion in the subsections below is structured around two key perspectives: **local** propagation within triplet items and **global** propagation across all the elements, as illustrated in Fig. 5.

Local Message Passing Within Triplets: Generally, features of the subject, predicate and object proposals are extracted for each triplet, and the information fusion within the triplets contribute to refine features and recognize visual relationships. ViP-CNN, proposed by Li et al. [13], uses a phrase-guided message passing structure to exchange the information between feature extraction branches of subject, object and predicate. Dai et al. [63] proposed an effective framework called Deep Relational Network (DR-Net), which captures the statistical relations between triplet components and outputs the posterior probabilities of *s*, *r*, and *o*. Another interesting model is ZoomNet [58], in which the Spatiality-Context-Appearance module consists of two spatiality-aware feature alignment cells for message passing between the different components of a triplet.

The local message passing within triplets ignores the surrounding context, while the joint reasoning with contextual information can often resolve ambiguities caused by local predictions made in isolation. The passing of global messages across all elements enhances the ability to detect finer visual relationships.

Global Message Passing Across All Elements: Considering that objects that have visual relationships are semantically related to each other, and that relationships which partially share objects are also semantically related, passing messages between related elements can be beneficial. Learning feature representation from a global view is helpful to scene-specific visual relationship detection. Scene graphs have a particular structure, so message passing on the graph or subgraph structures is a natural choice. Chain-based models (such as RNN or LSTM) can also be used to encode contextual cues due to their ability to represent sequence features. When taking into consideration the inherent parallel/hierarchical relationships between objects, dynamic tree structures can also be used to capture task-specific visual contexts. In the following subsections, message passing methods will be analyzed according to the three categories described below.

Message Passing on Graph Structures. Li et al. [16] developed an end-to-end Multi-level Scene Description Network (MSDN), in which message passing is guided by the dynamic graph constructed from objects and caption region proposals. In the case of a phrase proposal, the message comes from a caption region proposal that may cover multiple object pairs, and may contain contextual information with a larger

scope than a triplet. For comparison, the Context-based Captioning and Scene Graph Generation Network (C2SGNet) [88] also simultaneously generates region captions and scene graphs from input images, but the message passing between phrase and region proposals is unidirectional. Moreover, in an extension of MSDN model, Li et al. [52] proposed a subgraph-based scene graph generation approach called Factorizable Network (F-Net). F-Net clusters the fully-connected graph into several subgraphs and uses a Spatial-weighted Message Passing structure for feature refinement.

Even though MSDN and F-Net extended the scope of message passing, a subgraph is considered as a whole when sending and receiving messages. Liao et al. [89] proposed semantics guided graph relation neural network (SGRNN), in which the target and source must be an object or a predicate within a subgraph. The scope of messaging is the same as Feature Inter-refinement of objects and relations in [77].

Several other techniques consider SGG as a graph inference process because of its particular structure. By considering all other objects as carriers of global contextual information for each object, they will pass messages to each other's via a fully-connected graph. However, inference on a densely connected graph is very expensive. As shown in previous works [90,91], dense graph inference can be approximated by **mean field** in Conditional Random Fields (CRF). Zheng et al. [92,93] combines the strengths of CNNs with CRFs, and formulates mean-field inference as Recurrent Neural Networks (RNN). Therefore, it is reasonable to use CRF or RNN to formulate a scene graph generation problem [63,94].

Further, there are some other relevant works which proposed modeling methods based on a pre-determined graph [46,82,95–98].

Message Passing on Chain Structures. Dense graph inference can be approximated by mean fields in CRF, and it can also be dealt with using an RNN-based model. Xu et al. [99] generated a structured scene representation from an image, and solved the graph inference problem using GRUs to iteratively improve its predictions via message passing. This work is considered as a milestone in scene graph generation, demonstrating that RNN-based models can be used to encode the contextual cues for visual relationship recognition. At this point, Zellers et al. [64] presented a novel model, Stacked Motif Network (MOTIFNET), which uses LSTMs to create a contextualized representation of each object. Dhingra et al. [100] proposed an object communication module based on a bi-directional GRU layer. The Counterfactual critic Multi-Agent Training (CMAT) approach [101] is another important extension where each agent communicates with the others for T rounds to encode the visual context using LSTM.

Many other message passing methods based on RNN have also been developed. RNN and bi-directional RNN/LSTM models [88,102–104] are used to capture context.

Message Passing on Tree Structures. As previously stated, graph and chain structures are widely used for message passing. However, these two structures are sub-optimal. Chains are oversimplified and may only capture simple spatial information or co-occurring information. Even though fully-connected graphs are complete, they do not distinguish between hierarchical relations. Tang et al. [105] constructed a dynamic tree structure, dubbed VCTREE, that places objects into a visual context, and then adopted bidirectional TreeLSTM to encode the visual contexts. VCTREE has several advantages over chains and graphs, such as hierarchy, dynamicity, and efficiency. The ways of context encoding and decoding for objects and predicates are similar to [64], but they replace LSTM with TreeLSTM. In [64], Zellers et al. tried several ways to order the bounding regions in their analysis. Here, we can see the tree structure in VCTREE as another way to order the bounding regions.

3.1.5. Attention mechanisms

Attention mechanisms flourished soon after the success of Recurrent Attention Model (RAM) [106] for image classification. They enable models to focus on the most significant parts of the input [107]. With scene graph generation, as with iterative message passing models, there are two objectives: refine local features and fuse contextual information. On the basic framework shown in Fig. 2, attention mechanisms can be used both at the stage of feature representation and at the stage of feature refinement. At the feature representation stage, attention can be used in the spatial domain, channel domain or their mixed domain to produce a more precise appearance representation of object regions and unions of object-pairs. At the feature refinement stage, attention is used to update each object and relationship representation by integrating contextual information. Therefore, this section will analyze two types of attention mechanisms for SGG (as illustrated in Fig. 6), namely, **Self-Attention** and **Context-Aware Attention** mechanisms.

Self-Attention Mechanisms. Self-Attention mechanism aggregates multimodal features of an object to generate a more comprehensive representation. Zheng et al. [108] proposed a multi-level attention visual relation detection model (MLA-VRD), which uses multi-stage attention for appearance feature extraction and multi-cue attention for feature fusion. In another work, Zhou et al. [96] combined multi-stage and multi-cue attention to guide the generation of more efficient attention maps. Zhuang et al. [109] proposed a context-aware model, which applies an attention-pooling layer to the activations of the *conv5_3* layer of VGG-16 as an appearance feature representation of the union region. Han et al. [110] argued that the context-aware model pays less attention to small-scale objects. Therefore, they proposed a two-dimensional normal distribution attention scheme to effectively model small objects. Kolesnikov et al. [111] proposed the box attention and incorporated box attention maps in convolutional layers of the base detection model. Wang [112] incorporates the recurrent attention method into the detection pipeline, allowing the network to concentrate on a number of particular regions of an image when calculating the predicates for a given item pair.

Context-Aware Attention Mechanisms. Context-Aware Attention learns the contextual features using graph parsing. Yang et al. [113] proposed Graph R-CNN based on GCN [85], which can be factorized into three logical stages: **(1)** produce a set of localized object regions, **(2)** utilize a relation proposal network (RePN) to learn to efficiently compute relatedness scores between object pairs, and **(3)** apply an attentional graph convolution network (aGCN) to propagate a higher-order context throughout the sparse graph. From the derivation, it can be seen that the aGCN is similar to Graph Attention Network (GAT) [87].

Qi et al. [47] also leveraged a graph self-attention module to embed entities, and used the multi-layer perceptron (MLP) to learn to efficiently estimate the *relatedness* of an object pair. Lin et al. [114] designed a direction-aware message passing (DMP) module based on GAT to enhances the node feature with node-specific contextual information. Moreover, Zhang et al. [115] used context-aware attention mechanism directly on the fully-connected graph to refine object region feature. Dornadula et al. [97] introduced another interesting GCN-based attention model, which treats predicates as learned semantic and spatial functions.

3.1.6. Visual translation embedding

The **long-tail** problem heavily affects the scalability and generalization ability of learned models. Another problem is the **large intra-class divergence** [116], i.e., relationships that have the same predicate but from which different subjects or objects are essentially different. Therefore, there are two challenges for visual relationship detection models. **First**, is the right representation of visual relations to **handle the large variety in their appearance**, which depends on the involved entities. **Second**, is to handle the scarcity of training data for zero-shot visual relation triplets. Visual embedding approaches aim at learning

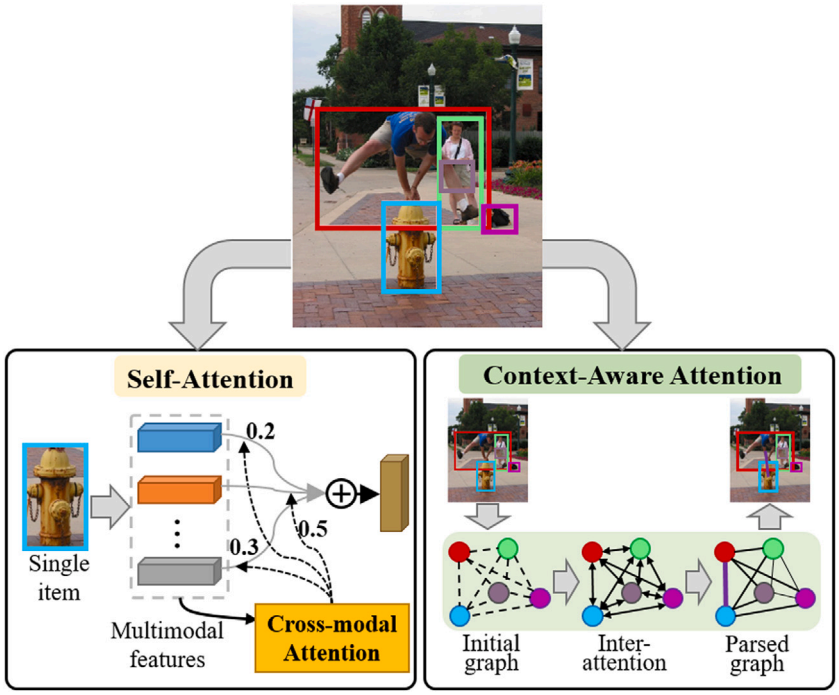


Fig. 6. Two kinds of attention mechanisms in SGG.

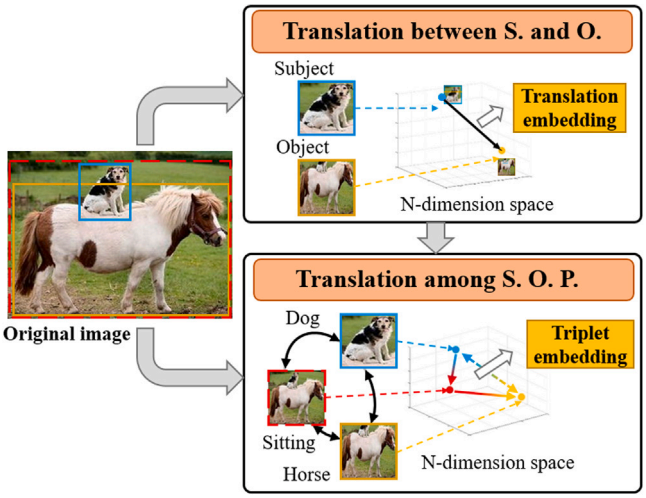


Fig. 7. Two types of visual translation embedding approaches according to whether to embed the predicate into N-dimensional space [117] or not [118].

a compositional representation for subject, object and predicate by learning separate visual-language embedding spaces, where each of these entities is mapped close to the language embedding of its associated annotation. By constructing a mathematical relationship of visual-semantic embeddings for subject, predicate and object, an end-to-end architecture can be built and trained to learn a visual translation vector for prediction. In this section, we divide the visual translation embedding methods according to the translations (as illustrated in Fig. 7), including **Translation between Subject and Object**, and **Translation among Subject, Object and Predicate**.

Translation Embedding between Subject and Object: Translation-based models in knowledge graphs are good at learning embeddings, while also preserving the structural information of the graph [119–121]. Inspired by Translation Embedding (TransE) [119] to represent large-scale knowledge bases, Zhang et al. [118] proposed a Visual Translation Embedding network (VTransE) which places objects in a low-dimensional relation space, where a relationship can be modeled as a simple vector translation, i.e., subject+predicate ≈ object.

Translation Embedding between Subject, Object and Predicate: In an extension of VTransE, Hung et al. [122] proposed the Union Visual Translation Embedding network (UVTransE), which learns three projection matrices $\boldsymbol{W}_s$, $\boldsymbol{W}_o$, $\boldsymbol{W}_u$ which map the respective feature vectors of the bounding boxes enclosing the *subject*, *object*, and *union* of subject and object into a common embedding space, as well as translation vectors t_p(to be consistent with VTransE) in the same space corresponding to each of the predicate labels that are present in the dataset. Another extension is ATR-Net, proposed by Gkanatsios et al. [117], which projects the visual features from the subject, object region and their union into a score space as S, O and P with multi-head language and spatial attention. The Multimodal Attentional Translation Embeddings (MATransE) model built upon VTransE [123] learns a projection of $\langle S, P, O\rangle$ into a score space where $S + P \approx O$. Subsequently, Qi et al. [47] introduced a semantic transformation module into their network structure to represent $\langle S, P, O\rangle$ in the semantic domain. Another TransE-inspired model is RLSV (Representation Learning via Jointly Structural and Visual Embedding) [45]. The architecture of RLSV is a three-layered hierarchical projection that projects a visual triple onto the attribute space, the relation space, and the visual space in order.

