Scene Graphs: A Survey of Generations and Applications

Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann

Abstract—Scene graph is a structured representation of a scene that can clearly express the objects, attributes, and relationships between objects in the scene. As computer vision technology continues to develop, people are no longer satisfied with simply detecting and recognizing objects in images; instead, people look forward to a higher level of understanding and reasoning about visual scenes. For example, given an image, we want to not only detect and recognize objects in the image, but also know the relationship between objects (visual relationship detection), and generate a text description (image captioning) based on the image content. Alternatively, we might want the machine to tell us what the little girl in the image is doing (Visual Question Answering (VQA)), or even remove the dog from the image and find similar images (image editing and retrieval), etc. These tasks require a higher level of understanding and reasoning for image vision tasks. The scene graph is just such a powerful tool for scene understanding. Therefore, scene graphs have attracted the attention of a large number of researchers, and related research is often cross-modal, complex, and rapidly developing. However, no relatively systematic survey of scene graphs exists at present. To this end, this survey conducts a comprehensive investigation of the current scene graph research. More specifically, we first summarized the general definition of the scene graph, then conducted a comprehensive and systematic discussion on the generation method of the scene graph (SGG) and the SGG with the aid of prior knowledge. We then investigated the main applications of scene graphs and summarized the most commonly used datasets. Finally, we provide some insights into the future development of scene graphs. We believe this will be a very helpful foundation for future research on scene graphs.

Index Terms—Scene Graph, Visual Feature Extraction, Prior Information, Visual Relationship Recognition

1 Introduction

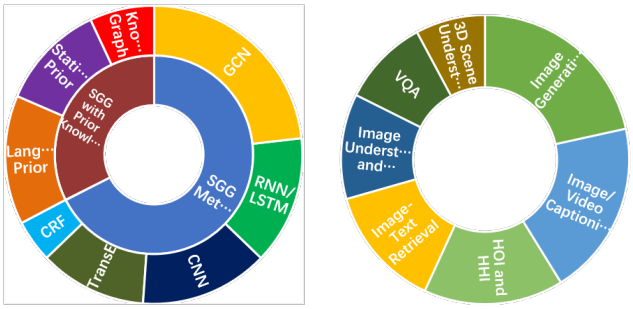
At present, deep learning [1], [2], [3], [4], [5], [6], [7] has substantially promoted the development of computer vision, such that people are no longer satisfied with simple visual understanding tasks such as target detection and recognition. Higher-level visual understanding and reasoning tasks are often required to capture the relationship between objects in the scene as a driving force. For this reason, the scene graph was first developed. Scene graphs were first proposed [8] as a data structure that describes the object instances in a scene and the relationships between these objects. A complete scene graph is able to represent the detailed semantics of a dataset of scenes, but not a single image or a video; moreover, it has powerful representations that encode 2D/3D images [8], [9] and videos [10], [11] into their abstract semantic elements without restricting either the types and attributes of objects or the relationships between objects. Related research into scene graphs greatly promotes the understanding of tasks such as vision, natural language, and their cross-domains.

As early as 2015, the idea of utilizing the visual features of different objects contained in the image and the relationships between them was proposed as a means of achieving the visual tasks of action recognition [12], image captioning [13] and other relevant computer vision tasks [14]. This type of visual relationship mining has been found to significantly improve the performance of related visual tasks, as well as to effectively enhance people's ability to understand and reason about visual scenes. Subsequently, this visual relationship was incorporated into scene graph theory in a paper by Johnson et al. [8], in which the definition of scene graphs was formally provided. In [8], a scene graph is generated manually from a dataset of real-world scene graphs, enabling the detailed semantics of a scene to be captured. Since then, the research on scene graphs has received extensive attention [15], [16], [17], [18]. When the scene graph concept was first proposed in 2015, it was initially applied to image retrieval; since then, the amount and scope of research into scene graphs has increased significantly.

In these research results, we primarily review scene graph generation (SGG) methods and the applications of the scene graph. More specifically, multiple scene graph datasets [19], [20], [21], [22], have been published since the method's inception, many scene graph generation methods based on these datasets have subsequently been proposed. These methods can be simply divided into SGG methods and SGG methods with prior knowledge. At present, the SGG methods mainly include CRF-based (conditional random field) SGG [8], [23], [24], TransE-based (visual translation embedding) SGG [25], [26], [27], CNN-based SGG [28], [29], [30], RNN/LSTM-based SGG [31], [32], [33], and graph-based SGG [34], [35], [36], among others. In addition, different types of prior information have been introduced for SGG, such as language priors [19], [37], [38], statistical priors [23], [39], knowledge graphs [23], [39], and so on. Fig. 1(a) presents the relevant works on SGG; as can be seen

- *X. Chang is with Department of Data Science and AI, Faculty of Information Technology, Monash University.*
- *P. Ren, P. Xu, and X. Chen are with School of Information Science and Technology, Northwest University.*
- *Z. Li is with Shandong Artificial Intelligence Institute, Qilu University of Technology.*
- *A. Hauptmann is with School of Computer Science, Carnegie Mellon University.*

Manuscript received March 8, 2021.



(a) SGG methods (b) SGG apllications

Fig. 1. SGG overall statistics. (a) The classification and statistics of SGG methods. (b) The classification and statistics of SGG applications.

from the figure, most research focuses on SGG that utilizes GCN (Graph Convolutional Network) [40]. Scene graph can provide powerful representations for the semantic features of a scene, and has thus been widely applied to related visual tasks, such as visual-textual transformers [15], [41], [42], [43], [44], [45], image-text retrieval [8], [46], visual question answering (VQA) [47], [48], [49], image understanding and reasoning [50], [51], 3D scene understanding [9], [17], [52], human-object interaction (HOI) and human-human interaction (HHI) [53], [54], among others (see Fig. 1(b)). We can accordingly determine that scene graphs have become a hot research topic in the field of vision, and will likely to continue to receive attention in the future.

1.1 Definition

Fig.2 presentes the overall process of building a scene graph. As shown in Fig.2 (bottom), the object instance in the scene graph can be a person (girl), a place (tennis court), a thing (shirt), or parts of other objects (arms). Attributes are used to describe the state of the current object, which may include its shape (a racket is a long strip), color (girl's clothes are white), and pose (a girl who is standing). Relations are used to describe the connection between pairs of objects, such as actions (e.g. girl swinging racket), and positions (cone placed in front of a girl). Formally, the scene graph G is a directed graph data structure; this can be defined as a tuple $G = (O, E)$, where $O = o_1, ..., o_N$ is the set of objects detected in the images. Each object has the form $o_i = (c_i, A_i)$, where c_i and A_i represent the category and attributes of the object respectively. Moreover, $E \subseteq O \times R \times O$ is a set of directed edges that repress the relationships between objects. This relationship is usually expressed as a $\langle subject - predicate - object \rangle$ triplet. At present, a scene graph is typically associated with an image dataset, but not with only one image; thus, it can be considered as representing a visual understanding of relevant images.

1.2 Construction Process

As shown in Fig. 2 (left), from the perspective of the scene graph generation process, the generations of scene graphs can be currently divided into two types [35]. The first approach has two stages, namely object detection and pairwise relationship recognition [19], [23], [39], [55]. The first stage involved in identifying the categories and attributes of the detected objects is typically achieved using Faster-RCNN [56]. This method is referred to as the *bottom-up* method. The other approach involves jointly detecting and recognizing the objects and their relationships [30], [34], [57]. This method is referred to as the *top-down* method. At the high level, the inference tasks and other visual tasks including recognizing objects, predicting the objects' coordinates, and detecting/recognizing pairwise relationship predicates between objects [58]. Therefore, most current work focuses on the key challenge of reasoning the visual relationship.

1.3 Challenge

Notably, however, the research into scene graphs still faces a number of challenges. At present, scene graph research focuses primarily on trying to solve the following two problems:

- *Accuracy of SGG.* The key question here is that of how to generate a more accurate and complete scene graph. Different learning models have a crucial impact on the accuracy and completeness of the scene graph generated by mining visual text information. Accordingly, the study of related learning models is very necessary for SGG.
- *The introduction of prior knowledge.* In addition to fully mining the objects in the current training set, along with their relationships, some additional prior knowledge is also crucial to the scene graph construction. Another important issue is that of how to make full use of this existing prior knowledge.

In response to the first question, a large number of methods for generating more accurate and complete scene graphs have been proposed, which are systematically discussed in Section 2; moreover, the introduction of additional prior knowledge is detailed in Section 3. Next, in Section 4, we compile a detailed summary of various applications of scene graphs. In Section 5, we summarized the relevant information about the datasets commonly used in SGG. In Section 6, we look forward to assess the future direction of SGG. Finally, we present our concluding remarks in Section 7.

2 Scene Graph Generation

A scene graph is a topological representation of a scene, the primary goal of which is to encode object and their relationships. The task of scene graph generation (SGG) involves constructing a graph structure that is best able to associate its nodes and edges with the objects and their relationships in a scene. Moreover, the key challenge task is to detect/recognize the relationships between the objects. The concept of the scene graph, which was first proposed by Johnson in [8], manually established the corresponding scene graph on the RW-SGD (Real-World Scene Graphs Dataset). However, it generating a scene graph manually is highly costly, while subjective factors exert an influence on the understanding of a scene. Therefore, automatically building an accurate and complete scene graph is greatly helpful to the understanding of related visual tasks.

Scene graph generation methods can be roughly divided into CRF-based SGG, TransE-based SGG, CNN-based SGG,

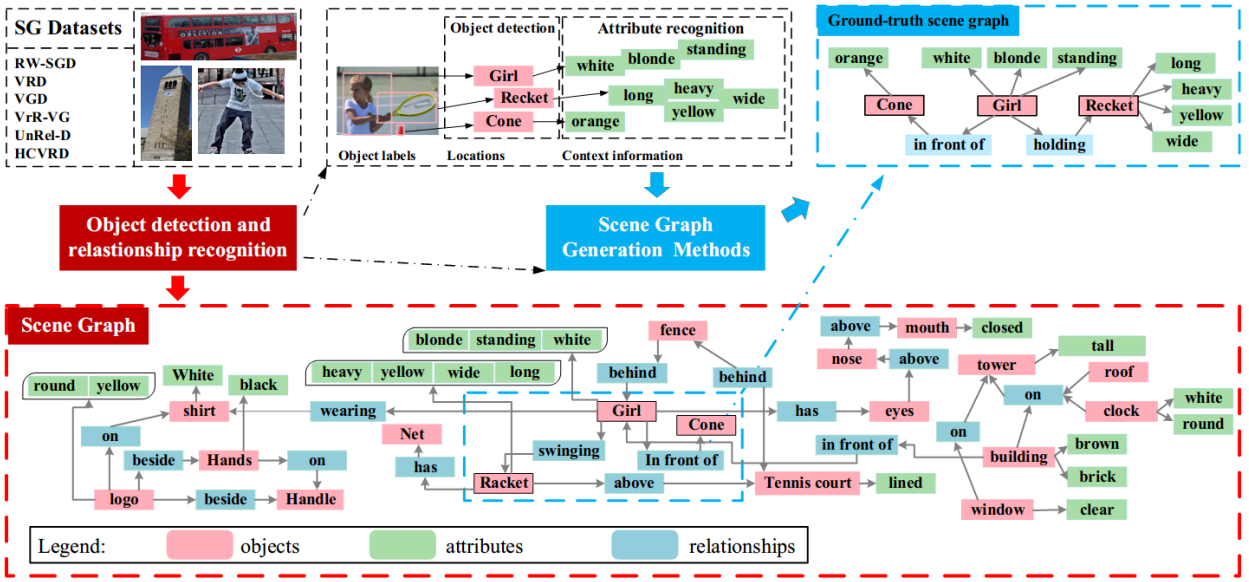


Fig. 2. An example of scene graph construction. Middle right: The ground-truth scene graph of a given image. Bottom: An example of a complete scene graph.

RNN/LSTM-based SGG, and Graph-based SGG. In this section, we conduct a detailed review of each of these methods.

