A Survey of Neurosymbolic Visual Reasoning with Scene Graphs and Common Sense Knowledge

M. Jaleed Khan [a,*], Filip Ilievski [b], John G. Breslin [a,c] and Edward Curry [a,c]

[a] *SFI Centre for Research Training in Artificial Intelligence, Data Science Institute, University of Galway, Ireland*
[b] *Center on Knowledge Graphs, Information Sciences Institute, University of Southern California, United States*
[c] *Insight SFI Research Centre for Data Analytics, Data Science Institute, University of Galway, Ireland*
E-mails: m.khan12@universityofgalway.ie, ilievski@isi.edu, john.breslin@universityofgalway.ie, edward.curry@universityofgalway.ie

Abstract. Combining deep learning and common sense knowledge via neurosymbolic integration is essential for semantically rich scene representation and intuitive visual reasoning. This survey paper delves into data- and knowledge-driven scene representation and visual reasoning approaches based on deep learning, common sense knowledge and neurosymbolic integration. It explores how scene graph generation, a process that detects and analyses objects, visual relationships and attributes in scenes, serves as a symbolic scene representation. This representation forms the basis for higher-level visual reasoning tasks such as visual question answering, image captioning, image retrieval, image generation, and multimodal event processing. Infusing common sense knowledge, particularly through the use of heterogeneous knowledge graphs, improves the accuracy, expressiveness and generalizability of the representation and allows for intuitive downstream reasoning. Neurosymbolic integration in these approaches ranges from loose to tight coupling of neural and symbolic components. The paper reviews and categorizes the state-of-the-art knowledge-based neurosymbolic approaches for scene representation based on the types of deep learning architecture, common sense knowledge source and neurosymbolic integration used. The paper also discusses the visual reasoning tasks, datasets, evaluation metrics, key challenges and future directions, providing a comprehensive review of this research area and motivating further research into knowledge-enhanced and data-driven neurosymbolic scene representation and visual reasoning.

Keywords: scene graph, image representation, deep learning, common sense knowledge, neurosymbolic integration, visual reasoning

1. Introduction

The field of Artificial Intelligence (AI) has seen significant advancements, particularly in scene representation and visual reasoning, with the integration of deep learning, common sense knowledge, and NeuroSymbolic (NeSy) approaches [1–4]. NeSy integration combines the strengths of neural and symbolic approaches, enhancing the performance of black-box neural networks and enabling large-scale symbolic reasoning. Scene Graph Generation (SGG), a process that constructs symbolic image representations, has become a widely used technique for higher-level visual reasoning tasks [5, 6]. Despite substantial progress in deep learning and multi-modal methods in computer

[*] Corresponding author. E-mail: m.khan12@universityofgalway.ie.

vision, data-centric techniques often fall short in complex visual reasoning problems that require semantic and relational information [7]. NeSy hybrid methodologies have emerged to address this, finding applications in areas such as visual narration [8], self-driving vehicles [9], mathematical logic [10], robotic manipulation [11], and medical diagnostics [12].

Despite the advancements in SGG, its practical applicability remains constrained by several challenges that directly impact its accuracy, expressiveness, and robustness. The quality of annotations and the skewed distribution of relationship predicates in crowd-sourced datasets have been identified as significant challenges for data-driven SGG methods. For instance, generic relationship predicates like "on", "has", and "in" dominate the Visual Genome dataset [13]. These generic predicates often fail to capture the nuanced visual relationships in scenes, thereby affecting the accuracy of visual relationship prediction in SGG. The expressiveness of SGG, reflecting its ability to depict scenes in a comprehensive and intuitive manner, is also compromised. For example, the relationship *(man, riding, bike)* is more accurate and expressive than *(man, on, bike)*. The task is further complicated by the vast variability in the visual appearances of relationships across different scenes. Consider the relationships *(man, holding, food)* and *(man, holding, bat)*; while they share the same predicate, their visual representations differ significantly. Furthermore, the robustness of SGG, which refers to its consistent performance across both familiar and unfamiliar scenes and regardless of the frequency of visual relationships in datasets, is also an important concern. Numerous efforts have been made to overcome these obstacles, exploring novel facets of visual relationships in images, such as heterophily [14] and saliency [15], and employing advanced techniques like knowledge transfer [16], linguistic supervision [17] and zero-shot learning [18].

Common sense knowledge infusion has evolved as a promising strategy to tackle these challenges [6]. Incorporating background details and related facts about scene components enhances the expressiveness of the representation and the performance of downstream reasoning [6]. While statistical and language priors have been widely used in SGG, they offer limited generalizability. Some KGs, such as ConceptNet [19], and WordNet [20], have been utilized in SGG. These KGs provide text-based and lexical knowledge representing different forms and notions of common sense. However, they do not provide broad common sense knowledge about visual concepts. Heterogeneous KGs, such as the Common Sense Knowledge Graph (CSKG) [21], provide a wider range of common sense knowledge about visual concepts and are crucial yet underutilized sources for the infusion of external common sense knowledge in scene representation and visual reasoning techniques.

This paper presents a comprehensive review of the combination of deep learning, common sense knowledge, and NeSy integration for semantic scene representation and visual reasoning. Such a comprehensive survey on this topic is inspired by the growing interest in combining deep learning, common sense knowledge, and NeSy integration in computer vision. The promise of this research direction requires a survey that clearly presents the current state of the literature on this subject and points out the current challenges, prospects, and applications to guide future research, ensuring that efforts are channeled effectively to address the challenges and elevate the performance of SGG to a practical level. Our survey reviews state-of-the-art techniques, datasets, and evaluation metrics, classifies existing SGG methods, and discusses key challenges and promising future research directions. This survey aims to serve as a valuable resource for researchers and practitioners in the field, guiding future research directions and contributing to the development of more effective and practical solutions for real-world applications.

1.1. Existing Surveys

Garcez et al. [22] and Wang et al. [23] provided comprehensive reviews of neuro-symbolic AI, discussing its development, forms of integration, the importance of representation, and promising future research directions. Ilievski et al. [24] analyzed multiple sources of common sense knowledge, categorizing them into 13 dimensions and suggesting a roadmap for developing a unified resource for neuro-symbolic methods. Kursuncu et al. [25] discussed the potential of hybrid neuro-symbolic learning approaches that integrate deep learning and knowledge graphs. Meanwhile, Chang et al. [5] provided a comprehensive review of SGG methods, applications, and datasets, while Zhu et al. [26] systematically summarized deep learning-based SGG methods and compared their performance across different datasets and representations [26]. These surveys are summarized in Table 1 and broadly classified into three domains, i.e. NeSy integration, common sense knowledge infusion and SGG, based on the main focus of each survey paper. The intersection of these domains is emerging as a promising research direction, showing significant potential for intuitive visual reasoning. There is a substantial need for a specialized survey paper on deep learning and common sense knowledge combined via NeSy integration for SGG and visual reasoning, which our paper addresses.

Table 1

Comparison with existing surveys

Domain	**Survey (Year)**	**Key Attributes**
Neurosymbolic (NeSy) Integration	Garcez et al. [22] (2023)	neurosymbolic AI, machine learning, reasoning, explainable AI, deep learning, trustworthy AI, cognitive reasoning
	Wang et al. [23] (2022)	neurosymbolic AI, symbolic AI, statistical AI, deep learning
Commonsense Knowledge Infusion	Ilievski et al. [24] (2021)	common sense knowledge, semantics, knowledge graphs, reasoning
	Kursuncu et al. [25] (2019)	knowledge-infused learning, knowledge graph, neural network, neurosymbolic AI
Scene Graph Generation (SGG)	Zhu et al. [26] (2022)	scene graph generation, visual relationship detection, object detection, scene understanding
	Chang et al. [5] (2021)	scene graph, visual feature extraction, prior information, visual relationship recognition
Intersection of the three domains	Ours (2023)	scene graph, visual reasoning, scene understanding, deep learning, common sense knowledge, neurosymbolic integration, VQA, image captioning

1.2. Contributions and Organization

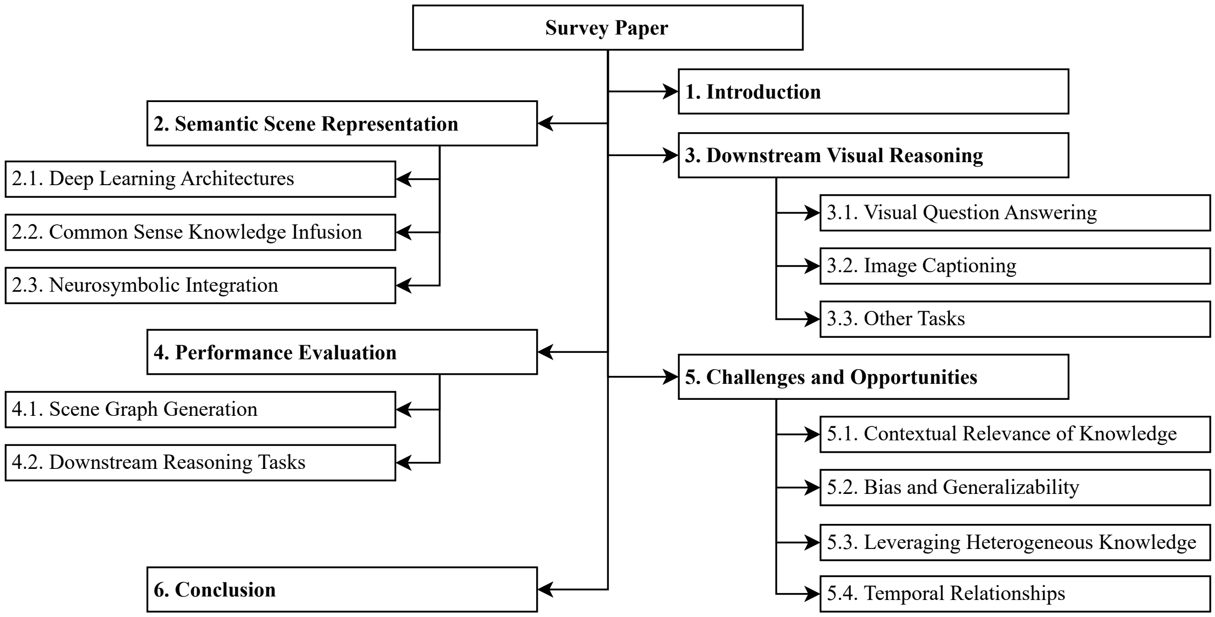
Survey Paper
1. Introduction
2. Semantic Scene Representation
2.1. Deep Learning Architectures
2.2. Common Sense Knowledge Infusion
2.3. Neurosymbolic Integration
3. Downstream Visual Reasoning
3.1. Visual Question Answering
3.2. Image Captioning
3.3. Other Tasks
4. Performance Evaluation
4.1. Scene Graph Generation
4.2. Downstream Reasoning Tasks
5. Challenges and Opportunities
5.1. Contextual Relevance of Knowledge
5.2. Bias and Generalizability
5.3. Leveraging Heterogeneous Knowledge
5.4. Temporal Relationships
6. Conclusion

Fig. 1. Structure of the survey paper.