In summary, while many of the above 2D SGG models use more than one method, we selected the method we felt best reflected the idea of the paper for our primary classification of methods. The aforementioned 2D SGG challenges can also be addressed in other ways utilizing different concepts. As an example, Knyazev et al. [124] used Generative Adversarial Networks (GANs) to synthesize rare yet plausible scene graphs to overcome the long-tailed distribution problem. Huang et al. [125] designed a Contrasting Cross-Entropy loss and a scoring module to address class imbalance. Suhail et al. [126] introduced a novel energy-based learning framework for generating scene graphs as an alternative to cross-entropy, which can effectively utilize structural information in the output space. Wang et al. [112] employs undersampling and negative sampling strategies to lessen the effects of an imbalanced data distribution. Tang et al. [127] propose a new SGG

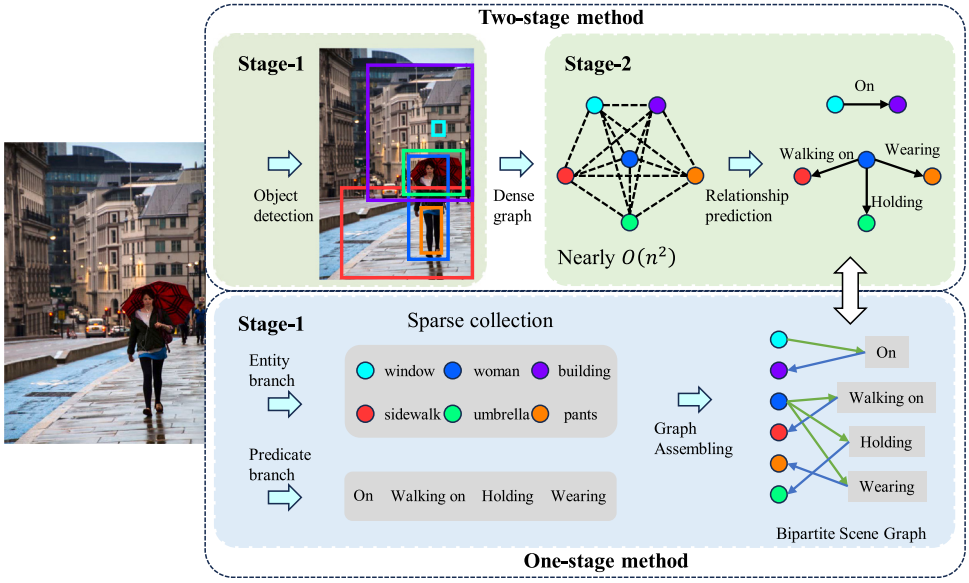


Fig. 8. Comparison of two-stage scene graph generation methods and one-stage scene graph generation methods.

task with the goal of generating unbiased scene graphs from biased training data. And proposed an unbiased prediction method based on causal inference, called Total Direct Effect (TDE). The idea of TDE is to use counterfactual thinking to distinguish between main effects and side effects in relation prediction, which affects the subsequent SGG method. Fukuzawa and Toshiyuki [128] introduced a pioneering approach to visual relationship detection by reducing it to an object detection problem, and they won the Google AI Open Images V4 Visual Relationship Track Challenge. A neural tensor network was proposed by Qiang et al. [129] for predicting visual relationships in an image. These methods contribute to the 2D SGG field in their own ways.

3.1.7. One-stage scene graph generation methods

Scene graph generation (SGG) is a complex visual understanding task that aims to detect objects and their relations in images and represent them with graph structures. Traditional SGG methods usually adopt a multi-stage pipeline that decomposes this task into sub-tasks such as object detection, relation graph construction, and relation prediction. However, these methods often require a large number of parameters and computational resources, and are prone to be affected by the performance of object detection. In recent years, some researchers have proposed one-stage SGG methods [48–51] that try to directly predict scene graphs from images, improving efficiency and accuracy. These methods can be divided into two or three groups according to whether they use transformer, a powerful self-attention mechanism.

RelTR [49] and SGTR [51] both treat SGG as a set prediction problem and use transformer to encode and decode the information of objects and relations. Their main differences are: RelTR [49] is a bottom-up method that first uses a transformer-based detector (such as DETR) to generate object candidates, then uses a relation transformer to predict the relations between object pairs. It also designs an integral-based relation representation method that encodes relations as two-dimensional vector fields. SGTR [51] is a top-down method that first uses a transformer-based generator to generate a set of learnable triplet queries (subject–predicate–object), then uses a cascaded triplet detector to progressively refine these queries and generate the final scene graph. It also proposes an entity-aware relation representation method based on structured generators that exploits the compositional property of relations.

Liu et al. proposes a fully convolutional scene graph generation (FCSGG) [48] model that simultaneously detects objects and relations. Unlike traditional detectors based on region proposal network (RPN) or anchors, FCSGG is a bottom-up method that encodes objects as bounding box center points, encodes relations as two-dimensional vector fields (called relation affinity fields), and uses spatial transformer network (STN) to extract features. FCSGG only uses visual features to generate scene graphs, without requiring additional semantic knowledge.

Teng et al. proposes a structured sparse r-cnn (SSR-CNN) [50] model that treats SGG as a bidirectional graph construction problem. SSR-CNN is a top-down method that first uses a set of learnable triplet queries to generate entity and predicate candidate sets, then uses a graph assembly module to infer the connectivity between nodes in the bidirectional scene graph. SSR-CNN also designs a relaxed and enhanced training strategy based on knowledge distillation (KD) to alleviate the training difficulty.

Since the overall structure of the one-stage model and the two-stage model are different, most of the prior knowledge or common sense knowledge methods used in the two-stage model cannot be directly applied to the one-stage model. However, Li et al. proposes a novel method [130] to leverage prior knowledge for both two-stage SGG and one-stage SGG. The prior knowledge is the label semantic distribution (LSD), which reflects the correlations between a subject–object instance and multiple predicate categories. [130] uses LSD to generate soft labels for each subject–object instance, which are more informative and less biased than the original one-hot target labels.

To summarize, these four papers are all one-stage SGG methods. They all try to directly generate scene graphs from images, without requiring prior object detection and relation graph construction. They all use the idea of set prediction, using learnable queries to encode the prior information of objects and relations, and matching them with true labels by optimizing loss functions. However, they differ in the network structures, relation representation methods, training strategies, etc. Transformer-based methods can better capture the long-distance dependencies between objects and relations, but also require more computational resources. FCN-based methods can more efficiently exploit spatial information, but may also ignore some fine-grained features. Sparse R-CNN-based methods can more flexibly adapt to different numbers and scales of objects and relations, but also need more complex training strategies.

The differences between two-stage and one-stage scene graph generation methods are show in Fig. 8 and can be compared from the

following aspects: Generation process: Two-stage methods first generate entity proposals, then predict predicates; one-stage methods directly generate relation candidate sets. Computation: Two-stage methods need to predict $O(n^2)$ relation triplets, where n is the number of entity proposals, which is computationally expensive; one-stage methods only need to predict a sparse relation candidate set, which is computationally efficient. Inference speed: Two-stage methods need to perform multiple post-processing steps, which is slow; one-stage methods can directly construct scene graphs, which is fast. Error propagation: Two-stage methods may be affected by the errors of the object detector, leading to error propagation; one-stage methods can avoid this problem, improving robustness. Model structure: Two-stage methods usually use simple classifiers to predict predicates, which is simple; one-stage methods usually use Transformer or other attention mechanisms to learn the queries of entities and predicates, which is complex.

3.2. Spatio-temporal scene graph generation

Recently, with the development of relationship detection models in the context of still images (ImgVRD), some researchers have started to pay attention to understand visual relationships in videos (VidVRD). Compared to images, videos provide a more natural set of features for detecting visual relations. The visual relations in a video are usually changeable over time.

Different from static images and because of the additional temporal channel, dynamic relationships in videos are often correlated in both the spatial and temporal dimensions. All the relationships in a video can collectively form a spatial–temporal graph structure, as mentioned in [131–136]. Therefore, we redefine the VidVRD as **Spatio-Temporal Scene Graph Generation (ST-SGG)**. ST-SGG relies on video object detection (VOD). The mainstream methods address VOD by integrating the latest techniques in both image-based object detection and multi-object tracking [137–139]. Although recent sophisticated deep neural networks have achieved superior performances in image object detection [10,11,140,141], object detection in videos [142] still suffers from a low accuracy. Inevitably, these problems have gone down to downstream video relationship detection and even are amplified.

Shang et al. [143] first proposed VidVRD task and introduced a basic pipeline solution, which adopts a bottom-up strategy. They firstly split videos into segments with a fixed duration and predict visual relations between co-occurrent short-term object tracklets for each video segment. Then they generate complete relation instances by a greedy associating procedure. Tsai et al. [132] proposed a Gated Spatio-Temporal Energy Graph (GSTEG) that models the spatial and temporal structure of relationship entities in a video by a spatial–temporal fully-connected graph. Shang et al. [144] has published another dataset, VidOR and launched the ACM MM 2019 Video Relation Understanding (VRU) Challenge [145]. In this challenge, Zheng et al. [146] use Deep Structural Ranking (DSR) [14] model to predict relations. Different from the pipeline in [143], they associate the short-term preliminary trajectories before relation prediction. Similarly, Sun et al. [147] also associate the preliminary trajectories on the front by applying a kernelized correlation filter (KCF) tracker.

Teng et al. [148] introduce a novel detect-to-track paradigm for VidSGG, which decouples the context modeling from the low-level entity tracking for relation prediction. The paradigm, named Target Adaptive Context Aggregation Network (TRACE), emphasizes on capturing spatio-temporal context information for relation recognition.

Cong et al. [149] introduce a neural network model called Spatial–temporal Transformer (STTran), which has two core modules for video understanding: a spatial encoder that extracts spatial context from an input frame and reasons about the visual relationships within the frame, and a temporal decoder that takes the output of the spatial encoder as input and captures the temporal dependencies between frames and infers the dynamic relationships.

3.3. 3D scene graph generation

In the computer vision field, one of the most important branches of 3D research is the representation of 3D information. To extend the concept of scene graph to 3D space, researchers are trying to design a structured text representation to encode 3D information. Although existing scene graph research concentrates on 2D static scenes, based on these findings as well as on the development of 3D object detection [150–154] and 3D Semantic Scene Segmentation [155–157], scene graphs in 3D have recently started to gain more popularity [28,158–162].

Stuart et al. [158] were the first to introduce the term of "3D scene graph" and defined the problem and the model related to the prediction of 3D scene graph representations across multiple views. However, there is no essential difference in structure between their 3D scene graph and a 2D scene graph. Zhang et al. [163] started with the *cardinal direction relations* and analyzed support relations between a group of connected objects grounded in a set of RGB-D images about the same static scene from different views. Kim et al. [159] proposed a 3D scene graph model for robotics. Johanna et al. [160] tried to understand indoor reconstructions by constructing 3D semantic scene graph. None of these works have proposed an ideal way to model the 3D space and multi-level semantics.

Until now, there is no unified definition and representation of 3D scene graph. However, as an extension of the 2D scene graph in 3D spaces, 3D scene graph should be designed as a simple structure which encodes the relevant semantics within environments in an accurate, applicable, usable, and scalable way, such as object categories and relations between objects as well as physical attributes. It is noteworthy that, Armeni et al. [28] creatively proposed a novel 3D scene graph model, which performs a hierarchical mapping of 3D models of large spaces. Recently, Rosinolet al. [164] defined 3D Dynamic Scene Graphs as a unified representation for actionable spatial perception.

In summary, this section includes a comprehensive overview of 2D SGG, followed by reviews of ST-SGG and 3D SGG. Researchers have contributed to the SGG field and will continue to do so, but the long-tail problem and the large intra-class diversity problem will remain hot issues, motivating researchers to explore more models to generate more useful scene graphs.

4. Datasets

In this section, we provide a summary of some of the most widely used datasets for visual relationship and scene graph generation. These datasets are grouped into three categories—2D images, videos and 3D representation.

4.1. 2D datasets

Several 2D image datasets are available and their statistics are summarized in Table 1. The following are some of the most popular ones:

Visual Phrase [53] is on visual phrase recognition and detection. The dataset contains 8 object categories from Pascal VOC2008 [170] and 17 visual phrases that are formed by either an interaction between objects or activities of single objects.

Scene Graph [26] is the first dataset of real-world scene graphs. The full dataset consists of 5000 images selected from the intersection of the YFCC100m [171] and Microsoft COCO [172] datasets and each of which has a human-generated scene graph.

Visual Relationship Detection (VRD) [41] dataset intends to benchmark the scene graph generation task. It highlights the long-tailed distribution of infrequent relationships. The public benchmark based on this dataset uses 4000 images for training and test on the remaining 1000 images. The relations broadly fit into categories, such as action, verbal, spatial, preposition and comparative.