2.1 CRF-based SGG

In the visual relationship triples $\langle s - r - o \rangle$, a strong statistical correlation exists between the relationship predicate and the object pair. Effective use of this information can greatly aid in the recognition of visual relationships. The CRF (Conditional Random Field) [59] is such a classical tool capable of incorporating statistical relationships into the discrimination task. CRF has been widely used in various graph inference tasks, including image segmentation [60], [61], [62], named-entity recognition [63], [64] and image retrieval [8]. In the context of visual relations, the CRF can be expressed as follows:

$$p(r, s, o|x_r, x_s, x_o) = \frac{1}{Z} exp(\Phi(r, s, o|x_r, x_s, x_o; W)), \quad (1)$$

where x_r refers to the appearance feature and spatial configuration of the object pair, x_s and x_o denote the appearance features of the subject and object respectively, W is the parameter of the model, Z is a normalization constant, and Φ represents the joint potential. Similar CRFs are widely utilized in computer vision tasks [62], [65] and have been proven effective in capturing statistical correlations in visual relationships. CRF-based scene graph generation is thus also a very valuable research direction.

Inspired by the success of deep neural networks [66], [67] and CRF [59] models, in order to explore statistical relationships in the context of visual relationships, DR-Net (Deep Relational Network) [23] opts to incorporate statistical relationship modeling into the deep neural network framework. DR-Net unrolls the inference of relational modeling into a feedforward network. In addition, DR-Net is obviously different from previous CRFs. More specifically, the statistical inference procedure in DR-Net is embedded in a deep relational network through iteration unrolling. The performance of the improved DR-Net is not only superior to classification-based methods, but also better than deep potential-based CRFs. Furthermore, SG-CRF (Scene Graph Generation via Conditional Random Fields) [24] works have observed that some previous methods [19], [20], [57], [68] tend to ignore the semantic compatibility (that is, the likelihood distribution of all 1-hop neighbor nodes of a given node) between instances and relationships, which results in a significant decrease in the model performance when faced with real-world data. For example, this may cause the model to incorrectly recognize $\langle dog - sitting\ inside - car \rangle$ as $\langle dog - driving - car \rangle$. Moreover, these models ignore the order of the two, leading to confusion between subject and object, which may produce absurd predictions such as $\langle car - sitting\ inside - dog \rangle$. In order to solve these problems, an end-to-end scene graph constructed via conditional random fields was proposed by SG-CRF to improve the quality of scene graph generation. More specifically, in order to learn the semantic compatibility of nodes in the scene graph, SG-CRF proposes a new semantic compatibility network based on conditional random fields. In order to distinguish between the subject and object in the relationship, SG-CRF proposes an effective relation sequence layer that can capture the subject and object sequence in the visual relationship.

In general, the CRF-based scene graph generation can effectively model the statistical correlation in the visual relationship, which aids in more accurately identifying the visual relationship. Although not many related works have been published on this subject, this statistically relevant information modeling is remains a classic tool in visual relationship recognition tasks.

2.2 TransE-based SGG

Since a relationship is a combination of objects and predicates, the complexity of this relationship is $\mathcal{O}(N^2R)$ for N objects and R predicates. Even if learning the models for the object and predicate separately reduces the complexity to $\mathcal{O}(N + R)$, the dramatic changes in the appearance of the predicate remain very challenging (for example, there is a significant difference in the visual appearance of

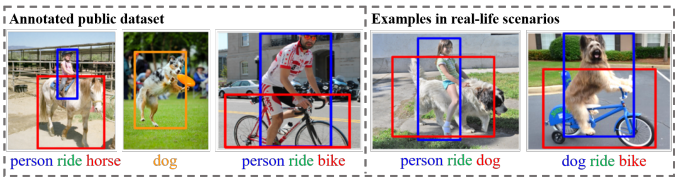


Fig. 3. Examples of the sparsity and variability of visual relationships. The left side presents examples of annotated visual relationships available in the training set, while the right side shows examples of visual relationships that are rare in the real world or test set but do actually exist.

$\langle person - ride - bike \rangle$ and $\langle person - ride - horse \rangle$. Fig. 3 presents an example of this case). The distribution of such object-predicate combinations tends to be more long-tailed than that of objects alone. In addition, the knowledge graph is similar to the scene graph; it also has a large number of fact triples, and these multi-relational data are denoted in the form $\langle head\ entity\ type - relation - tail\ entity\ type \rangle$. Knowledge graphs represent learning to embed triples into low-dimensional vector spaces, with TransE (Translation Embedding)-based models having been proven particularly effective. Furthermore, TransE (Translation Embedding) [69] regards the relationship as a translation between the head entity and the tail entity. This is in fact still helpful for learning the visual relationship representation in the scene graph.

Inspired by the advances made by TransE in the relational representation learning of knowledge bases [69], [70], VTransE [25] (based on TransE) explored how visual relations could be modeled by mapping the features of objects and predicates in low-dimensional space. More specifically, VTransE [25] is the first TransE-based SGG method that works by extending TransE networks [69]. VTransE maps entities and predicates into a low-dimensional embedding vector space, in which the predicate is interpreted as the translation vector between the embedded features of the subject and object's bounding box regions. The relationship is modeled as a simple vector transformation, i.e., $subject + predicate \approx object$. While this is a good start, VTransE considers only the features of the subject and the object, and not those of the predicate and context information [71], [72], despite these having been proven to be useful for the recognition of relations [73], [74]. To this end, MATransE (Multimodal Attentional Translation Embeddings) [27], an approach based on VTransE, combines the complementary nature of language and vision [19], along with an attention mechanism [75] and deep supervision [76], to propose a multimodal attention translation embedding method. MATransE designed two separate branches to deal directly with those of the predicate and the features of the subject-object, achieving good results.

In addition to the drastically changing visual appearance of the predicate, both the sparsity of the predicate representation in the training set [68], [77] and the very large predicate feature space also make the task of visual relationship detection increasingly difficult. Let us take the Stanford VRD dataset [19] as an example. This dataset contains 100 classes of objects, 70 classes of predicates, and a total of $30k$ training relationship annotations. The number of possible $\langle subject - predicate - object \rangle$ triplets is $100^2 * 70 = 700k$, which means that a large number of possible real relationships do not even have a training example. In fact, these invisible relationships should not be ignore even though they are not included in the training set. Fig. 3 presents an example of this case. However, VtransE and MATransE are not well-suited to dealing with this issue. Therefore, the detection of unseen/new relationships in scenes is very important to the building of a complete scene graph. Inspired by VTransE [25], the goal of UVTransE [26] is to improve generalization for rare or unseen relations. Based on VTransE, UVTransE also introduces a joint bounding box of subject and object to facilitate better capturing of contextual information and learns the embeddings guided by the constraint $predicate \approx union(subject, object) - (subject + object)$. UVtransE introduces the union of subject and object and uses a context-augmented translation embedding model to capture both common and rare relations in scenes. This type of exploration is highly beneficial for constructing a relatively complete scene graph. Finally, UVTransE combines the scores of the vision, language, and object detection modules to sort the predictions of the triple relationship. The architectural details of UVTRansE are illustrated in Fig. 4. In addition, to solve the same problem associated with new relationship discovery in scenes to generate incomplete scene graphs, RLSV (Representation Learning via Jointly Structural and Visual Embedding) [78] attempts to use existing scene graphs and images to predict the new relationship between two entities, enabling it to achieve scene graph completion. RLSV begins with the relevant knowledge of the knowledge graph, incorporating the characteristics of the scene graph, and proposes an end-to-end representation learning model of joint structure and visual embedding. Unlike TransE-based SGG methods, RLSV uses TransD [70] to project the entities (objects and subjects) from entity space to relation space by two mapping matrices. In RLSV, the structural information is introduced into the second branch by combining it with the visual features of objects and subjects, which is another improvement over the existing TransE-based SGG methods.

Unlike UVTransE and RLSV, which aim to find existing visual relationships but lack corresponding annotations in the image, AT (Analogies Transfer) [79] tries to detect those visual relationships that are not visible in the training set. As shown in Fig. 3, the individual entities of $\langle person - ride - dog \rangle$ and $\langle dog - ride - bike \rangle$ are available in the training set; however, their combination is not seen in the training set, or the visual relationship is extremely rare. As is evident, AT studies a more general phenomenon, specifically those unseen relationships that are visible in the training set for a single entity but not for the combination of $\langle subject - predicate - object \rangle$. The whole network model utilizes analogy transformation to compute the similarity between the unseen triplet and its similar triplets in order to estimate this unseen relationship and has achieved good results in unseen relationship detection. Compared with the commonly used TransD/TransE-based SGG methods, this SGG method using Analogies Transfer has good research prospects.

Building on the insights obtained from knowledge graph-related research, the TransE-based scene graph generation method has developed rapidly and attracted strong interest from researchers. Related research results have also

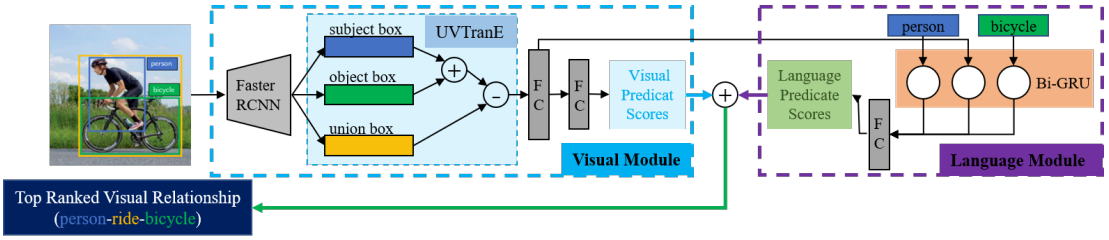


Fig. 4. The overall structure of UVTransE's [26] visual detection model. It is composed of two main parts: visual module and language module. In particular, UVTransE treats predicate embedding as $predicate \approx union(subject, object) - (subject + object)$. For comparison, VTransE [25] models visual relationships by mapping the features of objects and predicates in a low-dimensional space, where the relationship triples can be interpreted as vector translation; that is, assuming $subject + predicate \approx object$.

proven that this method is effective. In particular, the TransE-based SGG method is very helpful for the mining of unseen visual relationships, which will directly affect the integrity of the scene graph. Related research is thus still very valuable.

2.3 CNN-based SGG

CNN-based SGG methods attempt to extract the local and global visual features of the image using convolutional neural networks, then predict the relationships between the subjects and objects via classification. In these methods, the final features used for relationship identification are obtained by jointly considering the local visual features of multiple objects [58], or by carrying out feature interactions between local features [30]. As CNN performs well at extracting the visual features of the image the works related to CNN-based SGG are also very rich. In this part, we will elaborate on these CNN-based SGG methods.

The scene graph is generated by analyzing the relationship between multiple objects in the image dataset. It is accordingly necessary to consider the connection between related objects as much as possible, rather than focusing on a single object in isolation. LinkNet [58] was proposed to improve scene graph generation by explicitly modeling inter-dependency among all related objects. More specifically, LinkNet designs a simple and effective relational embedding module that jointly learns the connections between all related objects. In addition, LinkNet also introduces a global context encoding module and a geometrical layout encoding module, which extract global context information and spatial information between object proposals from the entire image and thereby further improve the performance of the algorithm. The specific LinkNet is divided into three main steps: bounding box proposal, object classification, and relationship classification. However, LinkNet considers the relation proposal of all objects, which causes it to have huge computational complexity.