This survey aims to review and summarize the literature on using deep learning, common sense knowledge and NeSy integration for scene representation and visual reasoning. The key contributions of this survey are as follows:

- To the best of our knowledge, this is the first paper to provide a comprehensive survey of the combination of deep learning, common sense knowledge and NeSy integration for semantic scene representation and visual reasoning.

- We provide a comprehensive review of the state-of-the-art techniques, datasets and evaluation metrics for knowledge-based scene representation and visual reasoning approaches. We also classify the existing scene graph generation methods based on deep learning architecture, common sense knowledge source and NeSy integration type used in each method.
- We discuss the key challenges of the existing knowledge-based scene representation and visual reasoning methods and present contextual relevance of knowledge, bias and generalizability, use of heterogeneous common sense knowledge and temporal visual relationships as promising future research directions.

The rest of this paper is organized as follows. Section 2 reviews the state-of-the-art in knowledge-based semantic scene representation in detail and classifies the existing approaches based on deep learning architecture, common sense knowledge source and NeSy integration type. Section 3 discusses the downstream tasks that leverage the structured scene representation for intuitive visual reasoning. Section 4 discusses the benchmark datasets and performance measures used for the evaluation of SGG and downstream reasoning methods. Section 5 provides a summary of the main challenges and promising future directions in this area of work. Finally, Section 6 summarizes and concludes the paper. The structure of this survey is presented in Figure 1.

2. Semantic Scene Representation

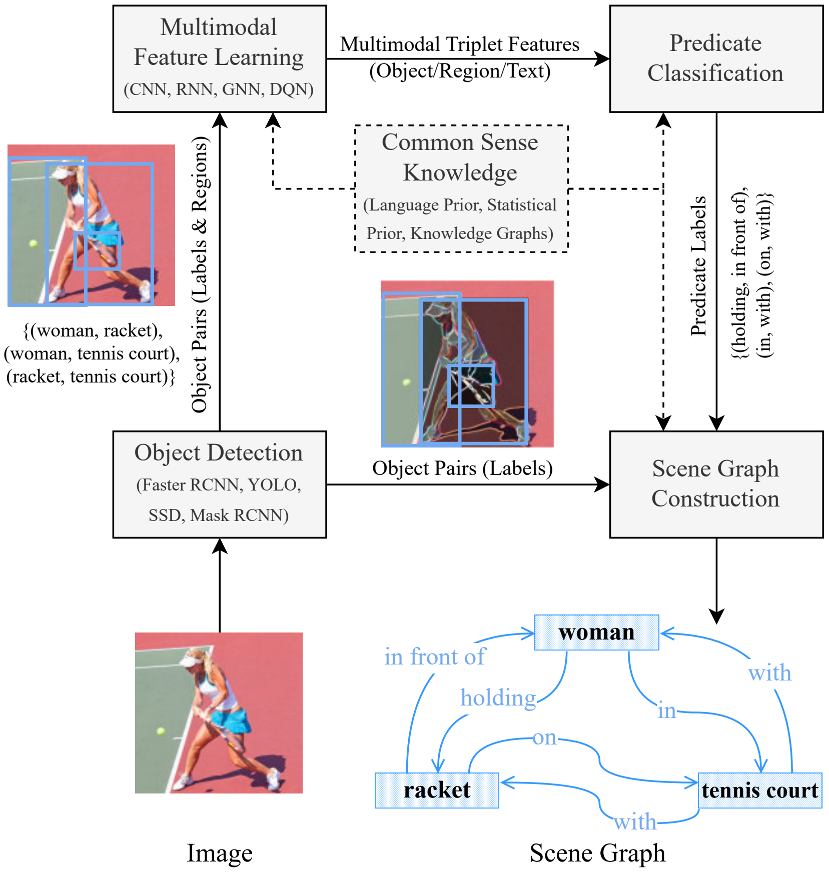
Multimodal Feature Learning
(CNN, RNN, GNN, DQN)
Multimodal Triplet Features (Object/Region/Text)
Predicate Classification
Common Sense Knowledge
(Language Prior, Statistical Prior, Knowledge Graphs)
Object Pairs (Labels & Regions)
{(woman, racket), (woman, tennis court), (racket, tennis court)}
Predicate Labels
{(holding, in front of), (in, with), (on, with)}
Object Detection
(Faster RCNN, YOLO, SSD, Mask RCNN)
Object Pairs (Labels)
Scene Graph Construction
woman
in front of
holding
with
in
on
racket
tennis court
with
Image
Scene Graph

Fig. 2. A schematic representation of the typical scene graph generation process comprising object detection, multimodal feature learning, common sense knowledge infusion, relationship predicate classification, and scene graph construction

High-level visual reasoning necessitates semantic and relational information, especially concerning object interactions within scene representations. Recently, there has been a surge in the adoption of knowledge-based and NeSy approaches for semantic scene representation. The effectiveness of downstream visual reasoning tasks is largely dependent on the expressiveness and quality of the semantic scene representation. Several efforts have been undertaken to capture visual features and object interactions in a systematic and explicit manner. The scene graph, which structures objects and pairwise relationships in a semantically-grounded manner, has emerged as a commonly used semantic scene representation [5]. SGG comprises the detection and contextual analysis of objects, visual relationships and attributes, leading to the construction of a symbolic representation of the scene, as demonstrated in Figure 2. These symbolic scene graphs lay the groundwork for advanced visual reasoning with examples of Visual Question Answering (VQA), image captioning, Multimedia Event Processing (MEP), image retrieval and image generation.

Deep learning is integral to the task of SGG. Deep learning architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs), can extract and understand complex visual features, handle large volumes of unstructured data, and capture intricate relationships between objects. These capabilities make deep learning essential for processing and interpreting the complex visual data involved in SGG. SGG is a complex task due to the extensive semantic space of potential pairwise relationships between objects in a scene. Capturing all these relationships in a finite training dataset is nearly impossible. Therefore, the integration of common sense knowledge, including statistical priors [27–29], language priors [30, 31], and KGs [32–36], becomes crucial. Common sense knowledge infusion helps bridge the gap between the limited training data and the vast semantic space, enabling a more accurate and comprehensive representation of relationships within a scene. NeSy integration in SGG techniques can be loosely or tightly coupled. In loose coupling, [28, 32, 33] the neural and symbolic components operate independently, interacting as needed, and focus on distinct yet complementary tasks. Meanwhile, tight coupling [27, 29–31, 34–36] deeply integrates symbolic and neural components, either incorporating symbolic knowledge directly into the neural network architecture or encoding it into the network's distributed representation. The following subsections present a detailed overview of various deep learning architectures, common sense knowledge sources and NeSy integration types used in knowledge-based SGG. Table 2 provides an overview of their characteristics, their main types and associated methods.

2.1. Deep Learning Architectures

SGG involves detecting and classifying objects in an image and understanding the relationships between them. This process requires the interpretation of complex visual data, a task that deep learning is uniquely equipped to handle. Deep learning architectures are capable of learning hierarchical representations from raw data. These architectures can automatically learn to extract and combine features, layer by layer, from raw pixels in images to form high-level features, such as edges, textures, and shapes. This ability to learn and understand features at various levels of abstraction allows these models to recognize and classify objects and visual relationships in an image, which is the core task in SGG. Moreover, deep learning architectures can handle large amounts of unstructured data, such as images, videos and text, and are capable of learning from this data in an end-to-end manner. Deep learning architectures can capture complex non-linear relationships and dependencies between variables, which is crucial for understanding the visual relationships between various objects detected in an image. For instance, the relationship between a person and a bicycle in an image might depend on various factors, such as the position and orientation of the person and the bicycle, and deep learning architectures can learn to capture these complex dependencies.