Table 1
The statistics of common 2D datasets.

Dataset	Object	Bbox	Relationship	Triplet	Image	Source Link
RW-SGD [26]	6745	93,823	1310	112,707	5000	https://github.com/google-research-datasets/dstc8-schema-guided-dialogue
HCVRD [165]	1824	–	28,323	256,550	52,855	https://github.com/bohanzhuang/HCVRD-a-benchmark-for-large-scale-Human-Centered-Visual-Relationship-Detection
Visual phrase [53]	8	3,71	9	1796	2769	http://vision.cs.uiuc.edu/phrasal/
Scene graph [26]	266	69,009	68	109,535	5000	http://imagenet.stanford.edu/internal/jcjohns/scene_graphs/sg_dataset.zip
VRD [41]	100	–	70	37,993	5000	https://cs.stanford.edu/people/ranjaykrishna/vrd/
Open Images v4 [42]	57	3,290,070	329	374,768	9,178,275	https://storage.googleapis.com/openimages/web/index.html
Visual Genome [40]	33,877	3,843,636	40,480	2347,187	108,077	http://visualgenome.org/
VrR-VG [166]	1600	282,460	117	203,375	58,983	http://vrr-vg.com/
UnRel [167]	–	–	18	76	1071	https://www.di.ens.fr/willow/research/unrel/
SpatialSense [168]	3679	–	9	13,229	11,569	https://github.com/princeton-vl/SpatialSense
SpatialVOC2K [169]	20	5775	34	9804	2026	https://github.com/muskata/SpatialVOC2K

Visual Genome [40] has the maximum number of relation triplets with the most diverse object categories and relation labels up to now. VG is annotated by crowd workers and thus a substantial fraction of the object annotations has poor quality and overlapping bounding boxes and/or ambiguous object names. Prior works have explored semi-automatic ways to clean up object and relation annotations and constructed their own VG versions. Of these, VG200 [118], VG150 [99], VG-MSDN [16] and sVG [63] have released their cleansed annotations. Other works [13,39,54,58,61,94,96,173–175] use a paper-specific and nonpublicly available split.

VG150 [99] is constructed by pre-processing VG to improve the quality of object annotations. On average, this annotation refinement process has corrected 22 bounding boxes and/or names, deleted 7.4 boxes, and merged 5.4 duplicate bounding boxes per image. The benchmark uses the most frequent 150 object categories and 50 predicates for evaluation.

VrR-VG [166] is also based on Visual Genome. Its pre-processing aims at reducing the duplicate relationships by hierarchical clustering and filtering out the visually-irrelevant relationships. As a result, the dataset keeps the top 1600 objects and 117 visually-relevant relationships of Visual Genome.

Open Images [42] is a dataset of 9M images annotated with image-level labels, object bounding boxes, object segmentation masks, visual relationships, and localized narratives. The images are very diverse and often contain complex scenes with several objects (8.3 per image on average).

UnRel [167] is a challenging dataset that contains 1000 images collected from the web with 76 unusual language triplet queries such as "person ride giraffe". All images are annotated at box-level for the given triplet queries. It is often used to evaluate the generalization performance of the algorithm.

GQA dataset [176] consists of 22,669,678 questions in 113,018 images, covering a wide range of reasoning skills. The dataset has a vocabulary of 3097 words and 1878 possible answers, and it covers 88.8% and 70.6% of VQA questions and answers, respectively, with a wider diversity.

SpatialSense [168] is a dataset specializing in spatial relation recognition. A key feature of the dataset is that it is constructed through *adversarial crowdsourcing*: a human annotator is asked to come up with adversarial examples to confuse a recognition system.

SpatialVOC2K [169] is the first multilingual image dataset with spatial relation annotations and object features for image-to-text generation. For each image, they provided additional annotations for each ordered object pair, i.e., (a) *the single best*, and (b) *all possible* prepositions that correctly describe the spatial relationship between objects.

4.2. Video datasets

So far there are several public datasets for video relational understanding.

ImageNet-VidVRD [143] is the first video visual relation detection dataset, which is constructed by selecting 1000 videos from the training set and the validation set of ILSVRC2016-VID [177]. Based on the 1000 videos, the object categories increase to 35. It contains a total of 3219 relationship triplets (*i.e.,* the number of visual relation types) with 132 predicate categories. All videos were decomposed into segments of 30 frames with 15 overlapping frames in advance, and all the predicates appearing in each segment were labeled to obtain segment-level visual relation instances.

VidOR [144] consists of 10,000 user-generated videos (98.6 h) together with dense annotations on 80 categories of objects and 50 categories of predicates. Specifically, objects are annotated with a bounding-box trajectory to indicate their spatio-temporal locations in the videos; and relationships are temporally annotated with start and end frames. The videos were selected from YFCC-100M multimedia collection and the average length of the videos is about 35 s.

Action Genome [133] provides frame-level scene graph labels for the components of each action. Overall, Action Genome provide annotations for 234,253 frames with a total of 476,229 bounding boxes of 35 object classes (excluding "person"), and 1,715,568 instances of 25 relationship classes.

Home Action Genome [178] is a large-scale multi-view video dataset that captures indoor daily activities from different perspectives, including an egocentric one. The dataset consists of 25.4 h of videos, covering 75 classes of daily activities and 453 classes of atomic actions, in 1752 synchronized sequences and 5700 videos in total. For scene graphs, one third-person view video in each synchronized sequence is annotated with bounding boxes of the subject and the object, as well as the relationship between them. The dataset has 86 object classes and 29 relationship classes.

PVSG [135] consists of 400 videos among which 289 are third-person videos and 111 are egocentric videos. Each video contains an average length of 76.5 s. In total, 152,958 frames are labeled with fine panoptic segmentation and temporal scene graphs. There are 126 object classes and 57 relation classes.

4.3. 3D datasets

Recently, several 3D datasets related to scene graphs have been released to satisfy the needs of SGG study.

3D Scene Graph is constructed by annotated the Gibson Environment Database [179] using the automated 3D Scene Graph generation pipeline proposed in [28]. Gibson's underlying database of spaces includes 572 full buildings composed of 1447 floors covering a total area of 211 km^2. It is collected from real indoor spaces using 3D scanning and reconstruction and provides the corresponding 3D mesh model of each building. Meanwhile, for each space, the RGB images, depth and surface normals are provided. A fraction of the spaces is annotated with semantic objects.

3DSGG, proposed in [159], is a large scale 3D dataset that extends 3RScan with semantic scene graph annotations, containing relationships, attributes and class hierarchies. A scene graph here is a set of tuples (N, R) between nodes N and edges R. Each node is defined by a hierarchy of classes $c = (c_1, \dots, c_d)$ and a set of attributes A that describe the visual and physical appearance of the object instance. The edges define the semantic relations between the nodes. This representation shows that a 3D scene graph can easily be rendered to 2D.

5. Performance evaluation

In this section, we first introduce some commonly used evaluation modes and criteria for the scene graph generation task. Then, we provide the quantitative performance of the promising models on popular datasets. Since there is no uniform definition of a 3D scene graph, we will introduce these contents around 2D scene graph and spatio-temporal scene graph.

5.1. Tasks

Most prior works often evaluated their SGG models on several of the following common sub-tasks. We preserve the names of tasks as defined in [41,99] here, despite the inconsistent terms used in other papers and the inconsistencies on whether they are in fact classification or detection tasks.

1. **Phrase Detection (PhrDet)** [41]: Outputs a label *subject–predicate–object* and localizes the entire relationship in one bounding box with at least 0.5 overlap with the ground truth box. It is also called **Union boxes detection** in [63].
2. **Predicate Classification (PredCls)** [99]: Given a set of localized objects with category labels, decide which pairs interact and classify each pair's predicate.
3. **Scene Graph Classification (SGCls)** [99]: Given a set of localized objects, predict the predicate as well as the object categories of the subject and the object in every pairwise relationship.
4. **Scene Graph Generation (SGGen)** [99]: Detect a set of objects and predict the predicate between each pair of the detected objects. This task is also called **Relationship Detection (RelDet)** in [41] or **Two boxes detection** in [63]. It is similar to phrase detection, but with the difference that both the bounding box of the *subject* and *object* need at least 50 percent of overlap with their ground truth. Since SGGen only scores a single complete triplet, the result cannot reflect the detection effects of each component in the whole scene graph. So Yang et al. [113] proposed the **Comprehensive Scene Graph Generation (SGGen+)** as an augmentation of SGGen. SGGen+ not only considers the triplets in the graph, but also the singletons (object and predicate). To be clear, SGGen+ is essentially a metric rather than a task.
5. **Panoptic Scene Graph Generation (PSG)** [180]: PSG is a novel scene understanding task that requires the model to generate a more comprehensive scene graph representation based on the panoptic segmentation of the image. Panoptic segmentation is a fine-grained object localization method that uses pixel-level masks to represent all the contents in the image, including "things" and "stuff". Scene graph is a structured representation that uses nodes and edges to describe the objects and relations in the image, which can bridge human language and visual scene. The goal of PSG task is to improve the quality and coverage of scene understanding, and address some of the issues of the conventional Scene Graph Generation (SGG) methods, such as the imprecision and redundancy of bounding boxes.

There are also some paper-specific task settings including **Triple Detection** [62], **Relation Retrieval** [118] and so on.

In the video based visual relationship detection task, there are two standard evaluation modes: **Relation Detection** and **Relation Tagging**. The detection task aims to generate a set of relationship triplets with tracklet proposals from a given video, while the tagging task only considers the accuracy of the predicted video relation triplets and ignores the object localization results.

5.2. Metrics

Recall@K. The conventional metric for the evaluation of SGG is the image-level $\boldsymbol{Recall@K}$($\boldsymbol{R@K}$), which computes the fraction of times the correct relationship is predicted in the top K confident relationship predictions. Some methods compute R@K with the constraint that merely one relationship can be obtained for a given object pair. There is a superparameter k, often not clearly stated in some works, which measures the maximum predictions allowed per object pair. Most works have seen PhrDet as a multiclass problem and they use $k = 1$ to reward the correct top-1 prediction for each pair. While other works [66,108,181] tackle this as a multilabel problem and they use a k equal to the number of predicate classes to allow for predicate co-occurrences [117]. Some works [61,64,117,182,183] have also identified this inconsistency and interpret it as whether there is graph constraint. The unconstrained metric (i.e., no graph constraint) evaluates models more reliably, since it does not require a perfect triplet match to be the top-1 prediction. Gkanatsios et al. [117] reformulated the metric as $\boldsymbol{Recall_k@K}$($\boldsymbol{R_k@K}$). $k = 1$ is equivalent to "graph constraints" and a larger k to "no graph constraints".

Given a set of ground truth triplets, GT, the image-level $\boldsymbol{R@K}$ is computed as:

$$R@K = |Top_K \cap GT| \ / \ |GT|, \tag{4}$$

where Top_K is the top-K triplets extracted from the entire image based on ranked predictions of a model [184]. However, in the PredCLs setting, which is actually a simple classification task, the $\boldsymbol{R@K}$ degenerates into the triplet-level Recall@K ($\boldsymbol{R_{tr}@K}$). $\boldsymbol{R_{tr}@K}$ is similar to the top-K accuracy. Furthermore, Knyazev et al. [184] proposed weighted triplet Recall($\boldsymbol{wR_{tr}@K}$), which computes a recall at each triplet and reweights the average result based on the frequency of the GT in the training set:

$$wR_{tr}@K = \sum_{t}^{T} w_t[rank_t \le K], \tag{5}$$

where T is the number of all test triplets, $[\cdot]$ is the Iverson bracket, $w_t = \frac{1}{(n_t+1)\sum_t 1/(n_t+1)\in[0,1]}$ and n_t is the number of occurrences of the tth triplet in the training set. It is friendly to those infrequent instances, since frequent triplets (with high n_t) are downweighted proportionally. To speak for all predicates rather than very few trivial ones, Tang et al. [105] and Chen et al. [46] proposed $\boldsymbol{meanRecall@K}$($\boldsymbol{mR@K}$) which retrieves each predicate separately then averages $\boldsymbol{R@K}$ for all predicates.

Notably, there is an inconsistency in Recall's definition on the entire test set: whether it is a *micro-* or *macro-Recall* [117]. Let N be the number of testing images and GT_i the ground-truth relationship annotations in image i. Then, having detected $TP_i = Top_{Ki} \cap GT_i$ true positives in the image i, micro-Recall micro-averages these positives as $\frac{\sum_i^N |TP_i|}{\sum_i^N |GT_i|}$ to reward correct predictions across dataset. Macro-Recall computed as $\frac{1}{N}\sum_i^N \frac{|TP_i|}{|GT_i|}$ macro-averages the detections in terms of

images. Early works use micro-Recall on VRD and macro-Recall on VG150, but later works often use the two types interchangeably and without consistency. In some special cases, micro- and macro-Recall@K will be affected differently. When there is an empty ground-truth scene graph, that is, some images without any object or relation annotations, this has differing effects on micro-Recall@K and macro-Recall@K. On the one hand, micro-Recall@K ignores these empty ground-truth scene graphs since they do not contribute to the total triplet count. On the other hand, macro-Recall@K will treat these empty ground-truth scene graphs as zero scores and include them in the final average calculation. In this way, macro-Recall@K is negatively affected by an empty ground-truth scene graph, leading to low evaluation results. When K is chosen to be lower than the total number of triplets in a given image, that is, there are more than K true triplets in some images, this will have negative impact on micro-Recall@K and macro-Recall@ K produces different effects. On the one hand, micro-Recall@K is negatively affected by these images, since they reduce the overall recall. On the other hand, macro-Recall@K normalizes these images so that their influence is relatively small. In the last case, either a nonexistent predicate was predicted, or the wrong object was assigned to the predicate, which had no effect on macro- or micro-recall.