On the other hand, as deep learning technology has developed, the corresponding object detection research has become increasingly mature [56], [82], [83], [84], [85], [86]. By contrast, the recognition of associations between different entities for higher-level visual task understanding has become a new challenge; this is also the key to scene graph construction. As analyzed in Section 2.2, in order to detect all relationships, it is both inefficient and unnecessary to first detect all single objects and then classify all pairs of relationships, as the visual relationship that really exists in the quadratic relationship is very sparse. Using visual

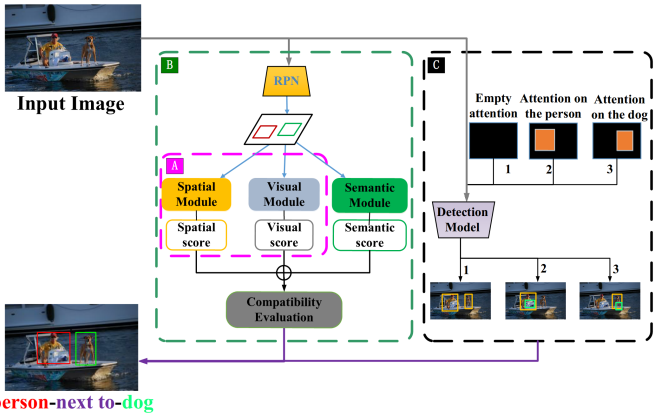


Fig. 5. Comparison of the brief schematic diagrams of three CNN-based SGG methods. (A) Rel-PN's [28] compatibility evaluation module uses two types of modules: a visual compatibility module and spatial compatibility module. The visual compatibility module is mainly used to analyzes the coherence of the appearance of the two boxes: the spatial compatibility module is primarily used to explore the locations and shapes of the two boxes. (B) IM-SGG (Interpretable Model for Scene Graph Generation) [80] adds a semantic module based on Rel-PN to capture the strong prior knowledge of the predicate. (C) BAR-CNN (Box Attention Relational CNN) [81] introduces a box attention mechanism to enhance object detection, which aids in detecting visual relationships without adding additional complex components.

phrases [68] to express this visual relationship may therefore be a good solution. Rel-PN [28] has conducted corresponding research in this direction. Similar to the region proposals of objects provided by Region Proposal Networks (RPN), Rel-PN [28] utilize a proposal selection module to select the meaningful subject-object pairs for subsequent relationship prediction. This operation will greatly reduce the computational complexity of scene graph generation. The model structure of Rel-PN is illustrated in Fig. 5(A). Furthermore, IM-SGG (Interpretable Model for Scene Graph Generation) [80], based on Rel-PN, considers three types of features, namely visual, spatial, and semantic, which are extracted by three corresponding models. Subsequently, similar to Rel-PN, these three types of features are fused for the final relationship identification. Different from Rel-PN, IM-SGG utilized an additional semantic module to learn the semantic features, and achieved better performances, (see Fig. 5(B)). This method effectively improves the interpretability of scene graph generation. More directly, using a similar method to Rel-PN, ViP-CNN (Visual Phrase Guided Convolutional Neural Network) [30] also clearly treats the visual relationship as a visual phrase containing three components. ViP-CNN [30] attempts to jointly learn the specific

Fig. 6. Schematic diagram of object detection. The use of object detection alone essentially cannot aid in the recognition of visual relationships. The interaction of features between different objects is crucial to the recognition of relationships.

visual features for the interaction to facilitate consideration of the visual dependency. In ViP-CNN, the PMPS (Phrase-guided Message Passing Structure) is proposed to model the interdependency information among local visual features using a gather-broadcast message passing flow mechanism. ViP-CNN has achieved significant improvements in speed and accuracy. In addition, to further improve the scene graph generation accuracy, some methods have also studied the interaction between different features with the goal of more accurately predicting the visual relationship between different entities. This is because the independent detection and recognition of a single object provide little assistance in fundamentally recognizing visual relationships. Fig. 6 presents an example of a case in which even the most perfect object detector finds it difficult to distinguish people standing beside horses from people feeding horses. Therefore, the information interaction between different objects is extremely important to the understanding of visual relationships. Many related works have been published on this subject. For example, the rich interactions between detected object pairs are used for visual relationship recognition in Zoom-Net [74]. Zoom-Net achieves compelling performance by successfully recognizing complex visual relationships through the use of deep message propagation and the interaction between local object features and global predicate features, without the use of any linguistic priors. VIP-CNN [30] also uses similar feature interactions. The key difference is that the CA-M (Context-Appearance Module) proposed by VIP-CNN attempts to directly fuse pairwise features to capture contextual information, while the SCA-M (Spatiality-Context-Appearance Module) proposed by Zoom-Net [74] performs spatially-aware channel-level local and global context information fusion. Therefore, SCA-M has more advantages when capturing the spatial and contextual relationships between the subject, predicate, and object features. Fig. 7 presents the structure comparison diagram of the Appearance Module (A-M) without information interaction, along with the Context-Appearance Module (CA-M) and Spatiality-Context-Appearance Module (SCA-M).

An attention mechanism is also a good tool for improving visual relationship detection. BAR-CNN (Box Attention Relational CNN) [81] observed that the receptive field of neurons in the most advanced feature extractors [67], [87] may still be limited, meaning that the model may cover the entire attention map. To this end, BAR-CNN proposes a box attention mechanism; this enables visual relationship detection tasks to use existing object detection models in order to complete the corresponding relationship recognition tasks without introducing additional complex components. This is a very interesting concept, and BAR-CNN has also obtained competitive recognition performance. A schematic illustration of BAR-CNN is presented in Fig. 5(C).

The CNN-based SGG method is very rich, and many interesting variations have been proposed. However, there are still many remaining challenges requiring further research, including how to reduce the computational complexity as much as possible while ensuring deep interaction between the triplet's different features, how to deal with the real but very sparse visual relationship in reality, etc. Identifying solutions to these problems will further deepen the research related to the CNN-based SGG method.

2.4 RNN/LSTM-based SGG

A scene graph is a structured representation of an image. The information interaction between different objects and the contextual information of these objects is crucial to the recognition of the visual relationship between them. Models based on RNN and LSTM have natural advantages in capturing the contextual information in the scene graph and reasoning on the structured information in the graph structure. RNN/LSTM-based methods are thus is also a popular research direction.

As discussed above, in order to make full use of the contextual information in the image to improve the accuracy of scene graph generation, IMP (Iterative Message Passing) [57] was proposed. IMP attempts to use standard RNN to solve the scene graph inference problem and iteratively improve the model's prediction performance through message passing. The main highlight of this approach is its novel primal-dual graph, which defines the channels for messages passing from node GRUs [88] to edge GRUs to achieve scene graph generation. This form of information interaction helps the model to more accurately identify the visual relationships between objects. Unlike cases of interaction between local information, such as IMP, MotifNet (Stacked Motif Network) [32] begins from the assumption that the strong independence assumption in the local predictor [30], [34], [57] actually limits the quality of global prediction. To this end, MotifNet encodes global context information through recurrent sequential architecture LSTMs (Long Short-term Memory Networks) [89]. However, MotifNet [32] only considers the context information between objects while failing to take scene information into account. There have also been some works [35], [57], [90] on the classification of relationships by exchanging the context between nodes and edges. However, the above-mentioned SGG methods focus primarily on the structural-semantic features in a scene while ignoring the correlations among different predicates. For this reason, [31] proposed a two-stage predicate association network (PANet). The main goal of the first stage to extract instance-level and scene-level context information, while the second stage is mainly used to capture the association between predicate alignment features. In particular, an RNN module is used in order to fully capture the association between alignment features. This kind of predicate association analysis has also achieved good results.

However, the methods discussed above often rely on object detection and predicate classification between objects.

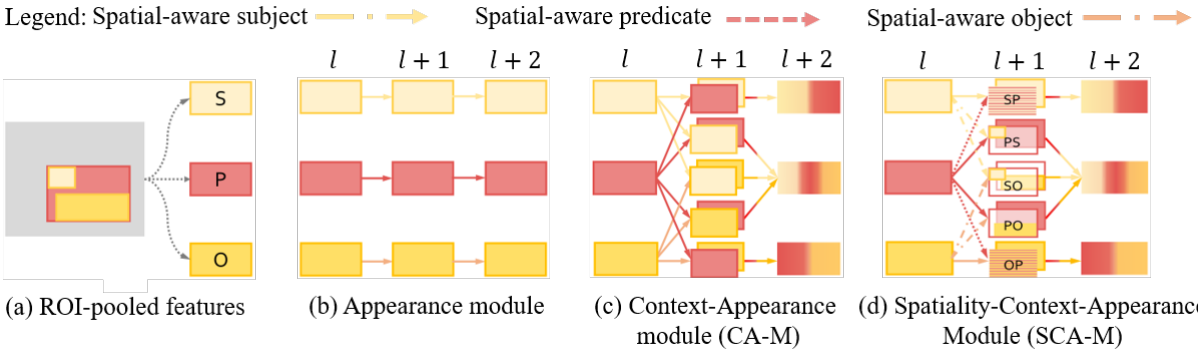


Fig. 7. (a) The ROI-pooled feature of the subject (S), predicate (P), and object (O) of a given input image. (b) Appearance Module (A-M) without information interaction. (c) Context-Appearance Module (CA-M) in ViP-CNN [30]. (d) Spatiality-Context-Appearance Module (SCA-M) in Zoom-Net [74].

There are two obvious limitations of this approach: first, the object bounding box or relationship pairs generated via the object detection method are not always necessary for the generation of the scene graph; secondly, scene graph generation depends on the probabilistic ranking of the output relationships, which will lead to semantically redundant relationships [91]. For this reason, AHRNN (attention-based hierarchical RNN) [92] proposed a hierarchical recurrent neural network based on a visual attention mechanism. This approach first uses the visual attention mechanism [93], [94] to resolve the first limitation. Secondly, AHRNN regards the recognition of relational triples as a sequence learning problem using recurrent neural networks (RNN). In particular, it employs hierarchical RNN to model relational triples in order to more effectively process long-term context information and sequence information [95], [96], thereby avoiding the need to rank the probability of output relationships.

On the other hand, VCTREE (Visual Context TREE model) [33] observed that the previous scene graphs either adopted chains [32] or a fully-connected graph [23], [35], [57], [74], [97], [98], [99]. However, VCTREE believes that these two prior structures may not be optimal, as the chain structure is too simple and may only capture simple spatial information or co-occurrence bias; moreover, the fully connected graph lacks the distinguishing structure of hierarchical and parallel relationships. In order to solve this problem, VCTREE proposed composite dynamic tree structures, which can use TreeLSTM [100] for efficient context coding and thus effectively represent the hierarchical and parallel relationships in visual relationships. This tree structure provides a new research direction for scene graph representation. Fig. 8 presents a comparison of the chain structure, fully connected graph structure, and dynamic tree structure of the scene graph. SIG (Sketching Image Gist) [101] also proposed a scene graph with a similar tree structure; the key difference stems from the observation that humans tend to describe the subjects and key relationships in the image first when analyzing scenes, meaning that a hierarchy analysis with primary and secondary order is more in line with human habits. To this end, SIG proposed a human-mimetic hierarchical scene graph generation method. Under this approach, the scene is represented by a human-mimetic HET (Hierarchical Entity Tree) composed of a series of image regions, while Hybrid-LSTM (Hybrid

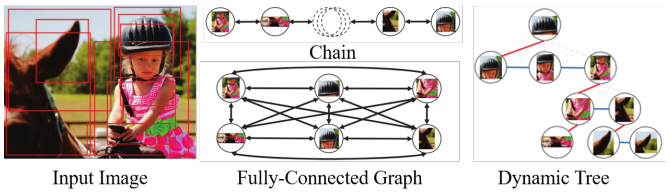


Fig. 8. A comparison of the chains [32], fully connected graph [23], [35], [57], [74], [97], [98], [99] and dynamic tree [33] structure of the scene graph. The dynamic tree structure on the left shows a left-child right-sibling binary trees, where the left branches (red) represents the hierarchical relationships, while the right branches (blue) represents the parallel relationship. Compared with chain and graph structures, dynamic tree have obvious natural advantages in the representation of hierarchical and parallel relationships.

Long Short-Term Memory) is used to parse HET, thereby enabling the hierarchical structure [33] and siblings context [32] information in HET to be obtained.

2.5 Graph-based SGG

The scene graph can be regarded as a graph structure. An intuitive approach would therefore be to improve the generation of scene graphs with the help of graph theory. The GCN [102], [103], [104] can operate on graph structure data by transmitting local information [105], [106]. GCN has been proven to be highly effective in tasks such as relational reasoning [107], graph classification [108], [109], [110], [111], node classification in large graphs [40], [112], and visual understanding [113], [114], [115]. Accordingly, many researchers have directly studied the scene graph generation method based on GCN.

As discussed in Section 1.2, the current scene graph generation methods can be roughly divided into two categories: *bottom-up* and *top-down* methods. However, types of frameworks build a quadratic number of objects, which is time-consuming. Therefore, an efficient subgraph-based framework for scene graph generation, called Factorizable Net [35], is proposed to promote the generation efficiency of the scene graph. Under this approach, the detected object region proposals are paired to facilitate the construction of a complete directed graph. Thereafter, a more precise graph is generated by merging edges corresponding to similar union regions into a sub-graph; each sub-graph has several objects and their relationships are represented as

edges. By substituting the original scene graph with these subgraphs, Factorizable Net can achieve higher generation efficiency of the scene graph. Adopting a different approach to Factorizable Net's attempt to decompose the scene graph in order to improve the efficiency of scene graph generation, Graph R-CNN [90] attempts to trim the original scene graph (removing those unlikely relationships) to generate a sparse candidate graph structure. Finally, an attention graph convolutional network (AGCN) is used to integrate global context information to promote more efficient and accurate scene graph generation.