2.1.1. Convolutional Neural Networks

CNN is a predominant deep learning architecture in SGG due to its exceptional capability in extracting visual features from images. It is employed to extract global and local visual features of an image, subsequently facilitating the prediction of relationships between subjects and objects through classification. Most of the knowledge-based SGG techniques [27–36] use Faster RCNN with a CNN-based backbone network for detecting objects in images prior to visual relationship detection. Khan et al. [32] used the feature maps extracted from the underlying CNN in Faster RCNN by applying RoIAlign to the image regions to obtain local and global region features of each detected object, which forms the basis for further processing and relationship prediction. DSGAT [27] incorporated Faster RCNN with a VGG16 backbone in its bounding box module to generate object proposals prior to visual relationship

Table 2
Summary of Main Characteristics of Knowledge-based NeSy SGG Methods

Characteristic	Type	Description	Methods
Deep Learning Architecture	CNN	– Used for local and global visual feature extraction and object detection – Learns hierarchical representations from large volumes of raw data.	[27–36]
	RNN	– Captures contextual information and learns dependencies between objects – Handles sequential data and maintains information over longer sequences	[28, 32, 34, 36]
	GNN	– Processes graph-structured data and facilitates message passing in SGG – Learns local information and captures the relationships between objects	[27, 29, 33, 35]
	DQN	– Formulates SGG as a sequential decision-making process – Handles high-dimensional continuous data and unseen inputs	[31]
Common Sense Knowledge Source	Statistical Prior	– Captures structural regularities in visual scenes – Models statistical correlations between object pairs	[27–29]
	Language Prior	– Refines relationship predictions using semantic information – Helps recognize relationships between semantically related objects	[30, 31]
	Knowledge Graph	– Provides a structured representation of common sense knowledge – Facilitates inference of unseen visual relationships	[32–36]
Neurosymbolic Integration	Loose Coupling	– Independent and sequential operation of neural and symbolic components – Flexibility in handling distinct yet complementary tasks	[28, 32, 33]
	Tight Coupling	– Symbolic knowledge is directly encoded into the neural networks – Unified symbolic reasoning and neural learning capabilities	[27, 29–31, 34–36]

detection. IRT-MSK [33] also employed Faster RCNN, however, the authors used a transformer to extract and fully explore the context of visual features rather than extracting visual features of each entity individually, enhancing the understanding of the visual scene. COACHER [34] used a pre-trained Faster RCNN for generating a set of region proposals, label probabilities distributions, and visual embeddings for each detected object as a part of the zero-shot SGG framework. GB-Net [35] represented objects detected using Faster-RCNN as scene entity nodes in the subsequent stages processing the detected objects, each with a label distribution, bounding box and RoI-aligned feature vector. KB-GAN [36] employs a Region Proposal Network (RPN), a type of CNN, for extraction of object proposals in images. The RPN module generates bounding boxes for potential objects in the image, which are then used to construct subgraph proposals. VRD Model [30] comprises a CNN to classify objects and predicates within an image by processing the image region representing the union of the bounding boxes of the interrelated objects.

2.1.2. Recurrent Neural Networks

The interaction of information among various objects within a scene, along with their contextual information, is vital for identifying pairwise visual relationships between these objects. Knowledge-based SGG models built on RNN and its variants, i.e. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, inherently excel at capturing this contextual information within the scene graph and reasoning based on the structured data within the graph. Khan et al. [32] used two sets of Bi-directional LSTM (BiLSTM) layers: one to encode the region features, image regions, and class labels as individual visual context features, and one to encode these individual visual context features of objects and concatenate them into combined pairwise object features for relationship classification in SGG. The COACHER model [34] utilizes a bi-directional LSTM to generate background embeddings that encapsulate information from both the region proposal and the global image level. A separate LSTM is then employed to decode each region proposal embedding, yielding a one-hot vector that signifies the refined class label of a region proposal. Once refined object labels for all region proposals are obtained, they are processed to generate context embeddings through a BiLSTM. These context embeddings are subsequently used to derive edge embeddings and predict the relationship between each pair of bounding boxes. MotifNet [28] leverages LSTMs to encode a global context that guides local predictors. The model sequences the prediction of bounding boxes, object classification, and relationship prediction in such a way that the global context encoding of earlier stages provides a rich context for prediction in the following stages. The global context across image regions is calculated and disseminated via BiLSTMs. This context is utilized by another LSTM layer that assigns labels to each region based on the overall context and the preceding labels. Subsequently, a dedicated BiLSTM layer computes and propagates the information for predicting edges, based on the regions, their labels, and the context. This approach allows the model to capture crucial dependencies between object labels and relation labels during the SGG process. KB-GAN [36] employed GRUs in the knowledge retrieval and embedding stage; the retrieved common sense relationships, transformed into a sequence of words, are fed into a bidirectional GRU for the effective encoding of the sequence, capturing both past and future context.

2.1.3. Graph Neural Networks

The graph structure of scene graphs makes graph-based architectures, such as GNN, Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT), a suitable choice to enhance SGG performance. GCN is used to effectively learn local information between neighbouring nodes in knowledge-based SGG. Its inherent graph-based structure plays a crucial role in guiding message passing in GNN- and GCN-based SGG methods. DSGAT [27] used a GAT component in its graphical message passing module for effective contextual learning and recognizing the object classes and visual relationships. The GAT component allows for effective message propagation and relationship prediction by facilitating interaction between object features and relational features via the inherent weight and attention weight of the multi-head GAT. IRT-MSK [33] leveraged GCNs to process the graph-structured knowledge, which includes both relational and common sense knowledge, and learn the semantic features of the entities embedded in the KG during the SGG process. GB-Net [35] employed Gated GNNs that iteratively propagate information within and between two graphs, i.e. the scene graph and a background common sense graph, successively inferring edges and nodes, leveraging the strengths of GNNs in handling graph-structured data and learning local information between neighbouring nodes. KERN [29] used a GNN to propagate messages through a graph built based on statistical object co-occurrence information, learning contextualized representations for each region and achieving better label prediction. A second GNN is used to explore the interplay between relationships and objects, with nodes representing objects and relationships, and edges representing the statistical co-occurrences between the corresponding object pair and all the relationships.

2.1.4. Deep Q-Networks

Deep Q-Networks, also contribute to the rich variety of deep learning techniques applied in knowledge-based SGG, further expanding the possibilities for scene understanding. DeepVRL[31] approaches the task of identifying visual relationships and attributes as a sequential process, managed using a Deep Q-Network (DQN). The DQN is employed to calculate three sets of Q-values, each corresponding to the action sets of attributes, predicates, and object categories. Using an ϵ-greedy strategy, the DQN sequentially identifies the optimal actions to uncover objects, relationships, and attributes in the provided image. The framework also incorporates a replay memory to retain information from previous episodes, which aids in stabilizing the training by averaging the training distribution over past experiences and minimizing the correlation among training examples.

2.2. Common Sense Knowledge Infusion

SGG is an inherently complex task due to the vast semantic space of possible relationships. The semantic space, in this context, refers to all possible relationships that can exist between different objects within a scene. This space is vast and complex, encompassing everything from simple relationships such as “cat-sits-on-mat" to more complex ones like “bird-perches-on-branch-of-tree". Given the infinite variety and complexity of these relationships, it is nearly impossible to capture all of them within a finite training dataset. This is where the infusion of common sense knowledge, a concept rooted in the understanding of the world as humans perceive it [24], becomes particularly crucial. Common sense knowledge refers to the basic, generally accepted information and reasoning that humans use to navigate the world around them. In the context of SGG, this includes understanding that birds are more likely to be found in trees than fish, or that people are more likely to sit on chairs than on clouds. By integrating this common sense knowledge into the SGG pipelines, we can bridge the gap between the limited scope of the training data and the vastness of the semantic space. This allows for a more accurate and comprehensive representation of the relationships within a scene, even when these relationships are not explicitly present or adequately represented in the training data. Common sense knowledge sources used in SGG can be broadly classified into three categories: statistical prior, language prior, and KG [6].

2.2.1. Statistical Priors

Statistical priors are a form of common sense knowledge that leverages the observed structural regularities and statistical correlations in visual scenes. For instance, certain relationships such as bird-flies-in-sky or dog-chases-cat are more frequently observed than others like bird-swims-in-water or cat-chases-dog. By modelling these statistical correlations, SGG can more accurately identify and predict visual relationships. It is similar to understanding the world through patterns and trends that are statistically significant, providing a probabilistic framework to predict relationships that are likely to occur based on past observations. For example, DSGAT [27] integrates statistical prior probabilities into the sparse graph component and graphical message propagation network to construct a sparse KG and learn statistical co-occurrence modelling for identifying and predicting visual relationships. MotifNet [28] also used statistical priors as a form of common sense knowledge, capturing dependencies between objects and relationships by leveraging structural regularities and statistical correlations observed in visual scenes. The model breaks down the likelihood of a graph into three distinct elements: bounding boxes, objects, and relations, making no independent assumptions during SGG. KERN [29] leverages statistical correlations between pairwise objects and visual relationships to regularize the semantic space and minimize the unbalanced distribution problem by explicitly representing these statistical correlations in a structured KG. The technique uses a routing mechanism to pass messages within the graph, exploring relationships between objects, thus integrating statistical prior knowledge into the deep learning process in SGG.

2.2.2. Language Priors

Language priors utilize the semantic information encapsulated in words to enhance the prediction of relationships. They aid in recognizing visual relationships by observing objects that are semantically correlated. For instance, even if the co-occurrence of “child" and “kite" is infrequent in the training data, the language prior from a more common example like “a child holding a toy" can help infer that a plausible relationship between a child and a kite could be “holding". This is because language priors understand the semantic context and use it to predict relationships that may not explicitly appear in the training data but are likely in the real world. VRD model [30] detects visual relationships within an image by leveraging visual appearance and language modules. The language module employs pre-trained word vectors (word2vec) to project the relationships onto an embedding space where semantically similar relationships are close together. This allows the model to infer less frequent relationships from similar, more common ones, effectively utilizing language priors as a source of common sense knowledge. DeepVRL [31] approaches the task of identifying visual relationships and attributes as a sequential process, utilizing language priors to progressively uncover object relationships and attributes within an image. It builds a directed semantic action graph that encapsulates semantic associations between object classes, attributes and predicates, using language priors. The system then employs a variation-structured traversal across the action graph, creating an adaptive action set at each stage, contingent on the current state and past actions. To address semantic ambiguity among object categories, an ambiguity-aware object mining strategy is implemented.