Zero-Shot Recall@K. Zero-shot relationship learning was proposed by Lu et al. [41] to evaluate the performance of detecting zero-shot relationships. Due to the long-tailed relationship distribution in the real world, it is a practical setting to evaluate the extensibility of a model since it is difficult to build a dataset with every possible relationship. Besides, a single $\boldsymbol{wR_{tr}@K}$ value can show zero or few-shot performance linearly aggregated for all $n \geq 0$.

Precision@K. In the video relation detection task, $\boldsymbol{Precision@K}$ ($\boldsymbol{P@K}$) is used to measure the accuracy of the tagging results for the relation tagging task.

mAP. In the OpenImages VRD Challenge, results are evaluated by calculating $Recall@50(R@50)$, mean AP of relationships (mAP_{rel}), and mean AP of phrases (mAP_{phr}) [183]. The mAP_{rel} evaluates AP of $\langle s, p, o\rangle$ triplets where both the subject and object boxes have an IOU of at least 0.5 with the ground truth. The mAP_{phr} is similar, but applied to the enclosing relationship box. mAP would penalize the prediction if that particular ground truth annotation does not exist. Therefore, it is a strict metric because we cannot exhaustively annotate all possible relationships in an image.

R-Precision. Cong et al. [185] argues that Recall@K cannot measure the similarity between an image and the generated scene graph. So he proposed R-Precision, a novel evaluation metric based on Graph Image Contrastive Learning (GICON) for scene graph generation evaluation.

GICON is a contrastive learning framework proposed in [185], which aims to learn the similarity between scene graphs and images. The main idea of GICON is to leverage two transformers, one is a graph transformer, which encodes scene graphs, and the other is an image transformer, which encodes images. Both transformers are neural network models based on the Transformer architecture, which can extract features from the input data and map them to a shared latent space.

The usage of GICON is as follows: Firstly, the scene graph is converted into a sequence with structural encoding, which is called graph serialization. The purpose of graph serialization is to enable the graph transformer to understand the structure of the scene graph and extract representative features. Secondly, the graph sequence and the image are fed into the graph transformer and the image transformer respectively, resulting in their feature vectors. These two feature vectors are in the same dimensional latent space. Thirdly, Given a batch of B scene graph-image pairs, the contrastive learning function Eq. (6) is used to train the two transformers, such that the similarity between matched scene graphs and images is higher than that between unmatched scene graphs and images.

$$L = -\frac{1}{B}\sum_{i=1}^{B}\left(log\frac{exp(sim(\mathbf{g}_i,\mathbf{i}_i))}{\sum_{j=1}^{N} exp(sim(\mathbf{g}_i,\mathbf{i}_j))} + log\frac{exp(sim(\mathbf{g}_i,\mathbf{i}_i))}{\sum_{j=1}^{N} exp(sim(\mathbf{g}_j,\mathbf{i}_i))}\right) \tag{6}$$

where $sim(\mathbf{g}_i,\mathbf{i}_i)$ indicates the cosine similarity between the ith scene graph representation $\mathbf{g}_i$ and the jth image representation $\mathbf{i}_i$ in a batch.

Finally, the generated scene graph is used as a query to retrieve the most matched image from a set of candidate images, or the given image is used as a query to retrieve the most matched scene graph from a set of candidate scene graphs. This process can be evaluated by R-Precision, which measures the retrieval accuracy. R-Precision measures the retrieval accuracy when retrieving the matching image from K image candidates using the generated scene graph as a query. A larger K implies that the retrieval task is more challenging, and the query scene graph's quality demands are higher. R-Precision can evaluate both location-free scene graphs and location-bound scene graphs.

5.3. Quantitative performance

We present the quantitative performance on Recall@K metric of some representative methods on several commonly used datasets in Tables 2 and 3. We preserve the respective task settings and tasks' names for each dataset, though SGGen on VG150 are the same to the RelDet on others. ‡ denotes the experimental results are under **"no graph constraints"**.

By comparing Tables 2 and 3, we notice that only a few of the proposed methods have been simultaneously verified on both VRD and VG150 datasets. The performance of most methods on VG150 is better than that on VRD dataset, because VG150 has been cleaned and enhanced. Experimental results on VG150 can better reflect the performance of different methods, therefore, several recently proposed methods have adopted VG150 to compare their performance metrics with other techniques.

On VG150, excellent performances have been achieved by using the Language Prior's model. Theoretically, Commonsense Knowledge can greatly improve the performance, but in practice, several models that use Prior Knowledge have unsatisfactory performance. We believe the main reason is the difficultly to extract and use the effective knowledge information in the scene graph generation model.

We also present the quantitative performance on R-precision metric of some representative methods on VG150 in Table 4. For both types of scene graphs, whether they are location-free or location-bound, most methods achieve higher R-Precision than using ground truth scene graphs. This indicates that the scene graphs generated by the existing models are more aligned with the visual features of the images than the manually labeled scene graphs, and thus perform better in image retrieval. The manual annotations of Visual Genome are noisy and incomplete. Some ground truth scene graphs only contain one or two relationships, while the generated scene graphs always include several triplets. The bias in the ground truth scene graphs has been noticed by previous studies and many models can overcome this bias and generate more diverse and unbiased scene graphs from the biased training data.

Due to the long tail effect of visual relationships, it is hard to collect images for all the possible relationships. It is therefore crucial for a model to have the generalizability to detect zero-shot relationships. VRD dataset contains 1877 relationships that only exist in the test set. Some researchers have evaluated the performance of their models on zero-shot learning. The performance summary of zero-shot predicate and relationship detection on VRD dataset are shown in Table 5.

Compared with the traditional Recall, meanRecall calculates a Recall rate for each relation. Therefore, meanRecall can better describe the performance of the model on each relation, which is obtained by averaging the Recall of each relation. Table 6 shows the meanRecall metric performance of several typical models.

Tables 7–9 show the performances of several ST-SGG methods on the ImageNet-VidVRD, VidOR datasets and Action Genome. Evaluation of performance is based on two tasks, namely Relation Detection and Relation Tagging. Because of its large size, VidOR presents many challenges to relation detection and tagging. ST-SGG is much more complex

Table 2
Performance summary of some representative methods on VRD dataset.

Models	PredCLs		PhrDet		RelDet		Year
	R@100	R@50	R@100	R@50	R@100	R@50	
LP [41]	47.87	47.84	17.03	16.17	14.70	13.86	2016
VRL [39]	–	–	22.60	21.37	20.79	18.19	2017
U+W+SF+L:S+T [61]	55.16	55.16	24.03	23.14	21.34	19.17	2017
DR-Net [63]	81.90	80.78	23.45	19.93	20.88	17.73	2017
ViP-CNN [13]	–	–	27.91	22.78	20.01	17.32	2017
AP+C+CAT [109]	53.59	53.59	25.56	24.04	23.52	20.35	2017
VTransE [118]	–	–	22.42	19.42	15.20	14.07	2017
Cues [186]	–	–	20.70	16.89	18.37	15.08	2017
Weakly-supervised [187]	–	46.80	–	16.00	–	14.10	2017
PPR-FCN [188]	47.43	47.43	23.15	19.62	15.72	14.41	2017
Large VRU [15]	–	–	39.66	32.90	32.63	26.98	2018
Interpretable SGG [43]	–	–	41.25	33.29	32.55	26.67	2018
CDDN-VRD [189]	93.76	87.57	–	–	26.14	21.46	2018
DSR [14]	93.18	86.01	–	–	23.29	19.03	2018
Joint VSE [190]	–	–	24.12	20.53	16.26	14.23	2018
F_o+L^m [191]	–	–	23.95	22.67	18.33	17.40	2018
SG-CRF [94]	50.47	49.16	–	–	25.48	24.98	2018
OSL [116]	56.56	56.56	24.50	20.82	16.01	13.81	2018
F-Net [52]	–	–	30.77	26.03	21.20	18.32	2018
Zoom-Net [58]	50.69	50.69	28.09	24.82	21.41	18.92	2018
CAI+SCA-M [58]	55.98	55.98	28.89	25.21	22.39	19.54	2018
VSA-Net [110]	49.22	49.22	21.65	19.07	17.74	16.03	2018
MF-URLN [78]	58.20	58.20	36.10	31.50	26.80	23.90	2019
LRNNTD [60]	–	–	30.92	28.53	25.87	24.20	2019
KB-GAN [77]	–	–	34.38	27.39	25.01	20.31	2019
RLM [96]	57.19	57.19	39.74	33.20	31.15	26.55	2019
NMP [95]	57.69	57.69	–	–	23.98	20.19	2019
MLA-VRD [108]	95.05	90.18	28.12	23.36	24.91	20.54	2019
ATR-Net [117]	58.40	58.40	34.63	29.74	24.87	22.83	2019
BLOCK [192]	92.58	86.58	28.96	26.32	20.96	19.06	2019
MR-Net [193]	61.19	61.19	–	–	17.58	16.71	2019
RelDN [183]	–	–	36.42	31.34	28.62	25.29	2019
UVTransE [122]	–	26.49	18.44	13.07	16.78	11.00	2020
AVR [182]	55.61	55.61	33.27	29.33	25.41	22.83	2020
GPS-Net [114]	–	63.40	39.20	33.80	31.70	27.80	2020
MemoryNet [194]	–	–	34.90	29.80	27.90	24.30	2020
HET [195]	–	–	42.94	35.47	24.88	22.42	2020
HOSE-Net [196]	–	–	31.71	37.04	23.57	20.46	2020
SABRA [197]	–	–	39.62	33.56	32.48	27.87	2020
Wang's [112]	–	–	28.46	24.12	25.01	21.25	2021
RelTR [49]	–	–	39.80	34.50	32.20	29.20	2023
NLGVRD‡ [66]	92.65	84.92	47.92	42.29	22.22	20.81	2017
U+W+SF+L:S+T‡ [61]	94.65	85.64	29.43	26.32	31.89	22.68	2017
Zoom-Net‡ [58]	90.59	84.05	37.34	29.05	27.30	21.37	2018
CAI+SCA-M‡ [58]	94.56	89.03	38.39	29.64	28.52	22.34	2018
LRNNTD‡ [60]	–	–	41.28	32.29	34.93	27.09	2019
RLM‡ [96]	96.48	90.00	46.03	36.79	37.35	30.22	2019
NMP‡ [95]	96.61	90.61	–	–	27.50	21.50	2019
ATR-Net‡ [117]	96.97	91.00	41.01	33.20	31.94	26.04	2019
RelDN‡ [183]	–	–	42.12	34.45	33.91	28.15	2019
AVR‡ [182]	95.72	90.73	41.36	34.51	32.96	27.35	2020
MemoryNet‡ [194]	–	–	39.80	32.10	32.40	26.50	2020
HET‡ [195]	–	–	43.05	35.47	31.81	26.88	2020
HOSE-Net‡ [196]	–	–	36.16	28.89	27.36	22.13	2020
SABRA‡ [197]	–	–	45.29	36.62	37.71	30.71	2020

than 2D SGG because additional steps such as object tracking, temporal segmentation, and merging the detected relationships in different segments are involved. It is expected that ST-SGG's performance will improve as more researchers contribute.

Since there are few works on 3D scene graph generation methods at present, we only list the results of several recent methods on the 3DSGG dataset in Table 10. We also adapt the constrained evaluation metric recall@K (R@K) and mean recall@K (mR@K) in three tasks: Predicate Classification (PredCls), Scene Graph Classification (SGCls) and Scene Graph Generation (SGDet). It can be seen from the data in the table that although the 3D SGG work started late, its development is quite rapid. It is believed that it will reach a level comparable to 2D SGG soon.

5.4. Qualitative performance

We compare different BGNN [208] and SGTR [51] by visualizing the relationship predictions in Fig. 9, mark the different relationship predictions between BGNN and SGTR with red color. Both BGNN and SGTR predict many reasonable relationships that are not found in ground truth. The difference between the two is that the relationships predicted by BGNN are basically frequent relationships, while SGTR uses long-tail learning strategies, thereby retrieving more relationships of less frequent semantic categories than BGNN. At the same time, it can also be seen that the relation predicted by the current model is not necessarily wrong, but the dataset annotation is not perfect enough.

Table 3
Performance summary of some representative methods on VG150 dataset.