The graph-based attention mechanism also has important research value in the generation of scene graphs. For example, previous scene graph generation work [34], [57], [90], [114], [116] often requires prior knowledge of the graph structure. In addition, these methods tend to ignore the overall structure and information of the entire image, as they capture the representation of nodes and edges in a step-by-step manner. Moreover, the one-by-one detection of the visual relationship of the paired regions [21], [30], [74], [117], [118], [119] is also poorly suited to describing the structure of the entire scene. For this reason, in ARN (Attentive Relational Network) [36], a semantic transformation module is proposed that produces semantic embeddings by transforming label embeddings and visual features into the same space, while a relation inference module is used to predict the entity category and relationship as the final scene graph result. In particular, to facilitate describing the structure of the entire scene, ARN proposed a graph self-attention-based model aimed at embedding a joint graph representation to describe all relationships. This module helps in the generation of more accurate scene graphs.

When predicting the visual relationship of the scene graph, the reading order of entities in the context encoding using RNN/LSTM [120] also has a crucial influence on the scene graph generation. A fixed reading order may not be optimal. A scene graph generator should reveal the connection between objects and relations in order to improve prediction precision, even if different types of inputs are present. Formally, given the same features, the same result should be obtained by a framework or a function $\mathcal{F}$ even if the input has been permuted. Motivated by this observation, the architecture of a neural network for SGG should ideally remain invariant to a particular type of input permutation. Herzig et al. [114] accordingly proved this property based on the fact that such an architecture or framework can gather information from the holistic graph in a permutation-invariant manner. Based on this feature, these authors proposed several common architecture structures and obtained competitive performance.

For most SGG approaches [19], [23], [30], [39], [55], [57], the long-tailed distribution of relationships remains a challenge to relational feature learning. Existing methods are often unable to deal with unevenly distributed predicates. Therefore, Dornadula et al. [121] attempted to construct a scene graph via few-shot learning of predicates, which can scale to new predicates. The SGG model based on few-shot Learning attempts to fully train the graph convolution model and the spatial and semantic shift functions on relationships with abundant data. For their part, new shift functions are fine-tuned with new, rare relationships of a few examples. When compared to conventional SGG methods, the novelty of this model is that predicates are defined as functions, such that object notations are useful for few-shot predicate forecasting; these include a forward function that turns subject notations into objects and a corresponding function that changes the object representation back into subjects. The model achieves good performance in the learning of rare predicates.

A comprehensive, accurate and coherent scene graph is what we expect to achieve, and the semantics of the same node in different visual relationships should also be consistent. However, the currently widely used supervised learning paradigm based on cross-entropy may not guarantee the consistency of this visual context. This is because this paradigm tends to predict pairwise relationships in an independent way [19], [25], [119], [122], while hub nodes (those that belong to multiple visual relationships at the same time) and non-hub nodes are given the same penalty. This is unreasonable. For this reason, [123] proposed a Counterfactual critic Multi-Agent Training (CMAT) approach. More specifically, CMAT is the first work to define SGG as a cooperative multi-agent problem. This approach solves the problems of graph consistency and graph-level local sensitivity by directly maximizing a graph-level metric as a reward (corresponding to the hub and non-hub nodes being given different penalties). Similarly, RelDN (Relationship Detection Network) [18] also found that applying only cross-entropy loss may have an adverse effect on predicate classification; for example, Entity Instance Confusion (confusion between different instances of the same type) and Proximal Relationship Ambiguity (subject-object pairing problems in different triples with the same predicate). RelDN is proposed to tackle these two problems. In RelDN, three types of features for semantic, visual, and spatial relationship proposals are combined by means of entity-wise addition. These features are then applied to obtain a distribution of predicate labels via softmax normalization. Thereafter, contrastive losses between graphs are specifically constructed to solve the aforementioned problems.

Scene graphs provide a natural representation for reasoning tasks. Unfortunately, their non-differentiable representations it difficult to use scene graphs directly as intermediate components of visual reasoning tasks. Therefore, DSG (Differentiable Scene-Graphs) [124] are employed solve the above obstacles. The visual features of objects are used as the inputs to the differentiable scene-graph generator module of DSGs, which is a set of the new node and edge features. The novelty of the DSG architecture lies in its decomposition of the scene graph components, enabling each element in a triplet to be represented by a dense descriptor. Thus, DSGs can be directly used as the intermediate representation of downstream inference tasks.

Although we have investigated many graph-based scene graph generation methods, there are still many other related methods. For example, [125] proposes a deep generative probabilistic graph neural network (DG-PGNN) to generate a scene graph with uncertainty. SGVST [126] introduces a scene graph-based method to generate story statements from image streams. This approach uses GCN to capture the local fine-grained region representation of objects in the scene graph. We can conclude from the above that the

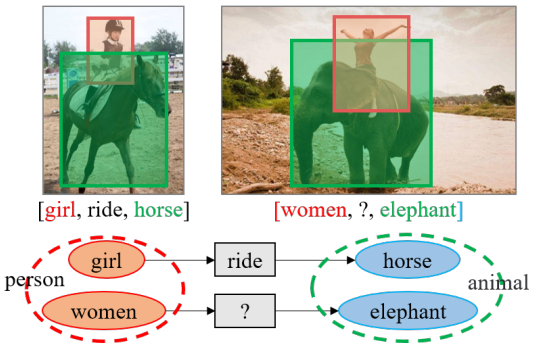


Fig. 9. Examples of semantic similarity helping with visual relationship inference. By mining object pairs with similar semantics [girl, woman] and [horse, elephant], language priors can help with the inference that the relationship between "woman" and "elephant" may be "riding". For example, if we know that [girl, woman] are subclasses of person, and [horse, elephant] are subclasses of animals, we can easily infer from [girl, ride, horse] that the relationship between "woman" and "elephant" also be "riding", even though these two relationships differ greatly in terms of their visual features.

graph-based scene graph generation method has attracted significant research attention due to its obvious ability to capture structured information.

3 SGG WITH PRIOR KNOWLEDGE

For SGG, relationships are combinations of objects, and its semantic space is wider than that of the objects. In addition, it is very difficult to exhaust all relationships from the SGG training data. It is therefore particularly critical to effectively learn relationship representations from a small amount of training data. The introduction of prior knowledge may thus greatly assist in the detection and recognition of visual relationships. Therefore, in order to efficiently and accurately generate a complete scene graph, the introduction of prior knowledge (such as language prior, visual prior, knowledge prior, context, etc.) is also crucial. In this section, we will introduce the related work of scene graph generation with prior knowledge.

3.1 SGG with Language Prior

Language priors typically use the information embedded in semantic words to fine-tune the possibility of relationship prediction, thereby improving the accuracy of visual relationship prediction. Language priors can help the recognition of visual relationships through the observation of semantically related objects. For example, horses and elephants may be arranged in semantically similar context, e.g., "a person riding a horse" and "a person riding an elephant". Therefore, although the co-occurrence of elephants and persons is not common in the training set, through the introduction of language priors and the study of the more common examples (such as "a person riding a horse"), we can still easily infer that the relationship between a person and an elephant may be one of riding. This idea is illustrated in Fig.9. This approach also helps to resolve the long tail effect in visual relationships.

Many researchers have conducted detailed studies on the introduction of language priors. For example, Lu et al. [19] suggested training a visual appearance module and a language module simultaneously, then combining the two scores to infer the visual relationship in the image. In particular, the language a priori module projects the semantic-like relationships into a tighter embedding space. This helps the model to infer a similar visual relationship ("person riding an elephant") from the "person riding a horse" example. Similarly, VRL (deep Variation-structured Reinforcement Learning) [37] and CDDN (Context-Dependent Diffusion Network) [38] also use language priors to improve the prediction of visual relationship; the difference is that [19] uses semantic word embedding [127] to fine-tune the possibility of predicting relationships, while VRL follows the variational-structured traversal scheme over a directed semantic action graph from the language prior, meaning that the latter can provide richer and more compact semantic association representation than word embedding. Moreover, CDDN finds that similar objects have close internal correlations, which can be used to infer new visual relationships. To this end, CDDN uses word embedding to obtain a semantic graph, while simultaneously constructing a spatial scene graph to encode global context interdependency. CDDN can effectively learn the latent representation of visual relations through the combination of prior semantics and visual scenes; furthermore, considering its isomorphic invariance to graphs, it can cater well to visual relation detection.

On the other hand, although the language prior can compensate for the difference between model complexity and dataset complexity, its effectiveness will also be affected when the semantic word embedding falls short [128]. For this reason, [117] further introduces a relation learning module with a priori predicate distribution on the basis of IMP [57] to better learn visual relations. In more detail, a pre-trained tensor-based relation module is added to [117] as a dense relation prior to fine-tune the relation estimation, while an iterative message-passing scheme with GRUs is used as a GCN method of promoting the scene graph generation performance with better feature representation. In addition to using language priors, [129] also combines visual cues to identify visual relationships in images and locate phrases. For its part, [129] models the appearance, size, location, and attributes of entities, along with the spatial relationship between object pairs connected by verbs or prepositions, and jointly infers visual relationships through automatically learning and combining the weights of these clues.

3.2 SGG with Statistical Prior

Statistical prior is also a form of prior knowledge widely used by SGG, as objects in the visual scene usually have strong structural regularity [32]. For example, people tend to wear shoes, while mountains tend to have water around them. In addition, $\langle cat - eat - fish \rangle$ is common, while $\langle fish - eat - cat \rangle$ and $\langle cat - ride - fish \rangle$ are very unlikely. This relationship can thus be expressed using prior knowledge of statistical correlation. Modeling the statistical correlation between object pairs and relationships can help us in correctly identifying visual relationships.

Due to the spatially large and long-tailed nature of relationship distributions, simply using the annotations contained in the training set would be insufficient. Moreover,

it is difficult to collect an adequate amount of labeled training data. For this reason, LKD (Linguistic Knowledge Distillation) [39] uses not only the annotations inside the training set, but also text publicly available on the Internet (Wikipedia) to collect external language knowledge. This is mainly achieved by tallying the vocabulary and expressions used by humans to describe the relationships between pairs of objects in the text, then calculating the conditional probability distribution ($P(pred|subj, obj)$) of the predicate given a pair of $\langle subj, obj \rangle$. A novel contribution is that of using knowledge distillation [130] to acquire prior knowledge from internal and external linguistic data in order to solve the problem of long-tail relationships.

Similarly, DR-Net (Deep Relational Networks) [23] also noticed the strong statistical correlation between the triples $\langle subj - pred - obj \rangle$. The difference is that DR-Net proposed a deep relationship network to take advantage of this statistical correlation. DR-Net first extracts the local regions and spatial masks of each pair of objects, then inputs them together with the appearance of a single object into the deep relational network for joint analysis, thereby obtaining the most likely relational category. In addition, MotifNet [32] performed a statistical analysis of the co-occurrences between the relationships and object pairs on the Visual Genome dataset [20], finding that these statistical co-occurrences can provide strong regularization for relationship prediction. To this end, MotifNet uses LSTM [89] to encode the global context of objects and relationships, thus enabling the scene graph to be parsed. However, although the above methods [23], [32] also observed the statistical co-occurrence of the triple, the depth model they designed implicitly mined this statistical information through message transmission. KERN (Knowledge-Embedded Routing Network) [131] also took note of this statistical co-occurrence. The difference is that KERN formally expresses this statistical knowledge in the form of a structured graph, which is incorporated into the deep propagation network as additional guidance. This can effectively regularize the distribution of possible relationships, thereby reducing the ambiguity of prediction.

In addition, similar statistical priors are also used in complex indoor scene analysis [132]. Statistical priors can effectively improve performance on corresponding scene analysis tasks.

3.3 SGG with Knowledge Graph

Knowledge graphs are a rich knowledge base that encode how the world is structured. Common-sense knowledge graphs have thus been used as prior knowledge to effectively help the generation of scene graphs.