2.2.3. *Knowledge Graphs*

KGs serve as extensive knowledge bases that encode the structure of the world and its relationships. They have been employed as a form of prior common sense knowledge to assist in the generation of scene graphs. KGs used for the infusion of prior common sense knowledge in SGG are presented in Table 3. By offering a rich source of common sense knowledge about how objects and entities relate to each other, KGs significantly enhance the accuracy and completeness of SGG. For instance, a KG can provide information that "a bird is likely to be found in a tree", thereby allowing the SGG to infer this relationship even if it's not explicitly present in the training data. Khan et al. [32] employed CSKG, a heterogeneous KG, consolidated from seven different knowledge bases, to generate expressive and semantically-rich scene graphs. The graph embeddings of object nodes were used to compute similarity metrics for scene graph refinement and knowledge enrichment. In IRT-MSK [33], the authors leveraged multiple structured knowledge sources, specifically relational knowledge and common sense knowledge, to encapsulate relationships between entities derived from images and to encode intuitive knowledge, such as “dog can guard yard", respectively. Infusing prior common sense knowledge from Visual Genome and ConceptNet KGs into the SGG process enhanced the accuracy and context awareness of the generated scene graphs. COACHER [34] employed graph mining pipelines to model neighbourhoods and paths around entities in ConceptNet and integrates them into the SGG framework. COACHER uses ConceptNet to generate common sense knowledge embeddings, which are then used to enhance zero-shot relation prediction. It develops three types of integrators: neighbour, path, and fused. The neighbour integrator generates common sense knowledge embeddings based on the neighbourhood information of a node in ConceptNet, while the path integrator retrieves a set of paths connecting two entities and learns a representation for each set of paths. The fused integrator combines the neighbour- and path-based common sense knowledge by initializing the path-based knowledge with the neighbour-based knowledge. GB-Net [35] leverages multiple KGs, i.e. ConceptNet, WordNet and Visual Genome, as sources of prior common sense knowledge. The method operates in an iterative manner, circulating data within and between a scene graph and a common sense graph, and enhancing their associations with each cycle. It sets up entity bridges by linking each scene entity to the common sense entity that aligns with the label predicted by Faster RCNN, followed by message dissemination among all nodes. It calculates the pairwise resemblance between every scene predicate node and every common sense predicate node, identifying pairs with maximum similarity to link scene predicates to their respective categories. This procedure is carried out for a predetermined number of iterations, with the final state of the bridge dictating the category to which each node is assigned, leading to the formation of the scene graph. KBGAN [36] used ConceptNet to retrieve and embed common sense knowledge for the refinement of object and phrase features in SGG through an attention-based knowledge fusion mechanism.

2.3. *Neurosymbolic Integration*

NeSy integration aims to combine neural and symbolic approaches to construct more powerful learning and reasoning approaches in AI. The neural approaches excel at identifying statistical patterns from data in raw form and are not susceptible to noise in data. However, these techniques are data-intensive and operate as black boxes, making their decision-making processes difficult to interpret. On the other hand, symbolic techniques excel at logical reasoning, offer high explainability and allow for the use of dynamic declarative languages for knowledge representation. However, they offer less trainability and can be brittle when faced with out-of-domain data. Given the complementary strengths and weaknesses, the integration of neural and symbolic techniques is a logical advancement towards AI approaches that are more robust, reliable and effective. A fine-grained classification of NeSy approaches with six different types is provided in [22, 23]. However, given the relatively few NeSy studies in the field of scene understanding and visual reasoning, we have streamlined the classification within this domain. We categorize NeSy approaches into two types, loosely coupled and tightly coupled [22], based on the degree of integration between the symbolic and neural components.

Table 3

Utilization of Knowledge Graphs for Infusing Common Sense Knowledge in SGG

Knowledge Graph	Nature of Knowledge	Dimensions	Examples	Methods Employed
ConceptNet [19]	Information about common objects, activities, relations, etc., in text format	8M nodes, 36 relations & 21M edges	(book, used for, reading), (pen, capable of, writing)	[33–36]
Visual Genome [13]	Visual information about image attributes, objects, and relations	3.8M nodes, 42k relations, 2.3M edges & 2.8M attributes	(cat, on, mat), (man, holding, umbrella)	[33, 35]
Wordnet [20]	Lexical information about words, concepts, relations, etc.	0.155M words, 10 relations & 0.176M synsets	(dog, has part, tail), (reading, part meronym, scanning)	[35]
CSKG [21]	Diverse common sense knowledge consolidated from seven distinct sources	2.16M nodes, 58 relations, 6M edges	(ball, located near, goalpost), (guitar, used for, playing music)	[32]

2.3.1. Loose Coupling

In the context of loosely-coupled NeSy approaches, the symbolic and neural components operate relatively independently, each adhering to its own processes and methodologies. While they interact as required, their operations are not deeply intertwined. This loose coupling allows each component to leverage its own strengths, while also benefiting from the capabilities of the other component. In some loosely-coupled NeSy approaches, the neural and symbolic elements concentrate on distinct yet complementary tasks within a large pipeline. They cooperate to accomplish the overall task, yet retain the ability to function independently. This arrangement harnesses the advantages of both neural and symbolic components while preserving the autonomy of each component when necessary. Such a setup proves particularly effective in handling complex tasks that necessitate a blend of symbolic reasoning and neural learning. The NeSy Concept Learner (NS-CL), proposed by Mao et al. [37], comprises a neural network designed to learn visual concepts and a symbolic module that processes symbolic programs on the features of visual concepts for the purpose of answering questions. The symbolic module furnishes feedback signals that facilitate the gradient-based optimization of the neural module. The symbolic and neural components function in a sequential chain of operations in several loosely-coupled NeSy approaches. The neural component transforms the raw input into a format that the symbolic component can process. The symbolic component processes the transformed input and subsequently passes its output back to the neural component for additional processing. This method allows for a fusion of symbolic reasoning and neural processing capabilities, despite the fact that each component largely maintains its own operational independence. For example, IRT-MSK [33] extracts knowledge-embedded semantic features from KGs to explore context information from visual features in SGG, with the symbolic and neural components operating relatively independently. MotifNet [28] uses a global context (symbolic) to inform each stage of its predictions (neural) in a sequential manner. The global context, represented via LSTMs, infuses symbolic knowledge into the neural component, enhancing its ability to make informed predictions. Khan et al. [32] proposed a loosely-coupled NeSy approach for SGG, knowledge enrichment and downstream visual reasoning. This approach leverages the representational and reasoning strengths of symbolic systems to represent images as scene graphs and supplement them with common sense knowledge. The neural modules employ the powerful learning capabilities of deep learning to predict semantic elements in images and to process the enhanced scene graphs for downstream visual reasoning tasks. In this setup, the neural and symbolic modules exhibit a degree of interdependence. The precision of the scene graph elements predicted by the neural module significantly impacts the effectiveness of the scene graph enrichment process, which in turn has a direct influence on the performance of downstream reasoning tasks.

2.3.2. Tight Coupling

The symbolic and neural components are deeply integrated within the tightly-coupled NeSy approaches, leveraging the strengths of both symbolic and neural approaches in a more unified way, potentially leading to improved performance and capabilities. In some tightly-coupled NeSy approaches, symbolic rules or knowledge are directly incorporated into the architecture or training regime of the neural networks. This means that the structure of the neural network or the way it is trained is influenced by the symbolic rules, leading to a deep integration of the symbolic and neural components. This allows the neural network to leverage the reasoning capabilities of the symbolic rules while also benefiting from the learning capabilities of the neural component. Some recent works [36, 38] employ GNNs to embed entities and relations from external knowledge bases, thereby enhancing performance in scene understanding and visual reasoning tasks. Moreover, several VQA methods [39–43] generate and execute symbolic programs, implemented as neural networks or fully differentiable operations, to answer questions. In several tightly-coupled NeSy systems, symbolic knowledge is encoded into the distributed representation of neural networks. This tight integration means that the data representation of the neural component is directly impacted by symbolic knowledge. This configuration allows the neural component to harness the reasoning capabilities inherent in the symbolic knowledge, while simultaneously leveraging its own learning capabilities. Such systems are particularly useful in tasks necessitating a deep understanding of the data, as it enables the system to exploit both the symbolic knowledge and the learning capabilities of neural networks. For instance, Li et al. [44] developed a hierarchical semantic segmentation network using compositional relations across semantic hierarchies as additional training targets. Zhou et al. [27] designed a three-module system for SGG, infusing statistical probabilities into the modules for a tightly-coupled NeSy approach. COACHER [34] integrated common sense knowledge for zero-shot relation prediction in SGG, embedding symbolic knowledge into the neural component's distributed representation. GB-Net [35] uses a graph-based neural network to refine connections between a scene graph and a common sense graph, exploiting the interconnected graphs' heterogeneous structure. KERN [29] incorporates statistical correlations between pairwise objects and visual relationships in deep neural networks, using a structured KG to propagate messages and explore object interactions. KBGAN [36] integrates symbolic knowledge from an external knowledge base into neural components for SGG, refining object and phrase features. The VRD model [30] integrates symbolic knowledge into CNNs to detect visual relationships from images. DeepVRL[31] employs a directed semantic action graph to capture semantic relationships. It utilizes a traversal structure with variations over the action graph and a scheme for mining objects that is aware of semantic ambiguity, which aids in distinguishing between object categories.

3. Downstream Visual Reasoning

Scene graphs are widely utilized in downstream visual reasoning tasks, including VQA, image captioning, MEP, image retrieval, and image generation, as shown in Figure 3. The efficacy of these downstream tasks is determined by the quality and expressiveness of the generated scene graphs. This section provides an overview of visual reasoning methods based on scene graphs and prior common sense knowledge.