Models	PredCLs		PhrDet		RelDet		Year
	R@100	R@50	R@100	R@50	R@100	R@50	
IMP [99]	53.08	44.75	24.38	21.72	4.24	3.44	2017
Px2graph [198]	86.40	82.00	38.40	35.70	18.80	15.50	2017
Interpretable SGG [43]	68.30	68.30	36.70	36.70	32.50	28.10	2018
IK-R_e [173]	77.60	67.71	42.74	35.55	–	–	2018
SK-R_e [173]	77.43	67.42	42.25	35.07	–	–	2018
TFR [71]	58.30	51.90	26.60	24.30	6.00	4.80	2018
MotifNet [64]	67.10	65.20	36.50	35.80	30.30	27.20	2018
Graph R-CNN [113]	59.10	54.20	31.60	29.60	13.70	11.40	2018
LinkNet [199]	68.50	67.00	41.70	41.00	30.10	27.40	2018
GPI [200]	66.90	65.10	38.80	36.50	–	–	2018
KERN [46]	67.60	65.80	37.40	36.70	29.80	27.1	2019
SGRN [89]	66.40	64.20	39.70	38.60	35.40	32.30	2019
Mem+Mix+Att [98]	57.90	53.20	29.50	27.80	13.90	11.40	2019
VCTREE [105]	68.10	66.40	38.80	38.10	31.30	27.90	2019
CMAT [101]	68.10	66.40	39.80	39.00	31.20	27.90	2019
VRasFunctions [97]	57.21	56.65	24.66	23.71	13.45	13.18	2019
PANet [102]	67.90	66.00	41.80	40.90	29.90	26.90	2019
ST+GSA+RI [47]	61.30	56.60	40.40	38.20	–	–	2019
Attention [115]	67.10	65.00	37.10	36.30	29.50	26.60	2019
RelDN [183]	68.40	68.40	36.80	36.80	32.70	28.30	2019
Large VRU [15]	68.40	68.40	36.70	36.70	32.50	27.9	2019
GB-NET [82]	68.20	66.60	38.80	38.00	30.00	26.40	2020
UVTransE [122]	67.30	65.30	36.60	35.90	33.60	30.10	2020
GPS-Net [114]	69.70	69.70	42.30	42.30	33.20	28.90	2020
RiFa [74]	88.35	80.64	44.38	37.62	26.68	20.86	2020
RONNIE [201]	69.00	65.00	37.00	36.20	–	–	2020
DG-PGNN [202]	73.00	70.10	40.80	39.50	33.10	32.10	2020
NODIS [203]	69.10	67.20	41.50	40.60	31.50	28.10	2020
HCNet [204]	68.80	66.40	37.30	36.60	31.20	28.00	2020
Self-supervision [205]	68.87	68.85	37.03	37.01	32.56	28.28	2020
MemoryNet [194]	69.30	69.20	37.10	37.10	32.40	27.60	2020
HET [195]	68.10	66.30	37.30	36.60	30.90	27.50	2020
HOSE-Net [196]	69.20	66.70	37.40	36.30	33.30	28.90	2020
PAIL [206]	69.40	67.70	40.20	39.40	32.70	29.40	2020
BL+SO+KT+FC [207]	68.80	66.20	38.30	37.50	31.40	28.20	2020
FCSGG [48]	45.00	41.00	25.70	23.50	25.10	21.30	2021
SGTR [50]	–	–	–	–	25.00	20.60	2022
SSR-CNN [50]	–	–	–	–	36.90	32.70	2022
RelTR [49]	–	64.20	–	36.60	–	27.50	2023
Interpretable SGG‡ [43]	97.70	93.70	50.80	48.90	36.40	30.10	2018
MotifNet‡ [64]	88.30	81.10	47.70	44.50	35.80	30.50	2018
GPI‡ [200]	88.20	80.80	50.80	45.50	–	–	2018
KERN‡ [46]	88.90	81.90	49.00	45.90	35.80	30.90	2019
CMAT‡ [101]	90.10	83.20	52.00	48.60	36.80	31.60	2019
PANet‡ [102]	89.70	82.60	55.20	51.30	36.30	31.10	2019
RelDN‡ [183]	97.80	93.80	50.80	48.90	36.70	30.40	2019
GB-NET‡ [82]	90.50	83.60	51.10	47.70	35.10	29.40	2020
HOSE-Net‡ [196]	89.20	81.10	48.10	44.20	36.30	30.50	2020
BL+SO+KT+FC‡ [207]	90.20	82.50	50.20	46.20	36.50	31.40	2020

Table 4
Performance summary of some representative methods on VG150 dataset using R-Precision.

Model		R-Precision (LF graph)			R-Precision (LB graph)		
		K = 10	K = 50	K = 100	K = 10	K = 50	K = 100
Ground truth		86.7	67.9	58	92.1	78	71
Two-stage	MOTIFS [64]	84.5	65.1	54.9	90.2	74.5	69.1
	RelDN [183]	88.9	72.5	62.7	93.5	83.1	75
	NODIS [203]	85	66.9	58.1	90.5	75.6	69.8
	MOTIFS-TDE [127]	94.5	82.9	74.7	97.8	91.8	87.1
	VCTree-EBM [126]	92.9	78	68.9	96.5	87.7	81.7
	BGNN [208]	93.3	79.4	70.9	97.1	89.7	84.3
One-stage	FCSGG [48]	79	59.2	49.8	87.4	70.1	66.7
	SGTR [51]	94.5	83.1	74.9	97.3	90.8	85.7
	RelTR [49]	93.1	80.1	71.1	97	90.1	84.8
	SSR-CNN [50]	91.6	74.4	64	95.9	86.5	79.6

Table 5
Performance summary of some representative methods for **zero-shot** visual relationship detection on the VRD dataset.

Models	PredCLs		PhrDet		RelDet		Year
	R@100	R@50	R@100	R@50	R@100	R@50	
LP [41]	8.45	8.45	3.75	3.36	3.52	3.13	2016
VRL [39]	–	–	10.31	9.17	8.52	7.94	2017
Cues [186]	–	–	15.23	10.86	13.43	9.67	2017
VTransE [118]	–	–	3.51	2.65	2.14	1.71	2017
Weakly-supervised [187]	–	19.00	–	6.90	–	6.70	2017
U+W+SF+L:S [61]	16.98	16.98	10.89	10.44	9.14	8.89	2017
AP+C+CAT [109]	16.37	16.37	11.30	10.78	10.26	9.54	2017
PPR-FCN [188]	–	–	8.22	6.93	6.29	5.68	2017
DSR [14]	79.81	60.90	–	–	9.20	5.25	2018
CDDN-VRD [189]	84.00	67.66	–	–	10.29	6.40	2018
Joint VSE [190]	–	–	6.16	5.05	5.73	4.79	2018
SG-CRF [94]	21.22	–	6.70	–	5.22	–	2018
MF-URLN [78]	26.90	26.90	7.90	5.90	5.50	4.30	2019
MLA-VRD [108]	88.96	73.65	13.84	8.43	12.81	8.08	2019
U+W+SF+L:S‡ [61]	74.65	54.20	17.24	13.01	16.15	12.31	2017

Table 6
Mean recall performance summary of some typical methods on VG150 dataset.

Models	PredCLs		SGCLs		SGGen		Year
	mR@50	mR@100	mR@50	mR@100	mR@50	mR@100	
IMP [99]	6.1	8.0	3.1	3.8	0.6	0.9	2017
IMP+ [64]	9.8	10.5	5.8	6.0	3.8	4.8	2018
FREQ [64]	13.0	16.0	7.2	8.5	6.1	7.1	2018
MotifNet [64]	14.0	15.3	7.7	8.2	5.7	6.6	2018
KERN [46]	17.7	19.2	9.4	10	6.4	7.3	2019
VCTREE-SL [105]	17.0	18.5	9.8	10.5	6.7	7.7	2019
VCTREE-HL [105]	17.9	19.4	10.1	10.8	6.9	8	2019
GPS-Net [114]	21.3	22.8	11.8	12.6	8.7	9.8	2020
MemoryNet [194]	22.6	22.7	10.9	11	7.4	9	2020
PAIL [206]	19.2	20.9	10.9	11.6	7.7	8.8	2020
GB-NET [82]	19.3	20.9	9.6	10.2	6.1	7.3	2020
GB-NET-β [82]	22.1	24.0	12.7	13.4	7.1	8.5	2020
FCSGG [48]	6.3	7.1	3.7	4.1	3.6	4.2	2021
SGTR [50]	–	–	–	–	15.8	20.1	2022
SSR-CNN [50]	–	–	–	–	8.4	10.0	2022
RelTR [49]	21.2	–	11.4	–	10.8	–	2023

Table 7
Performance for standard video relation detection and video relation tagging on ImageNet-VidVRD dataset [134].

Models	Relation detection			Relation tagging			Year
	R@50	R@100	mAP	P@1	P@5	P@10	
VP [53]	0.89	1.41	1.01	36.50	25.55	19.20	2011
Lu's-V [41]	0.99	1.80	2.37	20.00	12.60	9.55	2016
Lu's [41]	1.10	2.23	2.40	20.50	16.30	14.05	2016
VTransE [118]	0.72	1.45	1.23	15.00	10.00	7.65	2017
VidVRD [143]	5.54	6.37	8.58	43.00	28.90	20.80	2017
GSTEG [132]	7.05	8.67	9.52	51.50	39.50	28.23	2019
VRD-GCN [131]	8.07	9.33	16.26	57.50	41.00	28.50	2019
MHRA [209]	6.82	7.39	13.27	41.0	28.7	20.95	2019
VRD-STGC [134]	11.21	13.69	18.38	60.00	43.10	32.24	2020
Liu's [134]	9.14	11.39	14.81	55.5	38.9	28.9	2020
Teng's [148]	9.08	11.15	17.57	61	45.3	33.5	2021

Table 8
Performance for standard video relation detection and video relation tagging on VidOR dataset [134].

Models	Relation detection			Relation tagging		Year
	R@50	R@100	mAP	P@1	P@5	
RELAbuilder [146]	1.58	1.85	1.47	33.05	35.27	2019
OTD+CAI [147]	6.19	8.16	5.65	48.31	38.49	2019
OTD+GSTEG [147]	6.40	8.43	5.58	51.20	37.26	2019
MAGUS.Gamma [147]	6.89	8.83	6.56	51.20	40.73	2019
VRD-STGC [134]	8.21	9.90	6.85	48.92	36.78	2020

6. Challenges & future research directions

6.1. Challenges

There is not doubt that there are many excellent SGG models which have achieved good performance on the standard image datasets, such as VRD and VG150. However, there are still several challenges that have not been well resolved.

First, both the number of objects in the real world and the number of categories of relations are very large, but reasonable and meaningful relationships are scarce. Therefore, detecting all individual objects first and then classifying all pairs would be inefficient. Moreover, object classification networks require a fixed number of output categories, which does not scale with real-world images. Several works [13,15,56,63,89,113,188,212] have helped to filter out a set of object pairs with a low probability of interaction from the set of detected objects. An effective proposal network will definitely reduce the learning complexity and computational cost for the subsequent predicate classification, thus improving the accuracy of relationship detection.

The ***second*** main challenge comes from the long-tailed distribution of the visual relationships. Since interaction occurs between two objects, there is a greater skew of rare relationships, as object co-occurrence is infrequent in a real-world scenario. An uneven distribution makes it difficult for the model to fully understand the properties of some rare relationships and triplets. For example, if a model is trained to predict "on" 1000 times more than "standing on", then, during the test phase, "on" is more likely to prevail over "standing on". ***This phenomenon where the model is more likely to predict a simple and coarse relation than the accurate relation is called Biased Scene***

Table 9
Performance for standard video relation detection and video relation tagging on action Genome dataset [133].

Models	PredCLs			SGCLs			SGGen			Year
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	
VRD [41]	59.60	78.50	99.20	39.20	49.80	52.60	19.10	28.80	40.50	2016
MSDN [16]	74.90	92.70	99.00	51.20	61.80	65.00	23.10	34.70	46.50	2017
Motif Freq [64]	73.40	92.40	99.60	50.40	60.60	64.20	22.80	34.30	46.40	2018
G-RCNN [113]	–	88.73	93.73	–	45.57	49.75	–	34.28	44.47	2018
VC Tree [105]	75.50	92.90	99.30	52.40	62.00	65.10	23.90	35.30	46.80	2019
RelDN [183]	75.70	93.00	99.00	52.90	62.40	65.10	24.10	35.40	46.80	2019
GPS-NET [114]	76.00	93.60	99.50	53.60	63.30	66.00	24.40	35.70	47.30	2020
STTran [149]	77.90	94.20	99.10	54.00	63.70	66.40	24.60	36.20	48.80	2021
Teng's [148]	–	91.60	96.35	–	46.66	50.46	–	35.09	45.34	2021

Table 10
Performance for 3D scene graph prediction on the 3DSGG dataset [160].