To this end, GB-Net (Graph Bridging Network) [133] proposes a new perspective, which constructs scene graphs and knowledge graphs into a unified framework. More specifically, GB-Net regards the scene graph as the image-conditioned instantiation of the commonsense knowledge graph. Based on this perspective, the generation of scene graphs is redefined as a bridge mapping between scene graphs and common sense graphs. A schematic diagram of this idea is presented in Fig.10. In addition, the deviations in the existing label dataset on object pairs and relationship

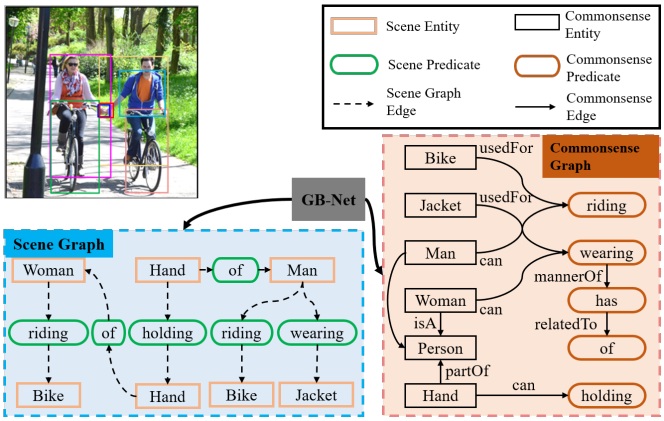


Fig. 10. Left: Sample image and truth scene graph. Right: The relevant part of the commonsense knowledge graph. GB-Net [133] defines scene graph generation as the problem of establishing a bridge between two graphs.

labels, along with the noise and missing annotations they contain, all increase the difficulty of developing a reliable scene graph prediction model. For this reason, KB-GAN (knowledge base and auxiliary image generation) [134] proposed a scene graph generation algorithm based on external knowledge and image reconstruction loss to overcome the problems found in datasets. More specifically, KB-GAN uses ConceptNet's [135] English subgraph as the knowledge graph; the knowledge-based module of KB-GAN improves the feature refinement process by reasoning on a basket of common sense knowledge retrieved from ConceptNet.

3.4 Discussion

Prior knowledge has been proven to significantly improve the quality of scene graph generation. Existing methods use either an external curated knowledge base, such as ConceptNet [134], [136], [137], [138], or the statistical information found in the annotation corpus to obtain commonsense data. However, the former is limited by incomplete external knowledge [32], [39], [131], [139], while the latter is often based on hard-coded heuristic algorithms such as the co-occurrence probability of a given category. Therefore, the latest research [140] attempts to use visual commonsense as a machine learning task for the first time, and automatically obtains visual commonsense data directly from the dataset to improve the robustness of scene understanding. While this exploration is very valuable, the question of how to acquire and make full use of this prior knowledge remains a difficult one that merits further attention.

4 Applications of Scene Graph

The scene graph can describe the objects in a scene and the relationships between the objects, meaning that it provides better representations for scene understanding-related visual and textual tasks and can greatly improve the model performance of these tasks. Fig.11 presents some examples of scene graph application scenarios. Next, we will conduct a detailed review of these scene graph applications one by one.

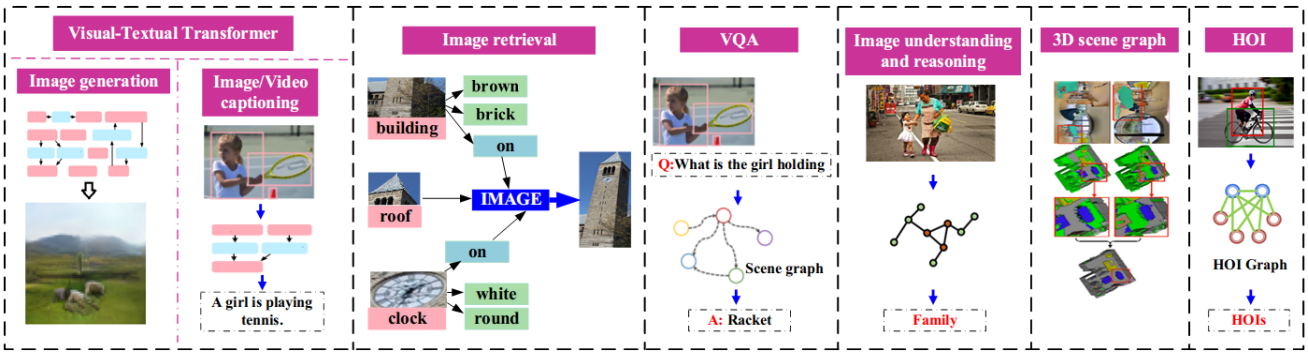


Fig. 11. Examples of scene graph application scenarios. These applications include visual-textual transformers [41], [141], [142], image text retrieval [8], [14], [143], visual question answering [48], [144], image understanding and reasoning [50], [51], [145], 3D scene graphs [9], [52], and the detection and recognition of human-object interaction and human-human interaction [22], [146], [147], [148], [149].

4.1 Visual-Textual Transformer

Scene graphs contain the structured semantic information in a visual scene, and this semantic information is mainly reflected in the representations of objects, attributes, and pairwise relationships in images. Thus, scene graphs can provide favorable reasoning information for the vision-text tasks of image generation and image/video captioning.

4.1.1 Image Generation

Although text-based image generation has made exciting progress in the context of simple textual description, it is still difficult to generate images based on complex textual descriptions containing multiple objects and their relationships. Image generation based on scene graphs is better able to deal with complex scenes with multiple objects and desired layouts.

Generating complex images from layouts is more controllable and flexible than text-based image generation. However, it remains a difficult one-to-many problem, and only limited information can be conveyed by a bounding box and its corresponding label. To generate a realistic image according to the corresponding scene graph with object labels and their relationships, Johnson et al. [41] proposed an end-to-end image generation network model. Compared with text-based image generation methods, a final complex images with many recognizable objects can be generated using this method by explicitly inferring the objects and relationships based on structured scene graphs. However, this image generation method [41] cannot introduce new or additional information to existing descriptions, and are limited to generating images at one time. Therefore, Mittal et al. proposed a recursive network architecture [141] that preserves the image content generated in previous steps and modifies the accumulated images based on newly provided scene information. This method allows the context in sequentially generated images to be preserved by subjecting certain information to subsequent image generation conditions. However, there are still problems associated with ensuring that the generated image conforms to the scene graph and measuring the performance of the image generation models. Subsequently, an image generation method was proposed that harnesses the scene graph context to improve image generation [142]. The scene context network encourages the generated images not only to appear realistic, but also to respect the scene graph relationships. Similarly, Layout2Im (Layout-based Image generation) [150] is also an end-to-end model for generating images from layouts. Different from other related methods, Layout2Im breaks down the representations of each object into specified and unspecified parts, and individual object representations are grouped together for encoding the layouts.

From [41], [141], [142], [150], we can see that generating a layout from a scene graph is an important step in layout-based image generation. Therefore, Herzig et al. [42] and Tripathi et al. [151] attempted to improve the quality of images generated from scene graphs by generating better layouts. Generating realistic images with complex visual scenes is a challenging task, especially when we want to control the layouts of the generated images. Herzig et al. [42] present a novel SRC (Soft Relations Closure) Module to inherently learn the canonical graph representations, with the weighted graph representations obtained from a GCN used to generate the scene layouts. SRC can canonicalize graphs to improve layout generation. Moreover, Tripathi et al. [151] proposed a scene layout generation system to generate structured scene layouts. Similarly, to solve the layout prediction problem, Schroeder et al. proposed a layout prediction framework based on Triplet-Aware Scene Graph Embeddings [152]. Triplet embeddings with supervisory signals are used to improve scene layout prediction, while a data augmentation technique is utilized to maximize triplet numbers during training. These two methods of additional supervision and data augmentation can enhance the embedding representation, enabling better layout outputs to be obtained.

The above image generation methods from the scene graph are based on layouts and aim at ensuring semantic consistency between generated images and scene graphs; however, the visual appearance of the obtained objects and images. PasteGAN [153] is a semi-parametric method for image generation based on scene graphs and object cropping, designed improve the visual quality of generated images. The proposed Crop Refining Network and Object-Image Fuser with attention mechanism in PasteGAN can encode objects' spatial arrangement, appearances, and interactions. Compared with most scene graph image generation methods, PasteGAN can parameterize control of the

appearance of objects in the generated images. In addition, one interesting approach to image generation utilizes image collections to generate a narrative collage based on scene graphs [154]. In this process, the object relationship is crucial to the object positions in the narrative collage. However, the scene graphs used here are primarily rule-based to build and evaluate. Furthermore, SS-SGN (Spatio-Semantic Scene Graph Network) [155] proposes a scene graph-based image modification method that can interact with users. More specifically, the user only needs to change a certain node or edge of the scene graph and then apply this change to edit the image. This provides users with more flexibility to modify or integrate new content into the image.

4.1.2 Image/Video Captioning

Traditional image captioning methods rely on the visual features of the objects detected in images. Here, the natural language description is generated by models of natural language reasoning, such as RNN or LSTM. However, these methods can not make full use of the semantic relationships between objects in images, which leads to the generated language description being inaccurate. The image captioning methods based on scene graph solve this problem to some extent by capturing the relationship information between objects.

The method proposed in [156] is an image captioning method with semantic representations that operates by embedding a scene graph as an intermediate state. This method is easy to execute, does not require complex image preprocessing, and is competitive with existing methods. Moreover, since graphical representations with conceptual positional binding can improve image captioning, TPsgtR (Tensor Product Scene-Graph-Triplet Representation) [157] is proposed for image caption generation using regional visual features. In TPsgtR, the technique of neuro-symbolic embedding can embed the relationships identified among different image regions into concrete forms (neural symbolic representations), rather than relying on the model to form all possible combinations. These neural symbolic representations aid in defining the neural symbolic space and can be transformed into better captions for images. In addition, to use the visual relations contained in scene graphs for the purpose of improving image captioning, the visual relational features are learned from a neural scene graph generator (Stacked Motif Network) [120], facilitating the grounding of language in visual relations.

From the perspective of human cognition, vision-based language generation is related to high-level abstract symbols. Abstracting the scenes into symbols will accordingly provide a clear path to language description generation. Therefore, SGAE (Scene Graph Auto-Encoder) [43] is proposed to incorporate these inductive biases into the encoder-decoder models for image captioning, an approach expected to help this encoder-decoder model exhibit less overfitting to the dataset bias. Similarly, to be able to generate the type of image descriptions desired by the human use, Chen et al. proposed an ASG (Abstract Scene Graph) structure [158] to represent user intentions, as well as to control the generated descriptions and detailed description in the scenes. ASG can identify users' intention and semantics in graphs, which enables it to generate the required caption based on the graph structure, and actively considers users' intention to produce the desired image caption, which significantly enhances the image caption diversity. Unlike ASG, which generates diversified captions by focusing on different combinations of objects in the scene graph, the core of SGD (Scene Graph Decomposition) [159] is to decompose the scene graph into a set of subgraphs, then use a deep model to select the most important subgraphs. SGD can obtain accurate, diverse, and controllable subtitles by using subgraphs.

In previous work, entities in images are often considered separately, which leads to the lack of structured information in the generated sentences. Scene graphs are structured by leveraging both visual features and semantic knowledge, and image captioning frameworks are proposed based on the structural-semantic information within an image [44], [160] and across different images [11]. In [44], an image caption framework based on scene graph is proposed to utilize the structural-semantic features in images. A hierarchical-attention-based module is designed to learn the correlation scores of the visual features and semantic relationship features, which are used to obtain the final context feature vector, instead of simply concentrating two feature vectors into a single vector. SGC (Scene Graph Captioner) [160] captures the integrated structural-semantic features from visual scenes, after which LSTM-based models translate these semantic features into the final text description. Furthermore, a scene graph can also be used to generate the story from an image stream; the proposed SGVST in [11] can model the visual relations both within one image and across images, which is conducive to image captioning. This method significantly improves the fluency and richness of the generated stories.

At present, most image captioning models rely heavily on image-caption pair dataset, while unpaired image captioning presents great challenges when it comes to extracting and mapping different features of visual and textual modalities. Therefore, there are high costs associated with obtaining large-scale paired data of images and texts. To solve this problem, an unpaired scene graph-based image captioning approach is presented in [45] to capture rich semantic information from scenes. It further proposes an unsupervised feature extraction method to learn the scene graph features by mapping from the visual features of the images to the textual features of the sentences.