3.1. Visual Question Answering

Models for VQA predict the most appropriate responses to queries regarding visual scenes by leveraging multimodal features, semantic relationships in scene graphs and factual knowledge. For instance, Zhang et al. [45] suggested the incorporation of scene graph structural data into GNNs, utilizing it as a basis for VQA. In a parallel work, Ziaeefard et al. [46] presented a VQA methodology based on GATs, which encodes scene graphs in conjunction with external knowledge derived from ConceptNet. Wu et al. [47] proposed a method that combines image content representation with a common knowledge base for image-based question answering, but it lacks explicit reasoning. Narasimhan et al. [48] developed a retrieval method based on learning, which embeds both facts and pairs of questions and images into a common space, thereby converting visual concepts into a vector in proximity to pertinent facts. Garderes et al. presented ConceptBert [49], a model that amalgamates pre-trained image and language features with embeddings from a knowledge graph, eliminating the need for external knowledge annotations or search queries. Shevchenko et al. [50] put forward a method that incorporates information from a knowledge base into a transformer for visual and language data, ensuring alignment between the learned representation and the knowledge embedding. Zhu et al. [51] developed Mucko, a model designed for multilayer cross-modal knowledge reasoning that builds a multimodal heterogeneous graph and employs a modality-aware heterogeneous graph convolutional network for the capture of evidence. Yu et al. [52] introduced a model that deciphers images using

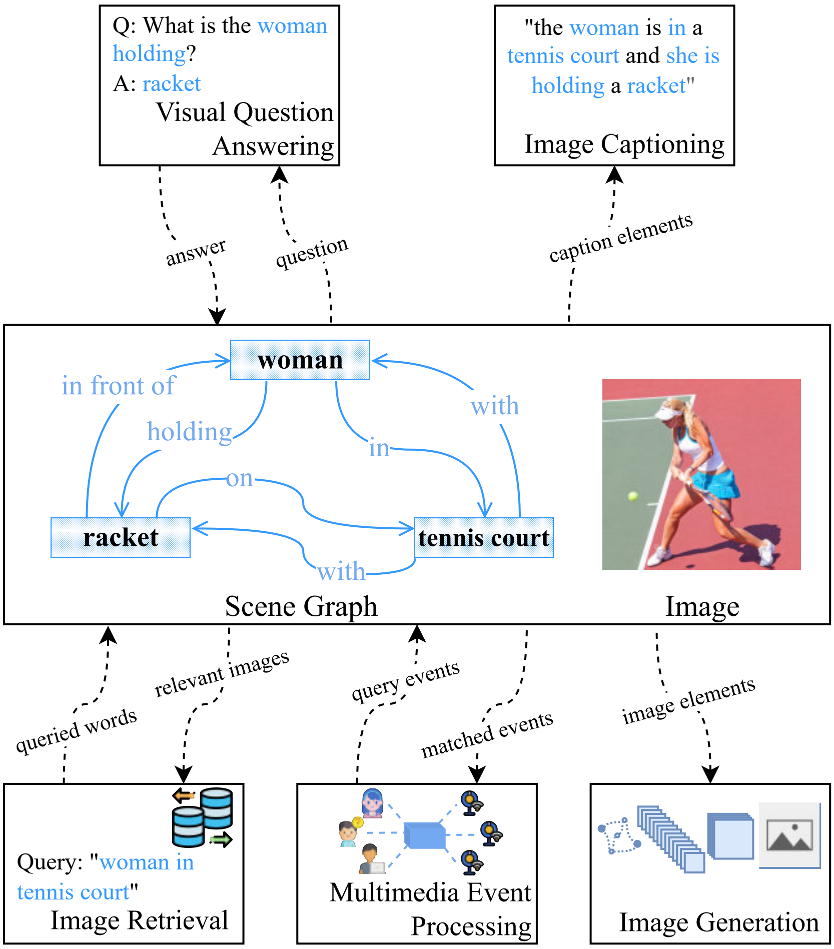

Fig. 3. An overview of the downstream visual reasoning tasks of scene graph generation, including VQA, image captioning, MEP, image retrieval, and image generation.

a multimodal knowledge graph and a memory-based recurrent network for cross-modal reasoning, representing knowledge across different modalities and selecting knowledge relevant to the problem for predicting answers. Anderson et al. [53] used Faster R-CNN to propose image regions and integrated bottom-up and top-down attention mechanisms to enhance the interpretability of attention weights and unified visual-linguistic understanding for VQA. Tan et al. [54] proposed the LXMERT framework employing a large-scale Transformer model with three encoders for scene graph-based VQA based on the understanding of visual concepts and language semantics, as well as, intra- and cross-modal relationships. Meta Module Network (MNM) [55] addresses the scalability and generalizability in VQA using a metamorphic meta module, which dynamically morphs into diverse instance modules, offers flexibility and allows for complex visual reasoning, while preserving the same model complexity as the function set expands. MDETR [56] is an end-to-end modulated detector that leverages a transformer-based architecture to fuse image and text modalities at an early stage for efficient extraction of visual concepts from the free-form text in multi-modal reasoning systems including VQA. Zhang et al. [57] performed visual reasoning for VQA based on their object detection model designed for visual-language tasks with richer visual representations of objects and concepts.

Among the scene graph-based VQA methods, Hudson et al. [41] presented a visual reasoning approach based on Neural State Machines (NSM) integrating visual and linguistic inputs into semantic concepts via a probabilistic scene graph for sequential reasoning and inference. Zhang et al. [45] embedded the structural features of scene graphs into a GNN for downstream VQA. Yang et al. [58] proposed Scene Graph Convolutional Network (SceneGCN) that incorporates object properties and semantic relationships into a structured scene representation for enhanced VQA via visual context and language priors. Graphhopper [59] addresses the challenge of performing multi-hop K reasoning over complex visual scenes to predict reasoning paths that lead to the answer in VQA. Dual Message-passing enhanced GNN (DM-GNN) [60] encodes multi-scale scene graph information into two diversified graphs focused on objects and relations, and uses a dual structure to encode them to achieve a balanced representation of object, relation, and attribute features in VQA. The Scene Graph Refinement network (SGR) [61] propose a transformer-based refinement network to enhance object and relation feature learning in VQA, utilizing question semantics to jointly learn multimodal representations and select the most relevant relations for better question answering.

3.2. Image Captioning

Scene graphs have been employed in image captioning techniques to generate effective scene descriptions, overcoming the limitations of relying solely on vision-language features. The Abstract Scene Graph (ASG) [62] approach, for instance, encodes the intentions of users and semantic information into scene graphs, facilitating the creation of diverse and desired text descriptions of scenes. The SMP method [15] generates scene graphs based on the saliency of visual relationships, which are then utilized for enhanced caption generation. Goel et al. [63] proposed that image captioning models could be improved by integrating prior knowledge through conditional latent topic attention and leveraging the semantic and syntactic structure of captions in regularization. The methods yield more human-like captions and significant improvements on the MSCOCO dataset, even in low data situations. The techniques are applicable to other vision tasks with limited data, indicating their potential for improving generalization. The R-SCAN model [64] explored learning visual relationship features in SGG for vision-and-language tasks and proposed pre-training SGG models with visual relationship data relevant to the scene. Liao et al. [65] proposed a 3D Scene Graph-based Change Captioning (SGCC) model to improve object location accuracy in change captioning tasks. Wu et al. [66] introduced a method incorporating high-level visual concepts and external knowledge into the deep learning cascade comprising CNN and RNN for enhanced image captioning and VQA performance. Yu et al. [67] developed a 3D-SceneCaptioner, a point clouds-based image captioning technique, which generates more accurate captions by fully utilizing the useful semantic information in point clouds. Zhang et al. [68] improved the image captioning performance of a transformer by leveraging a knowledge graph and augmenting maximum likelihood estimation with a Kullback-Leibler divergence term. A recent study [69] demonstrated that image captioning methods utilizing information extracted from knowledge graphs significantly outperformed the methods relying entirely on image information.

3.3. Other tasks

NeSy visual semantic models have found applications in the representation of multimedia streams for real-time multimodal event processing in the Internet of Multimedia Things (IoMT) [70, 71]. These models utilize deep neural networks for object and attribute detection, and symbolic rules are applied to discern spatiotemporal relations among the objects. These interactions are crucial for correlating high-level events queried by users. In the context of image retrieval, scene graphs serve to explicitly articulate the semantics and structured data of images, enabling efficient retrieval of images from extensive databases based on their contents. Schroeder et al. [72] introduced Structured Query-based Image Retrieval (SQIR), a method that represents visual relationships in scene graphs as directed subgraphs. This approach facilitates the task of graph matching in image retrieval using structured queries and scene graph embeddings. Ward et al. [73] proposed a NeSy approach based on deep learning and knowledge graphs to guide the colourization of black-and-white images, leveraging object classification and contextual knowledge to determine accurate colours for both simple and complex scenes. Compared to textual scene descriptions, scene graphs have proven to be more efficient and adaptable for image generation, especially when the number of objects and relationships increase [74]. Scene graphs, when infused with common sense knowledge, have been utilized in a scene graph-based image generation network, leading to the creation of more realistic images [32]. Gu et al. [36] made use of ConceptNet to refine objects and phrases based on prior common sense knowledge in an attention-based RNN technique for the reconstruction of images from scene graph representations.

4. Performance Evaluation

In this section, we present the benchmark datasets and standard metrics used for the performance evaluation of SGG and visual reasoning methods.

4.1. Scene Graph Generation

The knowledge-based SGG approaches and common datasets used for evaluation are summarized in Table 4 and Table 5, respectively. The benchmark dataset frequently employed for SGG evaluation is Visual Genome [13]. The standard metrics used to evaluate relationship prediction in SGG include Recall@K ($R@K$), mean Recall@K ($mR@K$), and zero-shot Recall@K ($zR@K$).

- $R@K$ is the proportion of instances where the correct relationship is among the top K relationship predictions with the highest confidence [30]. This metric requires not only accurate relationship label prediction but also a high confidence score.
- $mR@K$ is the average of $R@K$ values, each computed separately for every relationship category. This metric is designed to reduce evaluation bias towards frequently occurring relationships [29, 75].
- $zR@K$ is similar to $R@K$, but it is only computed for relationships that do not appear in the training dataset [30, 76].