Models	PredCLs		SGCLs		SGGen		Year
	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100	
SGPN [160]	57.71/58.05	38.12/38.67	28.39/28.74	22.23/22.57	–/–	–/–	2020
EdgeGCN [210]	58.42/59.11	38.84/39.35	28.58/28.93	22.67/23.33	–/–	–/–	2021
KISG [211]	64.47/64.93	63.19/63.52	29.46/29.65	28.20/28.64	–/–	–/–	2021
Feng's [154]	68.32/69.49	66.54/66.92	31.50/31.64	30.29/30.56	29.41/29.44	25.35/25.36	2023

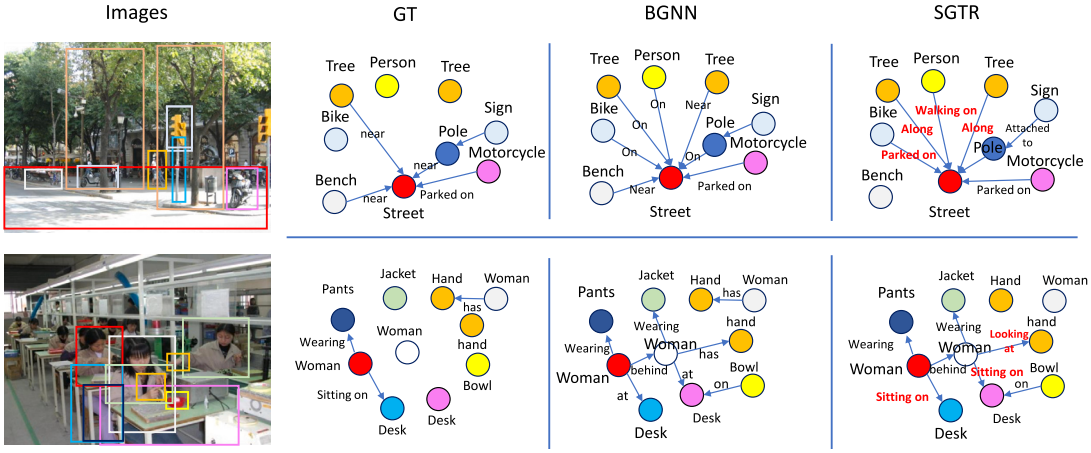


Fig. 9. Qualitative comparison between SGTR [51] and BGNN [208]. Both methods predict many reasonable relationships that are not annotated in the ground truth (GT). The rare semantic predicate categories retrieved by SGTR are marked in red color.

Graph Generation. Under this condition, even though the model can output a reasonable predicate, it is too coarse and obscure to describe the scene. However, for several downstream tasks, an accurate and informative pair-wise relation is undoubtedly the most fundamental requirement. Therefore, to perform a sensible graph reasoning, we need to distinguish between the more fine-grained relationships from the ostensibly probable but trivial ones, which is generally regarded as unbiased scene graph generation. A lot of works [14,61,94,108,118,186–189,213,214] have provided solutions for zero-shot relationship learning. Some researchers recently proposed unbiased SGG [127,215–218] to make the tail classes to receive more attention in a coarse-to-fine mode.

The **third** challenge is that the visual appearance of the same relation varies greatly from scene to scene (Fig. 3a and d). This makes the feature extraction phase more challenging. As we have described in Section 3.1.2, a great deal of methods focuses on semantic features, trying to make up for the lack of visual features. However, we have emphasized that visual relationships are incidental and scene-specific. This requires us to think from the bottom up and try to extract more discriminative visual features.

The **fourth** challenge is the lack of clarity/consensus in the definition of the relationships. It is always difficult to give mutually exclusive definitions to the predicate categories as opposed to objects, which have clear meaning. As a result, one relationship can be labeled with different but reasonable predicates, making the datasets noisy and the general SGG task ill-posed. Providing a well-defined relationship set is therefore one of the key challenges of the SGG task.

The **fifth** challenge is the evaluation metric. Even though many evaluation metrics are used to assess the performance of the proposed networks and Recall@K or meanRecall@K are common and widely adopted, none of them can provide perfect statistics on how well the model performs on the SGG task. When Recall@50 equals 100, does that mean that the model generates the perfect scene graph for an image? Of course not. The existing evaluation metrics only reveal the relative performance, especially in the current research stage. As the research on SGG progresses, the evaluation metrics and benchmark datasets will pose a great challenge.

6.2. Opportunities

The community has published hundreds of scene graph models and has obtained a wealth of research result. We think there are several avenues for future work. Researchers will be motivated to explore more models as a result of the above challenges. Besides, on one hand, from the learning point of view, building a large dataset with fine-grained labels and accurate annotations is necessary and significant. It contains as many scenes as possible, preferably constructed by computer vision

experts. The models trained on such a dataset will have better performance on visual semantic and develop a broader understanding of our visual world. However, this is a very challenging and expensive task. On the other hand, from the application point of view, we can design the models by subdividing the scene to reduce the imbalance of the relationship distribution. Obviously, the categories and probability distributions of visual relationships are different in different scenarios. Of course, even the types of objects are different. As a result, we can design relationship detection models for different scenarios and employ ensemble learning methods to promote scene graph generation applications.

Another area of research is 3D scene graphs. An initial step is to define an effective and unified 3D scene graph structure, along with what information it should encode. A 2D image is a two-dimensional projection of a 3D world scene taken from a specific viewpoint. It is the specific viewpoint that makes some descriptions of spatial relationships in 2D images meaningful. Taking the triplet of $\langle women, is\ behind, fire\ hydrant \rangle$ in Fig. 1 as an example, the relation "is behind" makes sense because of the viewpoint. But, how can a relation be defined as "is behind" in 3D scenes without a given viewpoint? Therefore, the challenge is how to define such spatial semantic relationships in 3D scenes without simply resorting to 2.5D scenes (for example, RGB-D data captured from a specific viewpoint). Armeni et al. augment the basic scene graph structure with essential 3D information and generate a 3D scene graph which extends the scene graph to 3D space and ground semantic information there [28,159]. However, their proposed structure representation does not have expansibility and generality. Second, because 3D information can be grounded in many storage formats, which are fragmented to specific types based on the visual modality (e.g., RGB-D, point clouds, 3D mesh/CAD models, etc.), the presentation and extraction of 3D semantic information has technological challenges.

7. Conclusion

This paper provides a comprehensive survey of the developments in the field of scene graph generation using deep learning techniques. We first introduced the representative works on 2D scene graph, spatio-temporal scene graph and 3D scene graph in different sections, respectively. Furthermore, we provided a summary of some of the most widely used datasets for visual relationship and scene graph generation, which are grouped into 2D images, video, and 3D representation, respectively. The performance of different approaches on different datasets are also compared. Finally, we discussed the challenges, problems and opportunities on the scene graph generation research. We believe this survey can promote more in-depth ideas used on SGG.

CRediT authorship contribution statement

Hongsheng Li: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Guangming Zhu:** Conceptualization, Formal analysis, Supervision, Writing – review & editing. **Liang Zhang:** Conceptualization, Writing – review & editing. **Youliang Jiang:** Formal analysis, Data curation. **Yixuan Dang:** Formal analysis, Data curation. **Haoran Hou:** Formal analysis, Data curation. **Peiyi Shen:** Supervision, Writing – review & editing. **Xia Zhao:** Supervision, Writing – review & editing. **Syed Afaq Ali Shah:** Supervision, Writing – review & editing. **Mohammed Bennamoun:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] J. Li, J.Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1075–1088.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[3] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[4] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[5] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[6] S. Asgari Taghanaki, K. Abhishek, J.P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: a review, Artif. Intell. Rev. 54 (1) (2021) 137–178.

[7] A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár, Panoptic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9404–9413.

[8] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015, pp. 234–241.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 801–818.

[10] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, Vol. 28, 2015.

[11] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, 2018, arXiv preprint arXiv:1804.02767.

[12] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al., Hybrid task cascade for instance segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4974–4983.

[13] Y. Li, W. Ouyang, X. Wang, 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.

[14] K. Liang, Y. Guo, H. Chang, X. Chen, Visual relationship detection with deep structural ranking, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

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

[16] Y. Li, W. Ouyang, B. Zhou, K. Wang, 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.

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

[18] S. Qi, W. Wang, B. Jia, J. Shen, 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.

[19] T. Wang, R.M. Anwer, M.H. Khan, F.S. Khan, Y. Pang, L. Shao, J. Laaksonen, Deep contextual attention for human-object interaction detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5694–5702.

[20] R. Grzeszick, G.A. Fink, Zero-shot object prediction using semantic scene knowledge, 2016, arXiv preprint arXiv:1604.07952.

[21] J. Deng, Y. Pan, T. Yao, W. Zhou, H. Li, T. Mei, Relation distillation networks for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7023–7032.

[22] Y. Liu, R. Wang, S. Shan, 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.

[23] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3588–3597.

[24] R. Krishna, I. Chami, M. Bernstein, L. Fei-Fei, Referring relationships, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6867–6876.

[25] S.K. Divvala, D. Hoiem, J.H. Hays, A.A. Efros, M. Hebert, An empirical study of context in object detection, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1271–1278.

[26] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, L. Fei-Fei, Image retrieval using scene graphs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3668–3678.

[27] S.Y. Bao, M. Bagra, Y.-W. Chao, S. Savarese, Semantic structure from motion with points, regions, and objects, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2703–2710.

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

[29] L. Gao, B. Wang, 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.

[30] X. Yang, K. Tang, H. Zhang, J. Cai, Auto-encoding scene graphs for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10685–10694.

[31] D.-J. Kim, J. Choi, T.-H. Oh, I.S. Kweon, Dense relational captioning: Triple-stream networks for relationship-based captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6271–6280.

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

[33] L. Li, Z. Gan, Y. Cheng, J. Liu, Relation-aware graph attention network for visual question answering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10313–10322.

[34] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, 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.

[35] J. Johnson, A. Gupta, 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] G. Mittal, S. Agrawal, A. Agarwal, S. Mehta, T. Marwah, Interactive image generation using scene graphs, 2019, arXiv preprint arXiv:1905.03743.

[37] S. Yang, G. Li, Y. Yu, Cross-modal relationship inference for grounding referring expressions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4145–4154.

[38] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, A. Hauptmann, A comprehensive survey of scene graphs: Generation and application, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2021) 1–26.

[39] X. Liang, L. Lee, 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.

[40] 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, Int. J. Comput. Vis. 123 (1) (2017) 32–73.

[41] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual relationship detection with language priors, in: European Conference on Computer Vision, 2016, pp. 852–869.

[42] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, et al., The open images dataset v4, Int. J. Comput. Vis. 128 (7) (2020) 1956–1981.

[43] J. Zhang, K. Shih, A. Tao, B. Catanzaro, A. Elgammal, An interpretable model for scene graph generation, 2018, arXiv preprint arXiv:1811.09543.

[44] S. Ji, S. Pan, E. Cambria, P. Marttinen, S.Y. Philip, A survey on knowledge graphs: Representation, acquisition, and applications, IEEE Trans. Neural Netw. Learn. Syst. 33 (2) (2021) 494–514.

[45] H. Wan, Y. Luo, B. Peng, W.-S. Zheng, Representation learning for scene graph completion via jointly structural and visual embedding, in: International Joint Conference on Artificial Intelligence, 2018, pp. 949–956.

[46] T. Chen, W. Yu, R. Chen, L. Lin, Knowledge-embedded routing network for scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6163–6171.

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

[48] H. Liu, N. Yan, M. Mortazavi, B. Bhanu, Fully convolutional scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11546–11556.

[49] Y. Cong, M.Y. Yang, B. Rosenhahn, Reltr: Relation transformer for scene graph generation, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[50] Y. Teng, L. Wang, Structured sparse r-cnn for direct scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19437–19446.

[51] R. Li, S. Zhang, X. He, Sgtr: End-to-end scene graph generation with transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19486–19496.

[52] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, 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.

[53] M.A. Sadeghi, A. Farhadi, Recognition using Visual Phrases, IEEE, 2011.

[54] Y. Zhu, S. Jiang, X. Li, Visual relationship detection with object spatial distribution, in: 2017 IEEE International Conference on Multimedia and Expo (ICME), 2017, pp. 379–384.

[55] S. Sharifzadeh, S.M. Baharlou, M. Berrendorf, R. Koner, V. Tresp, Improving visual relation detection using depth maps, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 3597–3604.

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

[57] N. Xu, A.-A. Liu, Y. Wong, W. Nie, Y. Su, M. Kankanhalli, Scene graph inference via multi-scale context modeling, IEEE Trans. Circuits Syst. Video Technol. 31 (3) (2020) 1031–1041.

[58] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, C.C. 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.

[59] J. Jung, J. Park, Visual relationship detection with language prior and softmax, in: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), 2018, pp. 143–148.

[60] M.H. Dupty, Z. Zhang, W.S. Lee, Visual relationship detection with low rank non-negative tensor decomposition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 10737–10744.

[61] R. Yu, A. Li, V.I. Morariu, 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.

[62] S. Baier, Y. Ma, V. Tresp, Improving visual relationship detection using semantic modeling of scene descriptions, in: International Semantic Web Conference, 2017, pp. 53–68.

[63] B. Dai, Y. Zhang, D. Lin, Detecting visual relationships with deep relational networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3076–3086.

[64] R. Zellers, M. Yatskar, S. Thomson, 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.

[65] F. Amodeo, F. Caballero, N. Díaz-Rodríguez, L. Merino, OG-SGG: ontology-guided scene graph generation—a case study in transfer learning for telepresence robotics, IEEE Access 10 (2022) 132564–132583.