4.2 Image-Text Retrieval

Image-text retrieval is a classic multi-modal retrieval task. For retrieving the target samples (images or textual sentences), the query could be the content of an image or the text describing the image. Most content-based image retrieval methods use low-level visual features. Recently, there has been increasing interest in models that jointly reason about visual and textual features. However, these models have limitations in terms of their expressiveness. For example, text-based image retrieval methods are impacted by the inherent referential uncertainty of textual representations. The scene graph is a structured representation of visual scenes that can explicitly represents the objects, attributes, and relationships in images. These structured feature representations provide a great deal of help to models when

they search the corresponding samples. Therefore, scene graph-based image-text retrieval has broad development prospects.

Image retrieval via scene graph was first applied in [8]. By replacing textual descriptions with scene graphs for image retrieval, the model can accurately describe the semantics of images without on unstructured texts, and can further retrieve related images in more open and interpretable retrieval tasks. Subsequently, another scene graph-based image retrieval method is proposed by Schuster et al [14]. These may be the two earliest methods to involve the construction and applications of the scene graph. The results of their experiments show that image retrieval using scene graphs achieves better results than traditional image retrieval methods based on low-level visual features.

At present, most existing cross-modal scene retrieval methods ignore the semantic relationship between objects in the scene and the embedded spatial layout information. Moreover, these methods adopt a batch learning strategy and are thus unsuitable for processing stream data. To solve these problems, an online cross-modal scene retrieval method [143] is proposed that utilizes binary representations and semantic graphs. The semantic graph can serve as a bridge between the scene graph and the corresponding text that enables measurement of the semantic correlation between different modal data. However, most text-based image retrieval models experience difficulties when searching large-scale image data, such that the model needs to resort to an interactive retrieval process through multiple iterations of question-answering. To solve this problem, Ramnath et al. proposed an image retrieval framework based on scene graph [161], which models the retrieval task as a learnable graph matching problem between query graphs and catalog graphs. Their approach incorporated the strategies and structural constraints of the retrieval task into inference modules using multi-modal graph representation. Similarly, SQIR (Structured Query-Based Image Retrieval Using Scene Graphs) [162] is also an image retrieval framework based on scene graphs. The difference is that SQIR determined that structured queries (e.g. "little girl riding an elephant") are more likely to capture the interaction between objects than single-object queries (e.g. "little girl", "elephant"). To this end, SQIR proposes an image retrieval method based on scene graph embedding, which treats visual relationships as directed subgraphs of the scene graph for the purposes of the structured query.

In addition, the image-text retrieval task is formulated as a cross-modal matching task. Given a query in one modality (a sentence query or an image query), the task of image-text cross-modal retrieval is to find the most similar sample from the database in another modality. In 2020, Wang et al. [46] proposed two kinds of scene graphs (visual scene graph (VSG) and textual scene graph (TSG)) to represent image and text respectively, while the scene graph matching model aims to evaluate the similarity of the image-text pairs by dissecting the input image and text sentence into scene graphs. The model collects object features and relationship features, then calculates the similarity score at the object-level and relationship-level, respectively. However, the above methods are often based on fixed text or images for cross-modal retrieval. GEDR (Graph Edit Distance Reward) [163] proposes a more creative and interactive image retrieval method. More specifically, similar to SS-SGN [155], GEDR attempts to edit the scene graph; in more detail, GEDR edits the scene graph according to the user's text instructions on the given image prediction scene graph to perform image retrieval. This makes image retrieval more flexible and promotes easier user interaction.

4.3 Visual Question Answering

VQA is also a multimodal feature learning task. Compared with traditional VQA methods, scene graphs can capture the essential information of images in the form of graph structures, which helps scene graph-based VQA methods to outperform traditional algorithms.

Inspired by the application of traditional QA systems on knowledge graphs, an alternative approach scene graph-based approach is investigated [164]. Zhang et al. explored how to effectively use scene graphs derived from images for visual feature learning, and further applied the graph networks (GN) for encoding the scene graph and performing reasoning according to the questions provided. Moreover, Yang et al. citeyang2018scene aimed improve performance on VQA tasks through the use of scene graphs, and accordingly proposed a new model named Scene GCN (Scene Graph Convolutional Network) [144] to solve the relationship reasoning problem in a visual question-and-answer context. To effectively represent visual relational semantics, a visual relationship encoder is built to yield discriminative and type-aware visual relationship embeddings, constrained by both the visual context and language priors. To confirm the reliability of the results predicted by VQA models, Ghosh et al. [48] proposed an approach named XQA (eXplainable Question Answering), which may be the first VQA model to generate natural language explanations. In XQA, natural language explanations comprised of evidence are generated to answer the questions, which are asked with regard to images using two sources of information: the entity annotations generated from the scene graphs and the attention map generated by a VQA model. As can be determined from these research works, since scene graphs can provide information regarding the relationships between objects in visual scenes, there is significant scope for future research into scene graph-based VQA.

4.4 Image Understanding and Reasoning

Fully understanding an image necessitates the detection and recognition of different visual components, as well as inferring the higher-level events and activities by combining visual modules, reasoning modules, and priors. Therefore, the scene graph with triplets of $\langle subject-relation-object \rangle$ contains information that is very important to image understanding and reasoning. Visual understanding requires the model to have visual reasoning ability. However, existing methods tend to pay less attention to how to make a machine (model) "think", and instead attempt to extract the pixel-level features directly from the images; this is despite the fact that it is difficult to carry out accurate reasoning using pixel-level visual features alone. The task of image reasoning should be based directly on the detected objects, rather than on pixel-level visual features.

More specifically, similar to traditional visual understanding methods, the objects, scenes, and other constituent visual components first need to be detected by a deep learning perception system from input images. A common-sense knowledge base is then built by a Bayesian Network based on the image annotations, while the object interactions are predicted by an intermediate knowledge structure called SDG (Scene Description Graph) [50]. These object interaction priors can be used as the input for image reasoning models and applied to other visual reasoning tasks. In addition, we should focus on teaching a machine (model) to "think" for visual reasoning tasks; for example, by using XNMS (Explicit and Explicit Neural Modules) [145]. XNMS defines several neural modules that are responsible for specific functions (such as object location, attention transformation, etc.) based on scene graphs. XNMS separates "high-level" visual reasoning from "low-level" visual perception and forces the model to focus on how to "think" rather than on simple visual recognition. Since image reasoning is based on object detection and recognition, we hope to learn the mapping from the shared visual feature space by objects and relations to two independent semantic embedding spaces (objects and relations). Moreover, in order to avoid confusion between these two feature spaces, the visual features of the relationships are not transferred to the objects; instead, only the object features are transferred [51]. Visual reasoning based on scene graphs has its applications for reasoning the civic issues [54], which are mainly reflected by the relationships between the objects. Furthermore, generating semantic layout from a scene graph is a crucial intermediate task in the process of connecting textual descriptions to the relevant images.

4.5 3D Scene Understanding

Similar to the 2D scene graph generated from 2D images, a scene graph can also be constructed from 3D scenes as a 3D scene graph, which can provide numerically accurate quantification of the object relationships in 3D scenes. A 3D scene graph succinctly describes the environments by abstracting the objects and their relationships in 3D space in the form of graphs. The construction of 3D scene is very helpful for the understanding of indoor complex environment and other tasks.

In order to construct a 3D scene graph, it is necessary to locate the different objects, identify the elements, attributes, and relationships between the objects in 2D images, and then use all of this information to construct a 3D scene. In [9] and [52], the basic process of used to generate 3D scene graphs is similar, and there are several similar methods (Faster RCNN or Mask RCNN) used to extract the required information from a number of 2D images. However, there are differences in the specific details. For example, different methods have been proposed for constructing 3D scene graphs using the relevant information obtained from 2D images. Specifically, in [9], Armeni et al. tried to construct a 3D scene graph of a building. The constructed 3D Scene Graph consists of four layers: the building, rooms, objects, and cameras. In each layer, there are a set of nodes with their attributes, and edges representing the relationships between nodes. Moreover, in [52], the 3D scene graph is defined by Kim et al. to promote the intelligent agents to gather the semantics within the environments, then apply the 3D scene graph to other downstream tasks. Furthermore, the applicability of the 3D scene graph [52] is verified by demonstrating two major applications of VQA (Visual Question and Answering) and task planning, achieving better performance than the traditional 2D scene graph-based methods. Similarly, 3DSSG (3D Semantic Scene Graphs) [165] and 3D-DSG (3D Dynamic Scene Graphs) [17] also studied the scene understanding of indoor 3D environments. More specifically, 3DSSG proposes a learning method based on PointNet and GCN that moves from the scene point cloud regression to the scene graph. This method has achieved good performance in the 3D scene retrieval task. 3D-DSG attempts to narrow the perception gap between robots and humans in a 3D environment by capturing the metrics and semantics of the dynamic environment. These works have effectively deepened people's understanding of 3D scenes and promoted related applications.

4.6 Human-Object / Human-Human Interaction

There are many fine-grained categories of things in scenes, which can be generally divided into humans and objects. Therefore, some scene graph-related research works have focused on the detection and recognition of HOI (Human-Object Interaction) [22], [146], [147], [148], [149] and HHI (Human-Human Interaction) [53], [166] in scenes. In these works, the long tail of relationships remains a problem to be solved [22], [147], while the detection and recognition of interpersonal relationships have also been proposed [53], [166]; these human-human relationships can be used to further infer the visual social relationships in a scene. In this section, we will discuss the existing methods-based scene graph for the detection and recognition of human-object interaction and human-human interaction.

For HOI, there are two main benchmarks: HICO-DET [167] and HCVRD [22]. This type of visual relational HOI dataset has a natural long-tail distribution, and also have one-shot or zero-shot detection of HOI, which makes it very difficult to conduct model training for most HOI methods in order to achieve better performance. In addition, the task of HOI relies on object detection and involves the construction of human and object pairs with high complexity [146]. The zero-shot learning approach is introduced to address the challenges of scaling HOI recognition to the long tail of categories in the HOI dataset [147]. In addition, HOI recognition is an important means of distinguishing between different types of human actions that happened in the real world. Most HOI methods consist of two steps: human-object pair detection and HOI recognition [146], [168]. The detected proposals of paired human-object regions are passed into a multi-stream network (HO-RCNN [168] and iCAN [146]) to facilitate classification of HOIs by extracting the features from the detected humans, objects, and the spatial relations between them. Moreover, the structural knowledge from the images is also beneficial for HOI recognition, with GCN being a commonly used model for learning the structural features. For example, GPNN (Graph Parsing Neural Network) is proposed in [169] to infer the HOI graph structure represented by an adjacency matrix and node labels. Furthermore, in order to reduce the number of human-object

pairs, some inter activeness priors can be explored for HOI detection; these indicate whether a human and object have interactions with each other [146], and can be learned from the HOI datasets, regardless of HOI category settings.

The above HOI approaches focus primarily on the detection, selection, and recognition of human-object pairs. However, they do not consider whether the approach adopted for corresponding HOI tasks should be human-centric or object centric. In a given scene, however, most human-object interactions are human-centric. Therefore, some HOI works have adopted human-centric approaches such as human-to-object [148], [149] and human-to-human [53], [166]. Inspired by a human-centric approach, we can first identify a human in a scene, then select the human-object pairs of interest to facilitate the recognition of human-object pairs using multi-stream networks; of these, HO-RCNN [148] is a representative example. In addition, the information of HOI can be used for action recognition. InteractNet [149] may be the first proposed multi-task network for human-centric HOI detection and action recognition. This network model can achieve the task of detecting $\langle human{-}verb{-}object \rangle$ triplets in challenging images. Moreover, it was hypothesized that the visual features of the detected persons have powerful cues for localizing the objects with which they intract, so that the model learns to predict the action-specific density over the object locations based on the visual features of the detected persons.

Furthermore, interactions can also take place between humans in a scene, which indicate social relationships. The identification of social relationships in a scene requires a deeper understanding of the scene, along with a focus on human-to-human rather than human-to-object interaction. Therefore, social relationship detection is a task of human-centric HHI, and related works mainly consist of human-human pair detection and social relationship recognition using two network branches [53], [166]. For social relationship recognition, contextual cues can be exploited by a CNN model with attention mechanisms [53]. Adaptive Focal Loss is designed by leveraging the ambiguous annotations so that the models can more effectively learn the relationship features; the goal here is to solve the problem of uncertainty arising during the visual identification of social relationships. The global visual features and mid-level details are also beneficial for social relationship recognition, and GCN is a commonly used model for predicting human social relationships by integrating the global CNN features [166].