Table 4
State-of-the-art Knowledge-based SGG methods evaluated using standard metrics on Visual Genome dataset

Method	Deep Learning Architecture	Common Sense Knowledge Source	Neurosymbolic Integration	SGG Performance		
				R@50/100	mR@50/100	zR@50/100
SGG-CKI [32]	CNN and LSTM	CSKG	Loose Coupling	**35.5/39.1**	**10.9/12.6**	-/-
DSGAT [27]	CNN and GAT	Statistical Prior	Tight Coupling	28.8/32.9	8.9/11.8	-/-
IRT-MSK[33]	CNN and GCN	ConceptNet and Visual Genome	Loose Coupling	27.8/31.0	-/-	-/-
COACHER[34]	CNN and LSTM	ConceptNet	Tight Coupling	-/-	-/-	**19.3/22.2**
MotifNet [28]	CNN and LSTM	Statistical Prior	Loose Coupling	27.2/30.3 (22.6/25.9*)	5.7/6.6 (5.2/6.3*)	19.0/21.9
GB-Net [35]	CNN and GNN	ConceptNet, WordNet and Visual Genome	Tight Coupling	26.4/30.0	6.1/7.3	-/-
KERN [29]	CNN and GNN	Statistical Prior	Tight Coupling	27.1/29.8	6.4/7.3	-/-
KB-GAN [36]	CNN and GRU	ConceptNet	Tight Coupling	13.6/17.6	-/-	18.1/21.1
DeepVRL[31]	CNN and DQN	Language Prior	Tight Coupling	13.3/12.6	-/-	6.3/7.1
VRD [30]	CNN	Language Prior	Tight Coupling	0.3/0.5	-/-	-/-

*on GQA dataset [77]

4.2. Downstream Reasoning Tasks

4.2.1. Visual Question Answering

The commonly used datasets for scene graph-based VQA include GQA [77], MS COCO [82], and Visual Genome [13]. The GQA dataset [77] is the standard dataset for scene graph-based VQA. It contains 113,018 images, 22 million questions, 1702 object classes and 310 relationship types, with an 80-10-10 split for training, validation and testing. The "binary" type questions are designed to have a 'yes' or 'no' answer, for example, questions that involve checking the presence, absence, or relationship between objects in the image. On the other hand, the "open" type questions require a more elaborate answer that needs deeper reasoning about the semantics of the visual content,

Table 5

Datasets for Evaluation of SGG and Downstream Reasoning Methods

Dataset	Size	Annotations for Scene Graph Generation		Annotations for Downstream Reasoning		External Knowledge
		Object categories	Relationship categories	Image Captions	Question-Answer Pairs	
Visual Genome [13]	108K images	33.8K	42K	✓	✓	✗
VG150 [78]	88K images	150	50	✓	✓	✗
VG200 [79]	99K images	200	100	✓	✓	✗
VG80k [80]	100K images	53K	29K	✓	✓	✗
VG-MSDN [81]	95K images	150	50	✓	✓	✗
MS COCO [82]	330K images	80	–	✓	✓	✗
Flickr30K [83]	30K images	–	–	✓	✗	✗
GQA [77]	113K images	1.7K	310	✗	✓	✗
VQA-v2 [84]	204K images	–	–	✗	✓	✗
VCR [85]	110K images	–	–	✗	✓	✗
KB-VQA [86]	700 images	–	–	✗	✓	✓
FVQA dataset [87]	2190 images	–	–	✗	✓	✓
OK-VQA [88]	14K images	–	–	✗	✓	✓
KRVQA [89]	33K images	–	–	✗	✓	✓

usually involving identifying, describing, or explaining objects and relationships in the image. Apart from the standard accuracy metric, the new performance metrics introduced in GQA are more robust to informed guesses as they need a deeper semantic understanding of questions and visual content. The following performance metrics [77] are used to quantify the reasoning capabilities of the VQA methods:

- “Accuracy" (Top-1) is the fraction of times the predicted answer with the highest probability matches the groundtruth; it is separately calculated for binary and open questions.
- “Consistency" measures the ability to answer multiple related questions consistently, indicating the level of understanding of the semantics of a question within the scene.
- “Validity" evaluates whether an answer aligns with the scope of the question, reflecting the ability to comprehend the question.
- “Plausibility" measures if an answer is reasonable within the context of the question and in line with real-world knowledge.
- “Distribution" (lower is better) checks the match between the distributions of predicted answers and groundtruth, showing the ability to predict the less frequent answers in addition to the common ones.

There are several knowledge-based datasets available for VQA. The KB-VQA dataset [86] evaluates the ability of a VQA model to answer questions requiring external knowledge. It comprises 2,402 questions generated from 700 MS COCO images, each question falling into one of three categories: visual, common-sense, or KB-knowledge. The FVQA dataset [87] pairs questions and answers with supporting facts in a structured triplet format, using a knowledge base built from DBpedia [90], WebChild [91, 92] and ConceptNet [19]. It includes 2,190 images, 5,286 questions, and 193,449 facts, with questions categorized by visual concept type, answer source, and supporting knowledge base. The OK-VQA dataset [88], comprising 14,031 images and 14,055 questions, requires reasoning based on uninstructed knowledge, unlike fact-based VQA datasets like KB-VQA and FVQA. Questions are categorized into one of 10 knowledge categories, or "Other" if they don’t fit into any specific category. The KRVQA dataset [89], the first large-scale set requiring knowledge reasoning on natural images, includes 32,910 images, 157,201 question-answer pairs, and 194,449 knowledge triplets. Questions are categorized by reasoning steps and knowledge involvement, and the dataset is built on the scene graph annotations of the Visual Genome dataset [13] and the knowledge base of FVQA dataset [87]. Although these datasets contain external knowledge to some extent, KB-VQA [86] and FVQA [87] have insufficient size and annotations for comprehensive visual reasoning and all these datasets lack scene graph annotations, ignoring the structural and relational features of visual concepts that are crucial for visual reasoning.

4.2.2. Image Captioning

The performance evaluation of scene graph-based image captioning methods is usually based on MS COCO [82], Flickr30k[83], and Visual Genome [13] datasets. Various metrics are used to assess the quality of the generated image captions, each focusing on different aspects.

- The BLEU score [93], originally developed for machine translation, measures the n-gram precision between sentences, considering n-grams up to a length of four. It is generally more suitable for comparing entire corpora rather than individual sentences.
- The METEOR score [94], another metric from the machine translation field, emphasizes the recall of matching unigrams from the candidate and reference sentences. It accounts for word alignment in their exact form, stemmed form, and semantics, making it particularly effective for corpus-level comparisons.
- The ROUGE score [95], initially designed for text summarization, and its variant ROUGE-L are frequently used in caption generation. ROUGE-L identifies the longest subsequence of tokens in the same relative order, potentially with other tokens in between, that exists in both the candidate and reference caption.
- The CIDEr score [96], specifically created for caption generation evaluation, calculates the cosine similarity between the Term Frequency-Inverse Document Frequency (TF-IDF) weighted n-grams in the candidate caption and the group of reference captions linked with the image. It considers both precision and recall.
- The SPICE score [97], the most recent evaluation metric, correlates best with human judgements and is particularly relevant for scene graph-based image captioning evaluation. The SPICE score considers matching tuples retrieved from the candidate and reference scene graphs. As a result, it favours semantic information over text fluency and more closely mirrors human judgment.

5. Challenges and Prospects

In this section, we present the main challenges faced by the existing knowledge-based SGG and visual reasoning methods, as well as future directions for addressing the challenges.

5.1. Contextual Relevance of Knowledge

Common sense knowledge has been shown to improve the accuracy and expressiveness of SGG and visual reasoning [6]. However, KGs, which are often used as a source of this common sense knowledge, have their own limitations, especially when it comes to understanding the context of a specific scene. KGs may not always supply contextually appropriate information about visual concepts in a specific scene due to their inherent contextual limitations [98]. For example, while a KG might correctly identify that birds "fly" and fish "swim," it might struggle in a scene where a bird is depicted as "swimming" in water after a dive for fish. The KG might not provide the most contextually appropriate information in such cases, leading to potential inaccuracies in the scene graph. Similarly, language priors and statistical priors can also have contextually limited or incorrect knowledge due to their inherent limitations [6]. Despite efforts to infuse relevant knowledge based on the semantic and structural similarity of concepts, the contextual relevance of external knowledge often remains overlooked. This results in the infusion of irrelevant knowledge, thereby restricting the contextual reasoning capability in downstream visual reasoning tasks. Moreover, current evaluation methods for SGG and downstream reasoning tasks do not assess directly the accuracy and relevance of this external knowledge.

These shortcomings underscore the need for new evaluation metrics capable of assessing the quality of knowledge infusion based on the proportion of accurate and contextually relevant knowledge integrated into neural networks. Additionally, the use of context-aware approaches [99, 100] can ensure that only relevant and contextually valid knowledge is added during the infusion process, leading to improved downstream visual reasoning. Future research in this area could explore approaches with feedback mechanisms [101], adaptive thresholds [102], and domain-specific knowledge [103]. For instance, feedback mechanisms can dynamically adjust the knowledge infusion process based on the performance of downstream tasks, ensuring that the knowledge remains relevant and useful. Adaptive thresholds can help fine-tune the amount and type of knowledge infused based on the specific requirements of the scene or downstream task at hand. Furthermore, integrating domain-specific knowledge can address specialized requirements within visual reasoning, ensuring that the knowledge is both broad-based for general contexts and tailored for specific scenarios. Such approaches will ensure the infused knowledge is contextually relevant in addition to being semantically and structurally related to the scene, leading to more reliable and precise scene representation and visual reasoning.

5.2. Bias and Generalizability

A main cause of the limited performance of existing SGG methods is the long-tailed distribution of crowdsourced datasets [5], restricting the SGG methods from generalizing to rare visual relationships. Many relationship predicates that carry significant meaning are underrepresented, making it challenging for SGG methods to learn their feature representations. Conversely, frequently occurring predicates are often quite generic and do not clearly express the actual visual relationships compared to less common predicates. Moreover, visual feature representations of relationships can significantly differ across various scenes, adding another layer of complexity [26]. Given the impracticality of collection and annotation of enough training examples for object-predicate combinations representing all possible visual relationships, there is a clear need to explore zero-shot approaches and augment the conventional data-driven SGG techniques with external common sense knowledge. Pivoting towards zero-shot and knowledge-centric strategies will enhance the prediction of unseen or infrequent visual relationships to improve generalizability in addition to solving the long-tailed distribution problem.

Approaches such as zero-shot [34, 104] and few-shot learning [33] have been investigated to address these challenges in SGG. Zero-shot learning leverages previously learned relationships to recognize visual relationships that have not been seen before. Conversely, few-shot learning utilizes a small number of labelled samples to learn new relationships, which is advantageous when the collection of extensive labelled training data is tedious, costly or impractical. By harnessing the power of heterogeneous KGs, these techniques can seamlessly integrate common sense knowledge, facilitating the extraction of relevant relationship triplets, and thereby enhancing the prediction of infrequent and unseen visual relationships. Additionally, knowledge transfer and distillation techniques [105, 106] present another promising direction. Previously learned visual relationships can be leveraged by employing models trained on diverse common sense knowledge bases, enhancing the generalization capabilities and practicality of SGG, and ensuring it remains relevant and effective in real-world scenarios.