[66] W. Liao, B. Rosenhahn, L. Shuai, M. Ying Yang, Natural language guided visual relationship detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 1–10.

[67] S. Abdelkarim, A. Agarwal, P. Achlioptas, J. Chen, J. Huang, B. Li, K. Church, M. Elhoseiny, Exploring long tail visual relationship recognition with large vocabulary, (2021) 15921–15930.

[68] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, C.-W. Chen, Learning to generate language-supervised and open-vocabulary scene graph using pre-trained visual-semantic space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2915–2924.

[69] T. He, L. Gao, J. Song, Y.-F. Li, Towards open-vocabulary scene graph generation with prompt-based finetuning, in: European Conference on Computer Vision, Springer, 2022, pp. 56–73.

[70] Y. Zhong, J. Shi, J. Yang, C. Xu, Y. Li, Learning to generate scene graph from natural language supervision, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1823–1834.

[71] S.J. Hwang, S.N. Ravi, Z. Tao, H.J. Kim, M.D. Collins, 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.

[72] I. Donadello, L. Serafini, Compensating supervision incompleteness with prior knowledge in semantic image interpretation, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.

[73] Y. Zhou, S. Sun, C. Zhang, Y. Li, W. Ouyang, Exploring the hierarchy in relation labels for scene graph generation, 2020, arXiv preprint arXiv:2009.05834.

[74] B. Wen, J. Luo, X. Liu, L. Huang, Unbiased scene graph generation via rich and fair semantic extraction, 2020, arXiv preprint arXiv:2002.00176.

[75] M.Y. Yang, W. Liao, H. Ackermann, B. Rosenhahn, On support relations and semantic scene graphs, ISPRS J. Photogramm. Remote Sens. 131 (2017) 15–25.

[76] J. Duan, W. Min, D. Lin, J. Xu, X. Xiong, Multimodal graph inference network for scene graph generation, Appl. Intell. 51 (12) (2021) 8768–8783.

[77] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, M. Ling, Scene graph generation with external knowledge and image reconstruction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1969–1978.

[78] Y. Zhan, J. Yu, T. Yu, D. Tao, On exploring undetermined relationships for visual relationship detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5128–5137.

[79] B. Lin, Y. Zhu, X. Liang, Atom correlation based graph propagation for scene graph generation, Pattern Recognit. 122 (2022) 108300.

[80] Y. Yao, A. Zhang, X. Han, M. Li, C. Weber, Z. Liu, S. Wermter, M. Sun, Visual distant supervision for scene graph generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15816–15826.

[81] J. Yu, Y. Chai, Y. Wang, Y. Hu, Q. Wu, Cogtree: Cognition tree loss for unbiased scene graph generation, 2020, arXiv preprint arXiv:2009.07526.

[82] A. Zareian, S. Karaman, S.-F. Chang, Bridging knowledge graphs to generate scene graphs, in: European Conference on Computer Vision, 2020, pp. 606–623.

[83] A. Zareian, Z. Wang, H. You, S.-F. Chang, Learning visual commonsense for robust scene graph generation, in: European Conference on Computer Vision, 2020, pp. 642–657.

[84] K.S. Tai, R. Socher, C.D. Manning, Improved semantic representations from tree-structured long short-term memory networks, 2015, arXiv preprint arXiv:1503.00075.

[85] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.

[86] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[87] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.

[88] D. Shin, I. Kim, Deep image understanding using multilayered contexts, Math. Probl. Eng. 2018 (2018).

[89] W. Liao, C. Lan, W. Zeng, M.Y. Yang, B. Rosenhahn, Exploring the semantics for visual relationship detection, 2019, arXiv preprint arXiv:1904.02104.

[90] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with gaussian edge potentials, in: Advances in Neural Information Processing Systems, Vol. 24, 2011.

[91] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[92] X. Liang, X. Shen, J. Feng, L. Lin, S. Yan, Semantic object parsing with graph lstm, in: European Conference on Computer Vision, 2016, pp. 125–143.

[93] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.

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

[95] Y. Hu, S. Chen, X. Chen, Y. Zhang, X. Gu, Neural message passing for visual relationship detection, in: ICML Workshop on Learning and Reasoning with Graph-Structured Representations, 2019, pp. 1–6.

[96] H. Zhou, C. Hu, C. Zhang, S. Shen, Visual relationship recognition via language and position guided attention, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2097–2101.

[97] A. Dornadula, A. Narcomey, R. Krishna, M. Bernstein, F.-F. Li, Visual relationships as functions: Enabling few-shot scene graph prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 1–10.

[98] W. Wang, R. Wang, S. Shan, X. Chen, Exploring context and visual pattern of relationship for scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8188–8197.

[99] D. Xu, Y. Zhu, C.B. Choy, 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.

[100] N. Dhingra, F. Ritter, A. Kunz, BGT-Net: Bidirectional GRU transformer network for scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2150–2159.

[101] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, S.-F. Chang, Counterfactual critic multi-agent training for scene graph generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4613–4623.

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

[103] Y. Dai, C. Wang, J. Dong, C. Sun, Visual relationship detection based on bidirectional recurrent neural network, Multimedia Tools Appl. 79 (47) (2020) 35297–35313.

[104] K. Masui, A. Ochiai, S. Yoshizawa, H. Nakayama, Recurrent visual relationship recognition with triplet unit for diversity, Int. J. Semant. Comput. 12 (04) (2018) 523–540.

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

[106] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: Advances in Neural Information Processing Systems, Vol. 27, 2014.

[107] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.

[108] S. Zheng, S. Chen, Q. Jin, Visual relation detection with multi-level attention, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 121–129.

[109] B. Zhuang, L. Liu, C. Shen, I. Reid, Towards context-aware interaction recognition for visual relationship detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 589–598.

[110] C. Han, F. Shen, L. Liu, Y. Yang, H.T. Shen, Visual spatial attention network for relationship detection, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 510–518.

[111] A. Kolesnikov, A. Kuznetsova, C. Lampert, V. Ferrari, Detecting visual relationships using box attention, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 1–5.

[112] L. Wang, P. Lin, J. Cheng, F. Liu, X. Ma, J. Yin, Visual relationship detection with recurrent attention and negative sampling, Neurocomputing 434 (2021) 55–66.

[113] J. Yang, J. Lu, S. Lee, D. Batra, D. Parikh, Graph r-cnn for scene graph generation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 670–685.

[114] X. Lin, C. Ding, J. Zeng, D. Tao, Gps-net: Graph property sensing network for scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3746–3753.

[115] L. Zhang, S. Zhang, P. Shen, G. Zhu, S. Afaq Ali Shah, M. Bennamoun, Relationship detection based on object semantic inference and attention mechanisms, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 68–72.

[116] L. Zhou, J. Zhao, J. Li, L. Yuan, J. Feng, Object relation detection based on one-shot learning, 2018, arXiv preprint arXiv:1807.05857.

[117] N. Gkanatsios, V. Pitsikalis, P. Koutras, P. Maragos, Attention-translation-relation network for scalable scene graph generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 1–11.

[118] H. Zhang, Z. Kyaw, S.-F. Chang, 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.

[119] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Advances in Neural Information Processing Systems, Vol. 26, 2013.

[120] Z. Wang, J. Zhang, J. Feng, Z. Chen, Knowledge graph embedding by translating on hyperplanes, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 28, 2014.

[121] G. Ji, K. Liu, S. He, J. Zhao, Knowledge graph completion with adaptive sparse transfer matrix, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[122] Z.-S. Hung, A. Mallya, S. Lazebnik, Contextual translation embedding for visual relationship detection and scene graph generation, IEEE Trans. Pattern Anal. Mach. Intell. 43 (11) (2020) 3820–3832.

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

[124] B. Knyazev, H. de Vries, C. Cangea, G.W. Taylor, A. Courville, E. Belilovsky, Generative graph perturbations for scene graph prediction, 2020, arXiv preprint arXiv:2007.05756.

[125] H. Huang, S. Saito, Y. Kikuchi, E. Matsumoto, W. Tang, P.S. Yu, Addressing class imbalance in scene graph parsing by learning to contrast and score, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[126] M. Suhail, A. Mittal, B. Siddiquie, C. Broaddus, J. Eledath, G. Medioni, L. Sigal, Energy-based learning for scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13936–13945.

[127] K. Tang, Y. Niu, J. Huang, J. Shi, H. Zhang, Unbiased scene graph generation from biased training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3716–3725.

[128] T. Fukuzawa, A problem reduction approach for visual relationships detection, 2018, arXiv preprint arXiv:1809.09828.

[129] Y. Qiang, Y. Yang, Y. Guo, T.M. Hospedales, Tensor composition net for visual relationship prediction, 2020, arXiv preprint arXiv:2012.05473.

[130] L. Li, J. Xiao, H. Shi, W. Wang, J. Shao, A.-A. Liu, Y. Yang, L. Chen, Label semantic knowledge distillation for unbiased scene graph generation, IEEE Trans. Circuits Syst. Video Technol. (2023).

[131] X. Qian, Y. Zhuang, Y. Li, S. Xiao, S. Pu, J. Xiao, Video relation detection with spatio-temporal graph, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 84–93.

[132] Y.-H.H. Tsai, S. Divvala, L.-P. Morency, R. Salakhutdinov, A. Farhadi, Video relationship reasoning using gated spatio-temporal energy graph, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10424–10433.

[133] J. Ji, R. Krishna, L. Fei-Fei, J.C. Niebles, Action genome: Actions as compositions of spatio-temporal scene graphs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10236–10247.

[134] C. Liu, Y. Jin, K. Xu, G. Gong, Y. Mu, Beyond short-term snippet: Video relation detection with spatio-temporal global context, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10840–10849.

[135] J. Yang, W. Peng, X. Li, Z. Guo, L. Chen, B. Li, Z. Ma, K. Zhou, W. Zhang, C.C. Loy, et al., Panoptic video scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18675–18685.

[136] S. Nag, K. Min, S. Tripathi, A.K. Roy-Chowdhury, Unbiased scene graph generation in videos, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22803–22813.

[137] F. Xiao, Y.J. Lee, Video object detection with an aligned spatial-temporal memory, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 485–501.

[138] M. Shvets, W. Liu, A.C. Berg, Leveraging long-range temporal relationships between proposals for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9756–9764.

[139] H. Wu, Y. Chen, N. Wang, Z. Zhang, Sequence level semantics aggregation for video object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9217–9225.

[140] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6154–6162.

[141] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, 2019, arXiv preprint arXiv:1904.07850.

[142] K. Kang, W. Ouyang, H. Li, X. Wang, Object detection from video tubelets with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 817–825.

[143] X. Shang, T. Ren, J. Guo, H. Zhang, T.-S. Chua, Video visual relation detection, in: Proceedings of the 25th ACM International Conference on Multimedia, 2017, pp. 1300–1308.

[144] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, T.-S. Chua, Annotating objects and relations in user-generated videos, in: Proceedings of the 2019 on International Conference on Multimedia Retrieval, 2019, pp. 279–287.

[145] X. Shang, J. Xiao, D. Di, T.-S. Chua, Relation understanding in videos: A grand challenge overview, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2652–2656.

[146] S. Zheng, X. Chen, S. Chen, Q. Jin, Relation understanding in videos, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2662–2666.

[147] X. Sun, T. Ren, Y. Zi, G. Wu, Video visual relation detection via multi-modal feature fusion, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2657–2661.

[148] Y. Teng, L. Wang, Z. Li, G. Wu, Target adaptive context aggregation for video scene graph generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13688–13697.

[149] Y. Cong, W. Liao, H. Ackermann, B. Rosenhahn, M.Y. Yang, Spatial-temporal transformer for dynamic scene graph generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16372–16382.

[150] B. Yang, W. Luo, R. Urtasun, Pixor: Real-time 3d object detection from point clouds, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7652–7660.

[151] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, A. El Sallab, Yolo3d: End-to-end real-time 3d oriented object bounding box detection from lidar point cloud, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 1–12.

[152] S. Shi, X. Wang, H. Li, Pointrcnn: 3d object proposal generation and detection from point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 770–779.

[153] S. Shi, Z. Wang, J. Shi, X. Wang, H. Li, From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network, IEEE Trans. Pattern Anal. Mach. Intell. 43 (8) (2020) 2647–2664.

[154] M. Feng, H. Hou, L. Zhang, Z. Wu, Y. Guo, A. Mian, 3D spatial multimodal knowledge accumulation for scene graph prediction in point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9182–9191.

[155] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 652–660.

[156] D. Rethage, J. Wald, J. Sturm, N. Navab, F. Tombari, Fully-convolutional point networks for large-scale point clouds, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 596–611.

[157] M. Jaritz, T.-H. Vu, R.d. Charette, E. Wirbel, P. Pérez, Xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12605–12614.

[158] P. Gay, J. Stuart, A. Del Bue, Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning, in: Asian Conference on Computer Vision, 2018, pp. 330–346.

[159] U.-H. Kim, J.-M. Park, T.-J. Song, J.-H. Kim, 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents, IEEE Trans. Cybern. 50 (12) (2019) 4921–4933.