5 Datasets for Scene Graphs

Datasets are the basis for driving research related to deep learning, and the research related to scene graph generation is no exception. In this section, we present a detailed summary of the datasets commonly used in scene graph generation tasks, so that interested readers can make their selection accordingly.

Real-World Scene Graphs Dataset (RW-SGD) [8]. RW-SGD is constructed by manually selecting 5,000 images from YFCC100m [170] and Microsoft COCO datasets [171], after which AMT (Amazon's Mechanical Turk) is used to produce a human-generated scene graph from these selected images. The Final RW-SGD contains over 93,832 object instances, 110,021 attribute instances, and 112,707 relationship instances.

Visual Relationship Dataset (VRD) [19] is constructed for the task of visual relationship prediction. VRD has 100 object classes detected from 5,000 images, and also contains 37,993 relationships. However, the distribution of the visual relationships is impacted by the common problem of the long tail of infrequent relationships in scene graph datasets.

Visual Genome Dataset (VGD) [20] is a large-scale visual consisting of various components, including objects, attributes, relationships, question-answer pairs, and so on. In addition, another scene graph dataset (VrR-VG (Visually-Relevant Relationships Dataset)) [29] is constructed based on VGD.

UnRel Dataset (UnRel-D) [21] is a new challenging dataset of unusual relations, and contains more than 1,000 images, which can be queried with 76 triplet queries.

HCVRD Dataset [22] contains 52,855 images with 1,824 object categories and 927 predicates, along with 28,323 relationship types. Similar to VRD, HCVRD also has a long-tail distribution of infrequent relationships.

VrR-VG [29] is constructed based on Visual Genome (VG) by filtering out the visually irrelevant relationships. The top 1,600 objects and 500 relationships are selected from VG by applying a hierarchical clustering algorithm on the relationships' word vectors. Therefore, VrR-VG is a scene graph dataset to highlight visually relevant relationships.

The information of these datasets is summarized in TABLE 1. This includes various attributes of datasets commonly used in scene graph generation tasks.

6 Future Research

Scene graph generation aims at mining the relationships between objects in images or scenes and forming relationship graphs. The generated scene graph transcends the simple understanding of traditional object detection and recognition in visual scenes such as images or videos, and the rich semantic relationships contained in the scene graph greatly improve the performance of related visual tasks. However, there is still significant scope for research to explore and improve the accuracy of the generated relationships, the completeness of relationship graphs, and the efficiency of generating scene graphs ; moreover, the applications of a scene graph in vision or other fields still need to be further explored.

The long-tail distribution in the scene graph. In scene graph datasets, the actually existing long-tailed distribution of relationships directly affects the accuracy and completeness of the generated scene graph, and is thus a problem that many scholars have been trying to solve for the scene graph generation context. For example, zero-shot, one-shot and few-shot learning approaches [121], [147], [172] try to address the challenges of scaling relationship recognition to the long tail of categories in the datasets. Moreover, the language prior information [19], [37], [38] and statistical prior [23], [32], [39] are used to project relationships, such that similar, rare relationships can be predicted to alleviate the problem of the long tail of infrequent relationships.

TABLE 1
Aggregate statistics for scene graph datasets.

Dataset	Images or Videos	Obj. Instances	Obj. Classes	Att. Instances	Att. Types	Rel. Instances	Rel. Types	Predicates. per Obj. Category	Pre.
RW-SGD [8]	5,000	93,832	6,745	110,021	3,743	112,707	1,310	3.3	-
VRD [19]	5,000	-	100	-	-	37,993	6,672	24.25	-
VGD [20]	100k	33,877	3,843,636	-	-	-	40,480	-	-
UnRel [21]	1,000	-	-	-	-	76	-	-	-
HCVRD [22]	52,855	-	1,824	-	-	256,550	28,323	10.63	927
VrR-VG [29]	58,983	282,460	1,600	-	-	203,375	117	-	-

Such types of prior information or analogies between similar relationships are very helpful for detecting infrequent relationships. In addition, [173] tries to solve the long-tail distribution problem in the scene graph by transferring the knowledge learned from the head relationship (relationship with a larger order of magnitude instance) to the tail (relationship with a smaller order of magnitude instance) by means of knowledge transfer. However, there are large numbers of potential, unfrequent, non-focused, or even unseen relationships in the scene that still need to be explored. Associative reasoning through similar objects or similar relationships across scenes may be a good research direction to pursue, as it may solve the long-tail distribution problem of relationships on the current scene graph dataset to a certain extent.

Relationships detection between distant objects. Currently, a scene graph is generated based on large numbers small-scale relationship graphs, which are abstracted from small scenes in scene graph datasets by means of relevant relationship prediction and reasoning models. The selection of potential effective relationships [28], [81] and the establishment of the final relationships in the scene graph are largely dependent on the spatial distance between objects, such that no relationships will exist between two distant objects. However, in the case of a large scene, there are still more such relationships [51]. Therefore, an appropriate proportion of large-scale images can be added to the existing scene graph datasets, while relationships between objects separated by a long distance can be properly considered during scene graph generation, which will improve the integrity of the scene graph.

Scene graph generation based on dynamic images. The scene graph is generated based on static images in scene graph datasets, and the object relationship prediction is also carried out for the static objects in the images by related reasoning models. In practice, however, it may be necessary to predict large numbers of relationships by means of successive actions or events; that is, relationship detection and reasoning based on video scenes. There are very few related research works [50], [145], and little attention has been paid to the role played by the dynamic behaviors of objects in the prediction and inference of the relationships. Moreover, there are difficulties associated with predicting the object relationships in videos, although the predicted relationships may be more accurate. Therefore, we believe that it will be necessary to focus on relationship prediction based on the dynamic actions of the objects in videos.

Social relationship detection based on scene graph. From Section 4.6, we can see that the detection of HOI (human-object interaction) and human-human interaction is an important application of scene graph, and that these types of relationships can be further extended to detect social relationships. We believe that social relationship detection can be used to understand the scenes in more depth, and is thus also a very important research direction. The scene graph generation models based on large-scale datasets can even mine unseen social relationships from the visual data, which has a wider range of practical application values.

Models and methods of visual reasoning. For scene graph generation, the mainstream methods have been developed based on object detection and recognition, visual reference, semantic reasoning, external information introduction, and so on; moreover, RNN, LSTM, and GCN models are the mainstream network models used for visual semantic feature reasoning. The process of relationship recognition also resembles a related mechanism utilized by humans. However, due to the limitations in the current scene graph datasets and the limited capability of relationship prediction models derived using these datasets, it difficult for existing models to continuously enhance their ability to predict relationships. Therefore, we believe that online learning, reinforcement learning, and active learning may be relevant methods or strategies that could be introduced into future scene graph generation methods, as this would enable the scene graph generation models to continuously enhance their relationship prediction abilities by drawing on a large number of constantly updated realistic datasets.

In general, the research in the field of scene graphs has developed rapidly and has broad application prospects. Scene graphs are expected to further promote the understanding and reasoning of higher-level visual scenes. At present, however, scene graph-related research is not sufficiently mature, meaning that it requires more effort and exploration.

7 Conclusion

As a powerful tool for high-level understanding and reasoning analysis of scenes, scene graphs have attracted an increasing amount of attention from researchers. However, research into scene graphs is often cross-modal, complex, and rapidly developing. At the same time, no comprehensive review of scene graph-related research could be found at time of writing. For this reason, we conducted a comprehensive and systematic survey of scene graph generation. In particular, we classified existing SGGs based on the scene graph generation method and the introduction

of additional prior knowledge. We then conducted a comprehensive survey of the application of scene graph generation. In addition, we presented detailed statistics on the datasets used in the context of the scene graph to facilitate the selection of interested readers. Finally, we discussed in detail the future development directions of the scene graph. Therefore, we have reason to believe that this survey will be very helpful for expanding readers' understanding of scene graph development and related research.

REFERENCES

[1] R. A. Rossi, R. Zhou, and N. K. Ahmed, "Deep inductive graph representation learning," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 3, pp. 438–452, 2018.

[2] C. H. Liu, J. Xu, J. Tang, and J. Crowcroft, "Social-aware sequential modeling of user interests: a deep learning approach," IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 11, pp. 2200–2212, 2018.

[3] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, X. Chen, and X. Wang, "A comprehensive survey of neural architecture search: Challenges and solutions," arXiv preprint arXiv:2006.02903, 2020.

[4] ——, "A survey of deep active learning," arXiv preprint arXiv:2009.00236, 2020.

[5] S. Pramanik, R. Haldar, A. Kumar, S. Pathak, and B. Mitra, "Deep learning driven venue recommender for event-based social networks," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 11, pp. 2129–2143, 2019.

[6] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 3, pp. 468–478, 2019.

[7] J. Xu, J. Zhao, R. Zhou, C. Liu, P. Zhao, and L. Zhao, "Predicting destinations by a deep learning based approach," IEEE Transactions on Knowledge & Data Engineering, vol. 33, no. 02, pp. 651–666, 2021.

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

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

[10] H. Qi, Y. Xu, T. Yuan, T. Wu, and S.-C. Zhu, "Scene-centric joint parsing of cross-view videos," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7292–7299.

[11] R. Wang, Z. Wei, P. Li, Q. Zhang, and X. Huang, "Storytelling from an image stream using scene graphs," pp. 9185–9192, 2020.

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

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

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

[15] S. Tripathi, K. Nguyen, T. Guha, B. Du, and T. Q. Nguyen, "Sg2caps: Revisiting scene graphs for image captioning," arXiv preprint arXiv:2102.04990, 2021.

[16] M. Andrews, Y. K. Chia, and S. Witteveen, "Scene graph parsing by attention graph," NIPS, 2018.

[17] A. Rosinol, A. Violette, M. Abate, N. Hughes, Y. Chang, J. Shi, A. Gupta, and L. Carlone, "Kimera: from slam to spatial perception with 3d dynamic scene graphs," arXiv preprint arXiv:2101.06894, 2021.

[18] J. Zhang, K. J. Shih, A. Elgammal, A. Tao, and B. Catanzaro, "Graphical contrastive losses for scene graph generation," arXiv preprint arXiv:1903.02728, 2019.

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

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

[21] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Weakly-supervised learning of visual relations," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5189—5198.

[22] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. van den Hengel, "Hcvrd: a benchmark for large-scale human-centered visual relationship detection," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 7631–7638.

[23] L. D. Dai Bo, Zhang Yuqi, "Detecting visual relationships with deep relational networks," in Proceedings of the IEEE conference on computer vision and Pattern recognition, 2017, pp. 3298–3308.

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

[25] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3107–3115.

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

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

[28] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal, "Relationship proposal networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5226–5234.

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

[30] Y. Li, W. Ouyang, X. Wang, and X. Tang, "Vip-cnn: Visual phrase guided convolutional neural network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7244–7253.

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

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

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

[34] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1270–1279.

[35] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable net: an efficient subgraph-based framework for scene graph generation," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 346–363.

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

[37] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4408–4417.

[38] Z. Cui, C. Xu, W. Zheng, and J. Yang, "Context-dependent diffusion network for visual relationship detection," in Proceedings of

the 26th ACM international conference on Multimedia, 2018, pp. 1475–1482.

[39] R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Visual relationship detection with internal and external linguistic knowledge distillation," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 1068–1076.

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

[41] J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1219–1228.

[42] R. Herzig, A. Bar, H. Xu, G. Chechik, T. Darrell, and A. Globerson, "Learning canonical representations for scene graph to image generation," no. 26, pp. 210–227, 2020.

[43] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 685–10 694.

[44] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," IEEE Transactions on Multimedia, vol. 21, no. 8, pp. 2117–2130, 2019.

[45] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10 322–10 331.

[46] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 1497–1506.

[47] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8102–8109.

[48] S. Ghosh, G. Burachas, A. Ray, and A. Ziskind, "Generating natural language explanations for visual question answering using scene graphs and visual attention," arXiv preprint arXiv:1902.05715, 2019.

[49] V. Damodaran, S. Chakravarthy, A. Kumar, A. Umapathy, T. Mitamura, Y. Nakashima, N. Garcia, and C. Chu, "Understanding the role of scene graphs in visual question answering," arXiv preprint arXiv:2101.05479, 2021.

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

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

[52] U.-H. Kim, J.-M. Park, T.-J. Song, and J.-H. Kim, "3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents," IEEE transactions on cybernetics, pp. 4921–4933, 2020.

[53] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Visual social relationship recognition," arXiv preprint arXiv:1812.05917, 2018.