5.3. Leveraging Heterogeneous Knowledge

Most existing techniques rely on statistical and language priors, as well as KGs, as sources of external common sense knowledge. However, the heuristic nature of statistical priors limits their generalizability, and the limitations of semantic word embeddings can impact the performance of language priors, especially when dealing with unseen or infrequent relationships. Individual KGs, such as WordNet [20] and ConceptNet [19], which have been employed in SGG, provide lexical and text-based knowledge, encapsulating a variety of common sense forms and notions. However, they fall short of providing a comprehensive understanding of visual concepts. In contrast, heterogeneous KGs, like CSKG [21], cover a significantly broader spectrum of common sense dimensions. Heterogeneous KGs are presently the most diverse and comprehensive repositories of common sense knowledge, encapsulating intricate structural and semantic characteristics of general concepts. The incorporation of these heterogeneous KGs to augment scene graphs has shown promising results in improving the overall efficacy of SGG within a loosely-integrated NeSy approach [32]. However, their application in tightly-coupled SGG approaches and mainstream visual reasoning tasks, such as VQA and image captioning, remains unexplored. These heterogeneous sources are essential but underutilized in the infusion of prior common sense knowledge in this field. Carefully integrating the heterogeneous KGs has the potential to deepen the interpretation of complex scenes, leading to comprehensive and precise scene representations for intuitive visual reasoning.

The integration of heterogeneous common sense knowledge directly into the structure or feedback mechanisms of deep neural networks for SGG can be an effective approach [107, 108]. This strategy can empower deep neural networks to learn the nuances of visual relationships more effectively, leading to more precise SGG that may eliminate the need for subsequent scene graph refinement. While some research has ventured into this area [35, 36], further exploration is needed to understand how the utilization of heterogeneous common sense knowledge can mitigate the challenges associated with SGG. Additionally, heterogeneous KGs can be instrumental in deriving rules about visual concepts and incorporating them into the learning process of deep neural networks [109, 110] for scene understanding and visual reasoning. Continued research in this direction could unlock the full potential of infusing common sense knowledge into scene understanding and visual reasoning techniques.

5.4. Temporal Relationships

The existing methods are proficient at processing images to extract semantic elements, infer visual relationships, and infuse common sense knowledge for enhanced downstream reasoning. However, these methods fall short when it comes to video data, where visual relationships can change over time. Current knowledge-based SGG methods can only process each video frame individually, which is computationally inefficient and overlooks the temporal patterns of visual relationships. This approach demands high computational resources and misses out on capturing the temporal dynamics of visual relationships. While there have been attempts to develop SGG methods for video data [111, 112] and corresponding datasets [113, 114], there is an opportunity to integrate external common sense knowledge into these methods.

Addressing these gaps requires a multi-faceted approach. Firstly, object tracking [115] can be integrated to maintain continuity in recognizing and following objects across frames. This ensures that the system understands the trajectories of objects and their interactions over time, rather than treating each appearance as isolated. Secondly, the temporal aspects of visual relationships [8] need to be incorporated. This would allow the system to understand sequences, patterns, and changes in relationships over time, offering a richer interpretation of video content. For instance, understanding the temporal relationship can help discern if a person is "picking up" or "putting down" an object. Thirdly, graph aggregation techniques [116, 117] can be employed to consolidate information from multiple frames into a unified scene graph. This would provide a holistic view of the video, capturing both spatial and temporal relationships in a compact representation. Such advancements would enhance scene understanding in videos and open doors to novel applications. For instance, by leveraging temporal dynamics and infusing common sense knowledge, systems could detect congestion patterns in traffic videos or pinpoint unusual activities in surveillance footage of smart cities [118]. This would be invaluable for many domains, including urban planning and security.

6. Conclusion

The integration of deep learning and common sense knowledge through neurosymbolic integration for scene representation and visual reasoning is a promising research direction. We investigated this research direction in detail by reviewing and classifying state-of-the-art knowledge-based neurosymbolic techniques for scene representation and discussing relevant datasets, evaluation methods, key challenges, and future research directions. The survey serves as a valuable resource for future research in the development of more effective scene representation and visual reasoning techniques at the intersection of deep learning, knowledge infusion and neurosymbolic integration.

References

[1] P. Hitzler, F. Bianchi, M. Ebrahimi and M.K. Sarker, Neural-symbolic integration and the semantic web, *Semantic Web* **11**(1) (2020), 3–11.

[2] A. Bennetot, J.-L. Laurent, R. Chatila and N. Díaz-Rodríguez, Towards explainable neural-symbolic visual reasoning, *arXiv preprint arXiv:1909.09065* (2019).

[3] W.W. Cohen, H. Sun, R.A. Hofer and M. Siegler, Scalable neural methods for reasoning with a symbolic knowledge base, *arXiv preprint arXiv:2002.06115* (2020).

[4] P. Hitzler, A. Eberhart, M. Ebrahimi, M.K. Sarker and L. Zhou, Neuro-symbolic approaches in artificial intelligence, *National Science Review* **9**(6) (2022), nwac035.

[5] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen and A. Hauptmann, Scene Graphs: A Survey of Generations and Applications, *arXiv preprint arXiv:2104.01111* (2021).

[6] M.J. Khan, J.G. Breslin and E. Curry, Common Sense Knowledge Infusion for Visual Understanding and Reasoning: Approaches, Challenges, and Applications, *IEEE Internet Computing* **26**(4) (2022), 21–27.

[7] M.J. Khan, J. Breslin and E. Curry, NeuSyRE: Neuro-Symbolic Visual Understanding and Reasoning Framework based on Scene Graph Enrichment, *Semantic Web* (2023).

[8] 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, 2020, pp. 9185–9192.

[9] J. Sun, H. Sun, T. Han and B. Zhou, Neuro-symbolic program search for autonomous driving decision module design, in: *Conference on Robot Learning*, PMLR, 2021, pp. 21–30.

[10] A. Paliwal, S. Loos, M. Rabe, K. Bansal and C. Szegedy, Graph representations for higher-order logic and theorem proving, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 2020, pp. 2967–2974.

[11] T. Silver, A. Athalye, J.B. Tenenbaum, T. Lozano-Perez and L.P. Kaelbling, Learning neuro-symbolic skills for bilevel planning, *arXiv preprint arXiv:2206.10680* (2022).

[12] M. Hassan, H. Guan, A. Melliou, Y. Wang, Q. Sun, S. Zeng, W. Liang, Y. Zhang, Z. Zhang, Q. Hu et al., Neuro-Symbolic Learning: Principles and Applications in Ophthalmology, *arXiv preprint arXiv:2208.00374* (2022).

[13] 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* **123**(1) (2017), 32–73.

[14] X. Lin, C. Ding, Y. Zhan, Z. Li and D. Tao, HL-Net: Heterophily Learning Network for Scene Graph Generation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19476–19485.

[15] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei and C.-W. Chen, Boosting Scene Graph Generation with Visual Relation Saliency, *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[16] 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).

[17] K. Ye and A. Kovashka, Linguistic Structures As Weak Supervision for Visual Scene Graph Generation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8289–8299.

[18] 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.

[19] R. Speer, J. Chin and C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: *Thirty-first AAAI conference on artificial intelligence*, 2017, pp. 4444–4451.

[20] G.A. Miller, WordNet: a lexical database for English, *Communications of the ACM* **38**(11) (1995), 39–41.

[21] F. Ilievski, P. Szekely and B. Zhang, Cskg: The commonsense knowledge graph, in: *European Semantic Web Conference*, Springer, 2021, pp. 680–696.

[22] A.d. Garcez and L.C. Lamb, Neurosymbolic AI: The 3rd wave, *Artificial Intelligence Review* (2023), 1–20.

[23] W. Wang and Y. Yang, Towards Data-and Knowledge-Driven Artificial Intelligence: A Survey on Neuro-Symbolic Computing, *arXiv preprint arXiv:2210.15889* (2022).

[24] F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D.L. McGuinness and P. Szekely, Dimensions of commonsense knowledge, *arXiv preprint arXiv:2101.04640* (2021).

[25] U. Kursuncu, M. Gaur and A. Sheth, Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning, *arXiv preprint arXiv:1912.00512* (2019).

[26] G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, M. Feng, X. Zhao, Q. Miao, S.A.A. Shah et al., Scene graph generation: A comprehensive survey, *arXiv preprint arXiv:2201.00443* (2022).

[27] H. Zhou, Y. Yang, T. Luo, J. Zhang and S. Li, A unified deep sparse graph attention network for scene graph generation, *Pattern Recognition* **123** (2022), 108367.

[28] 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.

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

[30] 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.

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

[32] M.J. Khan, J.G. Breslin and E. Curry, Expressive Scene Graph Generation Using Commonsense Knowledge Infusion for Visual Understanding and Reasoning, in: *European Semantic Web Conference*, Springer, 2022, pp. 93–112.

[33] Y. Guo, J. Song, L. Gao and H.T. Shen, One-shot Scene Graph Generation, in: *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3090–3098.

[34] X. Kan, H. Cui and C. Yang, Zero-shot scene graph relation prediction through commonsense knowledge integration, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2021, pp. 466–482.

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

[36] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1969–1978.

[37] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum and J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, *arXiv preprint arXiv:1904.12584* (2019).

[38] K. Marino, X. Chen, D. Parikh, A. Gupta and M. Rohrbach, Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14111–14121.

[39] S. Amizadeh, H. Palangi, A. Polozov, Y. Huang and K. Koishida, Neuro-symbolic visual reasoning: Disentangling, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 279–290.

[40] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick and R. Girshick, Inferring and executing programs for visual reasoning, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2989–2998.

[41] D. Hudson and C.D. Manning, Learning by abstraction: The neural state machine, *Advances in Neural Information Processing Systems* **32** (2019).

[42] J. Shi, H. Zhang and J. Li, Explainable and explicit visual reasoning over scene graphs, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8376–8384.

[43] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra and D. Parikh, Probabilistic neural symbolic models for interpretable visual question answering, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 6428–6437.

[44] L. Li, T. Zhou, W. Wang, J. Li and Y. Yang, Deep hierarchical semantic segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1246–1257.