[160] J. Wald, H. Dhamo, N. Navab, F. Tombari, Learning 3d semantic scene graphs from 3d indoor reconstructions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3961–3970.

[161] Z. Wang, B. Cheng, L. Zhao, D. Xu, Y. Tang, L. Sheng, VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D semantic scene graph prediction in point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 21560–21569.

[162] J. Bae, D. Shin, K. Ko, J. Lee, U.-H. Kim, A survey on 3D scene graphs: Definition, generation and application, in: International Conference on Robot Intelligence Technology and Applications, Springer, 2022, pp. 136–147.

[163] P. Zhang, X. Ge, J. Renz, Support relation analysis for objects in multiple view RGB-D images, in: International Joint Conference on Artificial Intelligence, 2019, pp. 41–61.

[164] A. Rosinol, A. Gupta, M. Abate, J. Shi, L. Carlone, 3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans, 2020, arXiv preprint arXiv:2002.06289.

[165] B. Zhuang, Q. Wu, C. Shen, I. Reid, A. van den Hengel, HCVRD: A benchmark for large-scale human-centered visual relationship detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[166] Y. Liang, Y. Bai, W. Zhang, X. Qian, L. Zhu, T. Mei, Vrr-vg: Refocusing visually-relevant relationships, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10403–10412.

[167] A. Zareian, S. Karaman, S.-F. Chang, Weakly supervised visual semantic parsing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3736–3745.

[168] K. Yang, O. Russakovsky, J. Deng, Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2051–2060.

[169] A. Belz, A. Muscat, P. Anguill, M. Sow, G. Vincent, Y. Zinessabah, Spatialvoc2k: A multilingual dataset of images with annotations and features for spatial relations between objects, in: Proceedings of the 11th International Conference on Natural Language Generation, 2018, pp. 140–145.

[170] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[171] B. Thomee, D.A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li, The new data and new challenges in multimedia research, 1 (2015) arXiv preprint arXiv:1503.01817.

[172] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.

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

[174] X. Sun, Y. Zi, T. Ren, J. Tang, G. Wu, Hierarchical visual relationship detection, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 94–102.

[175] M. Klawonn, E. Heim, Generating triples with adversarial networks for scene graph construction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[176] D.A. Hudson, C.D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6700–6709.

[177] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[178] N. Rai, H. Chen, J. Ji, R. Desai, K. Kozuka, S. Ishizaka, E. Adeli, J.C. Niebles, Home action genome: Cooperative compositional action understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11184–11193.

[179] F. Xia, A.R. Zamir, Z. He, A. Sax, J. Malik, S. Savarese, Gibson env: Real-world perception for embodied agents, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9068–9079.

[180] J. Yang, Y.Z. Ang, Z. Guo, K. Zhou, W. Zhang, Z. Liu, Panoptic scene graph generation, in: European Conference on Computer Vision, Springer, 2022, pp. 178–196.

[181] F. Plesse, A. Ginsca, B. Delezoide, F. Prêteux, Learning prototypes for visual relationship detection, in: 2018 International Conference on Content-Based Multimedia Indexing (CBMI), 2018, pp. 1–6.

[182] J. Lv, Q. Xiao, J. Zhong, Avr: Attention based salient visual relationship detection, 2020, arXiv preprint arXiv:2003.07012.

[183] J. Zhang, K.J. Shih, A. Elgammal, A. Tao, B. Catanzaro, Graphical contrastive losses for scene graph parsing, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11535–11543.

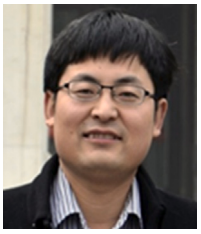
[184] B. Knyazev, H. de Vries, C. Cangea, G.W. Taylor, A. Courville, E. Belilovsky, Graph density-aware losses for novel compositions in scene graph generation, 2020, arXiv preprint arXiv:2005.08230.

[185] Y. Cong, W. Liao, B. Rosenhahn, M.Y. Yang, Learning similarity between scene graphs and images with transformers, 2023, arXiv preprint arXiv:2304.00590.

[186] B.A. Plummer, A. Mallya, C.M. Cervantes, J. Hockenmaier, 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.
[187] J. Peyre, J. Sivic, I. Laptev, C. Schmid, Weakly-supervised learning of visual relations, in: Proceedings of the Ieee International Conference on Computer Vision, 2017, pp. 5179–5188.
[188] H. Zhang, Z. Kyaw, J. Yu, S.-F. Chang, Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4233–4241.
[189] Z. Cui, C. Xu, W. Zheng, J. Yang, Context-dependent diffusion network for visual relationship detection, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 1475–1482.
[190] B. Li, Y. Wang, Visual relationship detection using joint visual-semantic embedding, in: 2018 24th International Conference on Pattern Recognition (ICPR), 2018, pp. 3291–3296.
[191] Y. Zhu, S. Jiang, Deep structured learning for visual relationship detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol.32, 2018.
[192] H. Ben-Younes, R. Cadene, N. Thome, 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.
[193] Y. Bin, Y. Yang, C. Tao, Z. Huang, J. Li, H.T. Shen, Mr-net: Exploiting mutual relation for visual relationship detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 8110–8117.
[194] W. Wang, R. Liu, M. Wang, S. Wang, X. Chang, Y. Chen, Memory-based network for scene graph with unbalanced relations, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2400–2408.
[195] W. Wang, R. Wang, S. Shan, X. Chen, Sketching image gist: Human-mimetic hierarchical scene graph generation, in: European Conference on Computer Vision, 2020, pp. 222–239.
[196] M. Wei, C. Yuan, X. Yue, K. Zhong, Hose-net: Higher order structure embedded network for scene graph generation, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1846–1854.
[197] D. Jin, X. Ma, C. Zhang, Y. Zhou, J. Tao, M. Zhang, H. Zhao, S. Yi, Z. Li, X. Liu, et al., Towards overcoming false positives in visual relationship detection, 2020, arXiv preprint arXiv:2012.12510.
[198] A. Newell, J. Deng, Pixels to graphs by associative embedding, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.
[199] S. Woo, D. Kim, D. Cho, I.S. Kweon, Linknet: Relational embedding for scene graph, in: Advances in Neural Information Processing Systems, Vol. 31, 2018.
[200] R. Herzig, M. Raboh, G. Chechik, J. Berant, A. Globerson, Mapping images to scene graphs with permutation-invariant structured prediction, Adv. Neural Inf. Process. Syst. 31 (2018).
[201] G.S. Kenigsfield, R. El-Yaniv, Leveraging auxiliary text for deep recognition of unseen visual relationships, 2019, arXiv preprint arXiv:1910.12324.
[202] M. Khademi, O. Schulte, Deep generative probabilistic graph neural networks for scene graph generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 11237–11245.
[203] C. Yuren, H. Ackermann, W. Liao, M.Y. Yang, B. Rosenhahn, NODIS: Neural Ordinary Differential Scene understanding, 2020, arXiv preprint arXiv:2001.04735.
[204] G. Ren, L. Ren, Y. Liao, S. Liu, B. Li, J. Han, S. Yan, Scene graph generation with hierarchical context, IEEE Trans. Neural Netw. Learn. Syst. 32 (2) (2020) 909–915.
[205] S. Inuganti, V.N. Balasubramanian, Assisting scene graph generation with self-supervision, 2020, arXiv preprint arXiv:2008.03555.
[206] H. Tian, N. Xu, A.-A. Liu, Y. Zhang, Part-aware interactive learning for scene graph generation, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3155–3163.
[207] T. He, L. Gao, J. Song, J. Cai, Y.-F. Li, Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation, 2020, arXiv preprint arXiv:2006.07585.
[208] R. Li, S. Zhang, B. Wan, X. He, Bipartite graph network with adaptive message passing for unbiased scene graph generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11109–11119.
[209] D. Di, X. Shang, W. Zhang, X. Yang, T.-S. Chua, Multiple hypothesis video relation detection, in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), IEEE, 2019, pp. 287–291.
[210] C. Zhang, J. Yu, Y. Song, W. Cai, Exploiting edge-oriented reasoning for 3d point-based scene graph analysis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9705–9715.
[211] S. Zhang, A. Hao, H. Qin, et al., Knowledge-inspired 3d scene graph prediction in point cloud, Adv. Neural Inf. Process. Syst. 34 (2021) 18620–18632.
[212] Y. Guo, J. Song, L. Gao, H.T. Shen, One-shot scene graph generation, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3090–3098.
[213] V.S. Chen, P. Varma, R. Krishna, M. Bernstein, C. Re, L. Fei-Fei, Scene graph prediction with limited labels, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2580–2590.
[214] J. Peyre, I. Laptev, C. Schmid, J. Sivic, Detecting unseen visual relations using analogies, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1981–1990.
[215] J. Yu, Y. Chai, Y. Wang, Y. Hu, Q. Wu, Cogtree: Cognition tree loss for unbiased scene graph generation, 2020, arXiv preprint arXiv:2009.07526.
[216] X. Yang, H. Zhang, J. Cai, Shuffle-then-assemble: Learning object-agnostic visual relationship features, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 36–52.
[217] S. Yan, C. Shen, Z. Jin, J. Huang, R. Jiang, Y. Chen, X.-S. Hua, Pcpl: Predicate-correlation perception learning for unbiased scene graph generation, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 265–273.
[218] T.-J.J. Wang, S. Pehlivan, J. Laaksonen, Tackling the unannotated: Scene graph generation with bias-reduced models, 2020, arXiv preprint arXiv:2008.07832.

Hongsheng Li is now studying for Ph.D. at Xidian University. His research interest is human-object interaction.

Guangming Zhu received the Ph.D. degree in instrument science and technology from Zhejiang University, China, in March 2015. He is currently associate Professor at Computer Science and Technology, Xidian University. His major research fields are information fusion, human action gesture recognition, scene recognition, and deep learning.



Liang Zhang professor, doctoral supervisor, studied for bachelor's, master's and doctoral degrees in biomedical engineering and instrument science at Zhejiang University from 1999.09 to 2009.09, visiting scholar at the University of Western Australia, leader of the first "Three Good and Three Have" tutorial team at Xidian University, employed as the first engineering elite talent title at Xidian University, key R&D plan evaluation expert of the Ministry of Science and Technology, project evaluation expert of Shanghai Science and Technology Commission. The main research directions include deep neural network structure design, intelligent environment perception system, human–computer interaction technology research, and embedded intelligent terminal system development. He has published more than 100 papers as the first author and corresponding author in top journals and international conferences in the field, such as IEEE TNNLS, IEEE TMI, IEEE TMM, IEEE TIP, NeurIPS, CVPR, ICCV, WWW, etc., authorized more than 20 patents, won the second prize of Shaanxi Provincial Science and Technology Progress Award (ranked second), and the first prize of Shaanxi Provincial Higher Education Technology Award (ranked second). He has presided over and participated in several projects such as the National Natural Science Foundation of China, the National Key R&D Program, etc. He serves as the guest editor of well-known SCI journals such as Neurocomputing, Sensors, etc., serves as the general chair of the first International Conference on Human–Computer Interaction and Software Engineering (HCISE) in 2022, and has published 5 monographs and textbooks.

Youliang Jiang received a master's degree in software engineering in 2022. Her main research direction is 2D scene graph generation.

Yixuan Dang received a master's degree in software engineering in 2022. Her main research direction is 2D scene graph generation.

Haoran Hou is pursuing a master's degree in software engineering at Xidian University, and his research direction is scene graph generation from 3D point clouds.

Peiyi Shen received the Ph.D. degree from Xidian University, in 1999, and the Ph.D. degree from the MTRC, Computer Science, University of Bath. He was a Research Officer with the MTRC, Computer Science, University of Bath, under the supervision of Prof. P. Willis, and a Research Fellow with CVSSP, University of Surrey, under the supervision of Prof. A. Hilton. He was with Agilent Technologies, USA, U.K., Malaysia, and Singapore, from 2000 to 2003. He is currently a Professor at the Computer Science and Technology, Xidian University. His research interests are in computer vision, volume visualization.

Xia Zhao is a professor at the Xidian University, whose main research direction is machine learning and graph convolution in action recognition.

Syed Afaq Ali Shah obtained his Ph.D. from the University of Western Australia in the area of computer vision and machine learning. He is currently working as a research associate in school of computer science and software engineering, the University of Western Australia, Crawley, Australia. He has been awarded Start Something Prize for Research Impact through Enterprise for 3D facial analysis project funded by the Australian Research Council. His research interests include deep learning, 3D object/face recognition, 3D modeling, and image processing.

Mohammed Bennamoun lectured in robotics at Queen, and then joined QUT in 1993 as an associate lecturer. He then became a lecturer in 1996 and a senior lecturer in 1998 at QUT. In January 2003, he joined The University of Western Australia as an associate professor. He was also the director of a research center from 1998–2002. He is the coauthor of the book Object Recognition: Fundamentals and Case Studies (Springer-Verlag, 2001). He has published close to 100 journals and 250 conference publications. His areas of interest include control theory, robotics, obstacle avoidance, object recognition, artificial neural networks, signal/image processing, and computer vision. He is currently a W/Professor at the School of Computer Science and Software Engineering at The University of Western Australia.