[54] S. Kumar, S. Atreja, A. Singh, and M. Jain, "Adversarial adaptation of scene graph models for understanding civic issues," in The World Wide Web Conference, 2019, pp. 2943–2949.

[55] L. Wentong, S. Lin, R. Bodo, and M. Y. Yang, "Natural language guided visual relationship detection," in In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 444–453.

[56] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn:towards real-time object detection with region proposal networks," IEEE Transactions on Pattern and Analysis Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[57] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3097–3106.

[58] S. Woo, D. Kim, D. Cho, and I. S. Kweon, "Linknet: Relational embedding for scene graph," in Advances in Neural Information Processing Systems, 2018, pp. 558–568.

[59] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[60] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," arXiv preprint arXiv:1210.5644, 2012.

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

[62] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1529–1537.

[63] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," arXiv preprint arXiv:1603.01360, 2016.

[64] A. McCallum and W. Li, "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," 2003.

[65] A. Quattoni, M. Collins, and T. Darrell, "Conditional random fields for object recognition," Advances in neural information processing systems, vol. 17, pp. 1097–1104, 2004.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[68] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases," in Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1745–1752.

[69] B. Antoine, U. Nicolas, and G.-D. Alberto, "Translating embeddings for modeling multi-relational data," in NIPS, 2013, pp. 2787–2795.

[70] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in Proceedings of the AAAI Conference on Artificial Intelligence, 2015, pp. 2181–2187.

[71] G. Ren, L. Ren, Y. Liao, S. Liu, B. Li, J. Han, and S. Yan, "Scene graph generation with hierarchical context," IEEE Transactions on Neural Networks and Learning Systems, 2020.

[72] J. Jiang, Z. He, S. Zhang, X. Zhao, and J. Tan, "Learning to transfer focus of graph neural network for scene graph parsing," Pattern Recognition, vol. 112, p. 107707, 2021.

[73] B. Zhuang, L. Liu, C. Shen, and 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.

[74] G. Yin, L. Sheng, B. Liu, N. Yu, X. Wang, J. Shao, and C. Change Loy, "Zoom-net: Mining deep feature interactions for visual relationship recognition," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 322–338.

[75] S. Jetley, N. A. Lord, N. Lee, and P. H. Torr, "Learn to pay attention," arXiv preprint arXiv:1804.02391, 2018.

[76] S. Xie and Z. Tu, "Holistically-nested edge detection," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1395–1403.

[77] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rosenberg, and L. Fei-Fei, "Learning semantic relationships for better action retrieval in images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1100–1109.

[78] H. Wan, Y. Luo, B. Peng, and W.-S. Zheng, "Representation learning for scene graph completion via jointly structural and visual embedding." in IJCAI, 2018, pp. 949–956.

[79] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1981–1990.

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

[81] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari, "Detecting visual relationships using box attention," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 1749–1753.

[82] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in CVPR, 2016, pp. 779–788.

[83] L. Wei, A. Dragomir, E. Dumitru, R. SzegedyScott, F. Cheng-Yang, and C. B. Alexander, "Ssd: Single shot multibox detector," in ECCV, 2016, pp. 21–37.

[84] K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang, "Object detection in videos with tubelet proposal networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 889–897.

[85] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang et al., "T-cnn: Tubelets with convolutional neural networks for object detection from videos," IEEE Transactions on Circuits and Systems for Video Technology, vol. 28, no. 10, pp. 2896–2907, 2018.

[86] K. Kang, W. Ouyang, H. Li, and 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.

[87] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[88] C. KyungHyun, v. M. Bart, B. Dzmitry, and B. Yoshua, "On the properties of neural machine translation: Encoder-decoder approaches," CoRR, vol. abs/1409.1259, 2014.

[89] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[90] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 690–706.

[91] K. Masui, A. Ochiai, S. Yoshizawa, and H. Nakayama, "Recurrent visual relationship recognition with triplet unit," in 2017 IEEE International Symposium on Multimedia (ISM). IEEE, 2017, pp. 69–76.

[92] W. Gao, Y. Zhu, W. Zhang, K. Zhang, and H. Gao, "A hierarchical recurrent approach to predict scene graphs from a visual-attention-oriented perspective," Computational Intelligence, vol. 35, no. 3, pp. 496–516, 2019.

[93] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," arXiv preprint arXiv:1406.6247, 2014.

[94] T. V. Nguyen, B. Ni, H. Liu, W. Xia, J. Luo, M. Kankanhalli, and S. Yan, "Image re-attentionizing," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 1910–1919, 2013.

[95] J. Li, M.-T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," arXiv preprint arXiv:1506.01057, 2015.

[96] R. Lin, S. Liu, M. Yang, M. Li, M. Zhou, and S. Li, "Hierarchical recurrent neural network for document modeling," in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 899–907.

[97] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang, "Scene dynamics: Counterfactual critic multi-agent training for scene graph generation," arXiv preprint arXiv:1812.02347, vol. 3, 2018.

[98] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7239–7248.

[99] Y. Xia, L. Zhang, Z. Liu, L. Nie, and X. Li, "Weakly supervised multimodal kernel for categorizing aerial photographs," IEEE Transactions on Image Processing, vol. 26, no. 8, pp. 3748–3758, 2016.

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

[101] W. Wang, R. Wang, S. Shan, and X. Chen, "Sketching image gist: Human-mimetic hierarchical scene graph generation," arXiv preprint arXiv:2007.08760, 2020.

[102] C. Goller and A. Kuchler, "Learning task-dependent distributed representations by backpropagation through structure," in Proceedings of International Conference on Neural Networks (ICNN'96), vol. 1. IEEE, 1996, pp. 347–352.

[103] M. Gori, G. Monfardini, and F. Scarselli, "A new model for learning in graph domains," in Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., vol. 2. IEEE, 2005, pp. 729–734.

[104] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," IEEE transactions on neural networks, vol. 20, no. 1, pp. 61–80, 2008.

[105] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in International Conference on Machine Learning. PMLR, 2017, pp. 1263–1272.

[106] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," Computer Science, 2015.

[107] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," arXiv preprint arXiv:1706.01427, 2017.

[108] J. Bruna and S. Mallat, "Invariant scattering convolution networks," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1872–1886, 2013.

[109] H. Dai, B. Dai, and L. Song, "Discriminative embeddings of latent variable models for structured data," in International conference on machine learning. PMLR, 2016, pp. 2702–2711.

[110] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," arXiv preprint arXiv:1606.09375, 2016.

[111] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in International conference on machine learning. PMLR, 2016, pp. 2014–2023.

[112] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," arXiv preprint arXiv:1706.02216, 2017.

[113] R. Herzig, E. Levi, H. Xu, E. Brosh, A. Globerson, and T. Darrell, "Classifying collisions with spatio-temporal action graph networks," arXiv preprint arXiv:1812.01233, vol. 2, 2018.

[114] R. Herzig, M. Raboh, G. Chechik, J. Berant, and A. Globerson, "Mapping images to scene graphs with permutation-invariant structured prediction," in Advances in Neural Information Processing Systems, 2018, pp. 7211–7221.

[115] X. Wang and A. Gupta, "Videos as space-time region graphs," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 399–417.

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

[117] S. Jae Hwang, S. N. Ravi, Z. Tao, H. J. Kim, M. D. Collins, and V. Singh, "Tensorize, factorize and regularize: Robust visual relationship learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1014–1023.

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

[119] X. Yang, H. Zhang, and 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.

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

[121] A. Dornadula, A. Narcomey, R. Krishna, M. Bernstein, and F.-F. Li, "Visual relationships as functions: Enabling few-shot scene graph prediction," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 1730–1739.

[122] X. Shang, T. Ren, J. Guo, H. Zhang, and T.-S. Chua, "Video visual relation detection," in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1300–1308.

[123] L. Chen, H. Zhang, J. Xiao, X. He, S. Pu, and S.-F. Chang, "Counterfactual critic multi-agent training for scene graph generation," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4612–4622.

[124] M. Raboh, R. Herzig, J. Berant, G. Chechik, and A. Globerson, "Differentiable scene graphs," in In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1477–1486.

[125] M. Khademi and O. Schulte, "Deep generative probabilistic graph neural networks for scene graph generation," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 11 237–11 245.

[126] R. Wang, Z. Wei, P. Li, Q. Zhang, and X. Huang, "Storytelling from an image stream using scene graphs," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, 2020, pp. 9185–9192.

[127] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[128] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, "Learning to generalize to new compositions in image understanding," arXiv preprint arXiv:1608.07639, 2016.

[129] B. A. Plummer, A. Mallya, C. M. Cervantes, J. Hockenmaier, and S. Lazebnik, "Phrase localization and visual relationship detection with comprehensive image-language cues," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1946–1955.

[130] Z. Hu, X. Ma, Z. Liu, E. Hovy, and E. Xing, "Harnessing deep neural networks with logic rules," pp. 2410–2420, 2016.

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

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

[133] A. Zareian, S. Karaman, and S.-F. Chang, "Bridging knowledge graphs to generate scene graphs," no. 23, pp. 606–623, 2020.

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

[135] R. Speer and C. Havasi, "Conceptnet 5: A large semantic network for relational knowledge," in The People's Web Meets NLP. Springer, 2013, pp. 161–176.

[136] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1576–1585.

[137] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6857–6866.

[138] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, no. 1, 2017.

[139] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, and J. Li, "Learning visual knowledge memory networks for visual question answering," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7736–7745.

[140] A. Zareian, H. You, Z. Wang, and S.-F. Chang, "Learning visual commonsense for robust scene graph generation," arXiv preprint arXiv:2006.09623, 2020.

[141] G. Mittal, S. Agrawal, A. Agarwal, S. Mehta, and T. Marwah, "Interactive image generation using scene graphs," CoRR, vol. abs/1905.03743, 2019.

[142] T. Subarna, B. Anahita, B. Alexei, and T. Hanlin, "Using scene graph context to improve image generation," CoRR, vol. abs/1901.03762, 2019.

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

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

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

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

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

[148] C. Yu-Wei, L. Yunfan, L. Xieyang, Z. Huayi, and D. Jia, "Learning to detect human-object interactions," WACV2018, pp. 381–389, 2018.

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

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

[151] S. Tripathi, S. Nittur Sridhar, S. Sundaresan, and H. Tang, "Compact scene graphs for layout composition and patch retrieval," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 676–683.

[152] B. Schroeder, S. Tripathi, and H. Tang, "Triplet-aware scene graph embeddings," in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 1783–1787.

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

[154] F. Fei, Y. Miao, F. Hui, H. Shenghong, and X. Chunxia, "Narrative collage of image collections by scene graph recombination," IEEE Trans. Vis. Comput. Graph., vol. 24, no. 9, pp. 2559–2572, 2018.

[155] H. Dhamo, A. Farshad, I. Laina, N. Navab, G. D. Hager, F. Tombari, and C. Rupprecht, "Semantic image manipulation using scene graphs," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5213–5222.

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

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

[158] C. Shizhe, J. Qin, W. Peng, and W. Qi, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," CVPR, pp. 9959–9968, 2020.

[159] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, "Comprehensive image captioning via scene graph decomposition," in European Conference on Computer Vision. Springer, 2020, pp. 211–229.

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

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

[162] B. Schroeder and S. Tripathi, "Structured query-based image retrieval using scene graphs," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 178–179.

[163] L. Chen, G. Lin, S. Wang, and Q. Wu, "Graph edit distance reward: Learning to edit scene graph," in European Conference on Computer Vision. Springer, 2020, pp. 539–554.

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

[165] J. Wald, H. Dhamo, N. Navab, and 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.

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

[167] C. YuWei, W. Zhan, H. Yugeng, W. Jiaxuan, and D. Jia, "Hico: A benchmark for recognizing human-object interactions in images," ICCV2015, pp. 1017–1025, 2015.

[168] B. Xu, Y. Wong, J. Li, Q. Zhao, and M. S. Kankanhalli, "Learning to detect human-object interactions with knowledge," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2019–2028.

[169] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 407–423.

[170] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "The new data and new challenges in multimedia research," arXiv preprint arXiv:1503.01817, vol. 1, no. 8, 2015.

[171] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.
[172] X. Lin, C. Ding, J. Zeng, and 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.
[173] T. He, L. Gao, J. Song, J. Cai, and Y.-F. Li, "Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation," arXiv preprint arXiv:2006.07585, 2020.