[45] 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).

[46] M. Ziaeefard and F. Lécué, Towards Knowledge-Augmented Visual Question Answering, in: *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1863–1873.

[47] Q. Wu, P. Wang, C. Shen, A. Dick and A. Van Den Hengel, Ask me anything: Free-form visual question answering based on knowledge from external sources, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4622–4630.

[48] M. Narasimhan and A.G. Schwing, Straight to the facts: Learning knowledge base retrieval for factual visual question answering, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 451–468.

[49] F. Gardères, M. Ziaeefard, B. Abeloos and F. Lecue, Conceptbert: Concept-aware representation for visual question answering, in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 489–498.

[50] V. Shevchenko, D. Teney, A. Dick and A.v.d. Hengel, Reasoning over vision and language: Exploring the benefits of supplemental knowledge, *arXiv preprint arXiv:2101.06013* (2021).

[51] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu and Q. Wu, Mucko: multi-layer cross-modal knowledge reasoning for fact-based visual question answering, *arXiv preprint arXiv:2006.09073* (2020).

[52] J. Yu, Z. Zhu, Y. Wang, W. Zhang, Y. Hu and J. Tan, Cross-modal knowledge reasoning for knowledge-based visual question answering, *Pattern Recognition* **108** (2020), 107563.

[53] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[54] H. Tan and M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, *arXiv preprint arXiv:1908.07490* (2019).

[55] W. Chen, Z. Gan, L. Li, Y. Cheng, W. Wang and J. Liu, Meta module network for compositional visual reasoning, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 655–664.

[56] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra and N. Carion, Mdetr-modulated detection for end-to-end multi-modal understanding, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.

[57] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi and J. Gao, Vinvl: Revisiting visual representations in vision-language models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5579–5588.

[58] Z. Yang, Z. Qin, J. Yu and T. Wan, Prior visual relationship reasoning for visual question answering, in: *2020 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2020, pp. 1411–1415.

[59] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp and S. Günnemann, Graphhopper: Multi-hop Scene Graph Reasoning for Visual Question Answering, in: *International Semantic Web Conference*, Springer, 2021, pp. 111–127.

[60] H. Li, X. Li, B. Karimi, J. Chen and M. Sun, Joint learning of object graph and relation graph for visual question answering, in: *2022 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2022, pp. 01–06.

[61] T. Qian, J. Chen, S. Chen, B. Wu and Y.-G. Jiang, Scene graph refinement network for visual question answering, *IEEE Transactions on Multimedia* (2022).

[62] S. Chen, Q. Jin, P. Wang and Q. Wu, Say as you wish: Fine-grained control of image caption generation with abstract scene graphs, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9962–9971.

[63] A. Goel, B. Fernando, T.-S. Nguyen and H. Bilen, Injecting prior knowledge into image caption generation, in: *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, Springer, 2020, pp. 369–385.

[64] 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).

[65] Z. Liao, Q. Huang, Y. Liang, M. Fu, Y. Cai and Q. Li, Scene graph with 3D information for change captioning, in: *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5074–5082.

[66] Q. Wu, C. Shen, P. Wang, A. Dick and A. Van Den Hengel, Image captioning and visual question answering based on attributes and external knowledge, *IEEE transactions on pattern analysis and machine intelligence* **40**(6) (2017), 1367–1381.

[67] Q. Yu, X. Pan, S. Xiang and C. Pan, 3D-SceneCaptioner: Visual Scene Captioning Network for Three-Dimensional Point Clouds, in: *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part II*, Springer, 2021, pp. 275–286.

[68] Y. Zhang, X. Shi, S. Mi and X. Yang, Image captioning with transformer and knowledge graph, *Pattern Recognition Letters* **143** (2021), 43–49.

[69] Y. Zhou, Y. Sun and V. Honavar, Improving Image Captioning by Leveraging Knowledge Graphs, in: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 283–293. doi:10.1109/WACV.2019.00036.

[70] M.J. Khan and E. Curry, Neuro-symbolic Visual Reasoning for Multimedia Event Processing: Overview, Prospects and Challenges., in: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'2020) Workshops*, 2020.

[71] E. Curry, D. Salwala, P. Dhingra, F.A. Pontes and P. Yadav, Multimodal Event Processing: A Neural-Symbolic Paradigm for the Internet of Multimedia Things, *IEEE Internet of Things Journal* (2022).

[72] 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.

[73] R. Ward, M.J. Khan, J.G. Breslin and E. Curry, Knowledge-Guided Colorization: Overview, Prospects and Challenges, in: *17th International Workshop on Neural-Symbolic Learning and Reasoning*, 2023.

[74] 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.

[75] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6619–6628.

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

[77] D.A. Hudson and C.D. Manning, Gqa: A new dataset for real-world visual reasoning and compositional question answering, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[78] D. Xu, Y. Zhu, C.B. Choy and L. Fei-Fei, Scene graph generation by iterative message passing, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5410–5419.

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

[80] 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.

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

[82] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C.L. Zitnick and P. Dollár, Microsoft COCO: Common Objects in Context, 2015.

[83] B.A. Plummer, L. Wang, C.M. Cervantes, J.C. Caicedo, J. Hockenmaier and S. Lazebnik, Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[84] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra and D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

[85] R. Zellers, Y. Bisk, A. Farhadi and Y. Choi, From recognition to cognition: Visual commonsense reasoning, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731.

[86] P. Wang, Q. Wu, C. Shen, A.v.d. Hengel and A. Dick, Explicit knowledge-based reasoning for visual question answering, *arXiv preprint arXiv:1511.02570* (2015).

[87] P. Wang, Q. Wu, C. Shen, A. Dick and A. Van Den Hengel, Fvqa: Fact-based visual question answering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(10) (2017), 2413–2427.

[88] K. Marino, M. Rastegari, A. Farhadi and R. Mottaghi, Ok-vqa: A visual question answering benchmark requiring external knowledge, in: *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[89] Q. Cao, B. Li, X. Liang, K. Wang and L. Lin, Knowledge-routed visual question reasoning: Challenges for deep representation embedding, *IEEE Transactions on Neural Networks and Learning Systems* **33**(7) (2021), 2758–2767.

[90] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, Dbpedia: A nucleus for a web of open data, in: *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, Springer, 2007, pp. 722–735.

[91] N. Tandon, G. De Melo, F. Suchanek and G. Weikum, Webchild: Harvesting and organizing commonsense knowledge from the web, in: *Proceedings of the 7th ACM international conference on Web search and data mining*, 2014, pp. 523–532.

[92] N. Tandon, G. Melo and G. Weikum, Acquiring comparative commonsense knowledge from the web, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28, 2014.

[93] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[94] S. Banerjee and A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[95] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text summarization branches out*, 2004, pp. 74–81.

[96] R. Vedantam, C. Lawrence Zitnick and D. Parikh, Cider: Consensus-based image description evaluation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[97] P. Anderson, B. Fernando, M. Johnson and S. Gould, Spice: Semantic propositional image caption evaluation, in: *European conference on computer vision*, Springer, 2016, pp. 382–398.

[98] A. Ettorre, A. Bobasheva, C. Faron and F. Michel, A systematic approach to identify the information captured by Knowledge Graph Embeddings, in: *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2021, pp. 617–622.

[99] N. Heist, Towards Knowledge Graph Construction from Entity Co-occurrence, in: *EKAW (Doctoral Consortium)*, 2018.

[100] S. Moon, P. Shah, A. Kumar and R. Subba, Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 845–854.

[101] G. Tamašauskaitė and P. Groth, Defining a knowledge graph development process through a systematic review, *ACM Transactions on Software Engineering and Methodology* **32**(1) (2023), 1–40.

[102] M. Qiao, H. Gui and K. Tang, Recommender system based on adaptive threshold filtering GCN, in: *International Conference on Neural Networks, Information, and Communication Engineering (NNICE)*, Vol. 12258, SPIE, 2022, pp. 26–31.

[103] B. Abu-Salih, Domain-specific knowledge graphs: A survey, *Journal of Network and Computer Applications* **185** (2021), 103076.

[104] 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.

[105] X. Yang, H. Zhang and J. Cai, Auto-encoding and distilling scene graphs for image captioning, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[106] 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.

[107] H. Dai, Y. Tian, B. Dai, S. Skiena and L. Song, Syntax-directed variational autoencoder for structured data, *arXiv preprint arXiv:1802.08786* (2018).

[108] M. Allamanis, P. Chanthirasegaran, P. Kohli and C. Sutton, Learning continuous semantic representations of symbolic expressions, in: *International Conference on Machine Learning*, PMLR, 2017, pp. 80–88.

[109] M. Nayyeri, C. Xu, M.M. Alam, J. Lehmann and H.S. Yazdi, LogicENN: a neural based knowledge graphs embedding model with logical rules, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[110] N. Hoernle, R.M. Karampatsis, V. Belle and K. Gal, Multiplexnet: Towards fully satisfied logical constraints in neural networks, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, 2022, pp. 5700–5709.

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

[112] K. Gao, L. Chen, Y. Niu, J. Shao and J. Xiao, Classification-then-grounding: Reformulating video scene graphs as temporal bipartite graphs, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19497–19506.

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

[114] X. Shang, T. Ren, J. Guo, H. Zhang and T.-S. Chua, Video Visual Relation Detection, in: *ACM International Conference on Multimedia*, Mountain View, CA USA, 2017.

[115] G. Bhat, M. Danelljan, L. Van Gool and R. Timofte, Know your surroundings: Exploiting scene information for object tracking, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, Springer, 2020, pp. 205–221.

[116] P. Yadav and E. Curry, VEKG: Video Event Knowledge Graph to Represent Video Streams for Complex Event Pattern Matching, in: *2019 First International Conference on Graph Computing (GC)*, IEEE, 2019, pp. 13–20.

[117] B. Zhao, H. Li, X. Lu and X. Li, Reconstructive sequence-graph network for video summarization, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(5) (2021), 2793–2801.

[118] A. Usmani, M.J. Khan, J. G. Breslin and E. Curry, Towards Multimodal Knowledge Graphs for Data Spaces, in: *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 1494–1499.