Neurosymbolic Artificial Intelligence -1 (2024) 1–24
DOI 10.3233/NAI-240719
IOS Press
CORRECTED PROOF

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

Editor: Md Kamruzzaman Sarker, Bowie State University, USA
Solicited reviews: Ernesto Jimenez-Ruiz, City, University of London, UK; One anonymous reviewer

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 reasoning ability 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 categorises 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 witnessed significant advancements, particularly in scene representation and visual reasoning, with the integration of deep learning, common sense knowledge, and NeuroSymbolic

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

(NeSy) integration [9,18,36,37]. 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 [13]. Despite substantial progress in deep learning and multi-modal methods in computer vision, data-centric techniques often fall short in complex visual reasoning problems that require semantic and relational information [51]. NeSy integration and common sense knowledge infusion have emerged to address this, finding diverse applications such as visual narration [114], self-driving vehicles [101], mathematical logic [84], robotic manipulation [99], and medical diagnostics [31].

Consider a scenario where a Deep Neural Network (DNN) trained for SGG encounters an image of a bustling street scene. Traditionally, it excels in identifying objects like cars, pedestrians, and traffic lights. However, by integrating relational and background information via NeSy approaches, the network goes beyond mere identification. It begins to understand complex interactions, such as a pedestrian waiting to cross the road or a car stopping at a traffic light, by infusing common sense knowledge from knowledge graphs [50]. This NeSy integration imbues the network with an enhanced ability to reason about the scene, recognising not just the objects but also their interrelations and implied actions. For instance, it can be inferred that a person with a shopping bag is likely coming from a store, or a car slowing down near a pedestrian crossing implies yielding. This enriched understanding, combining DNN-based vision with common sense knowledge from knowledge bases, significantly elevates the capabilities of scene graph-based visual reasoning, leading to more accurate, context-aware, and semantically rich scene interpretations.

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 [57]. 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 [70] and saliency [137], and employing advanced techniques like knowledge transfer [32], linguistic supervision [126] and zero-shot learning [61].

Common sense knowledge infusion, particularly, has evolved as a promising strategy to tackle these challenges [51]. Incorporating background details and related facts about scene components enhances the expressiveness of the representation and the performance of downstream reasoning [51]. While statistical and language priors have been widely used in SGG, they offer limited generalisability. Some KGs, such as ConceptNet [100], and WordNet [79], have been utilised 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. A heterogeneous KG, such as the Common Sense Knowledge Graph (CSKG) [43], integrates entities and relationships from multiple sources. Each source contributes different aspects of common sense knowledge, resulting in a comprehensive resource covering a broad spectrum. These heterogeneous KGs offer rich and diverse common sense knowledge related to visual concepts. However, they are underutilised in enhancing scene representations for visual reasoning.

This paper presents a comprehensive review of the promising intersection of deep learning, common sense knowledge and NeSy integration, particularly focusing on semantic scene representation and visual reasoning. Given the growing interest and significant potential in this area, our survey aims to provide a clear and thorough overview of the current literature. We also point out the current challenges, prospects, and applications to guide future research, ensuring that efforts are channelled 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, categorises existing SGG methods, and discusses key challenges and promising future research directions. This survey aims to serve

Table 1
Comparison with existing surveys

Domain	Survey (year)	Key attributes
Neurosymbolic (NeSy) integration	Garcez et al. [24] (2023)	Neurosymbolic AI, machine learning, reasoning, explainable AI, deep learning, trustworthy AI, cognitive reasoning
	Wang et al. [115] (2022)	Neurosymbolic AI, symbolic AI, statistical AI, deep learning
Commonsense knowledge infusion	Ilievski et al. [42] (2021)	Common sense knowledge, semantics, knowledge graphs, reasoning
	Kursuncu et al. [58] (2019)	Knowledge-infused learning, knowledge graph, neural network, neurosymbolic AI
Scene Graph Generation (SGG)	Zhu et al. [142] (2022)	Scene graph generation, visual relationship detection, object detection, scene understanding
	Chang et al. [13] (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

as a valuable resource for researchers and practitioners in the field, guiding future research and contributing to the development of more effective and practical solutions for real-world applications.

1.1. Existing surveys

Garcez et al. [24] and Wang et al. [115] provided comprehensive reviews of NeSy AI, discussing its development, forms of integration, the importance of representation, and promising future research directions. Ilievski et al. [42] analysed multiple sources of common sense knowledge, categorising them into 13 dimensions and suggesting a roadmap for developing a unified resource for NeSy methods. Kursuncu et al. [58] discussed the potential of hybrid NeSy learning approaches that integrate deep learning and knowledge graphs. Meanwhile, Chang et al. [13] provided a comprehensive review of SGG methods, applications, and datasets, while Zhu et al. [142] systematically summarised deep learning-based SGG methods and compared their performance across different datasets and representations [142]. These surveys are summarised 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 specialised survey paper on deep learning and common sense knowledge combined via NeSy integration for SGG and visual reasoning, which our paper addresses.

1.2. Contributions and organisation

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 generalisability, use of heterogeneous common sense knowledge and temporal visual relationships as promising future research directions.

The rest of this paper is organised 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

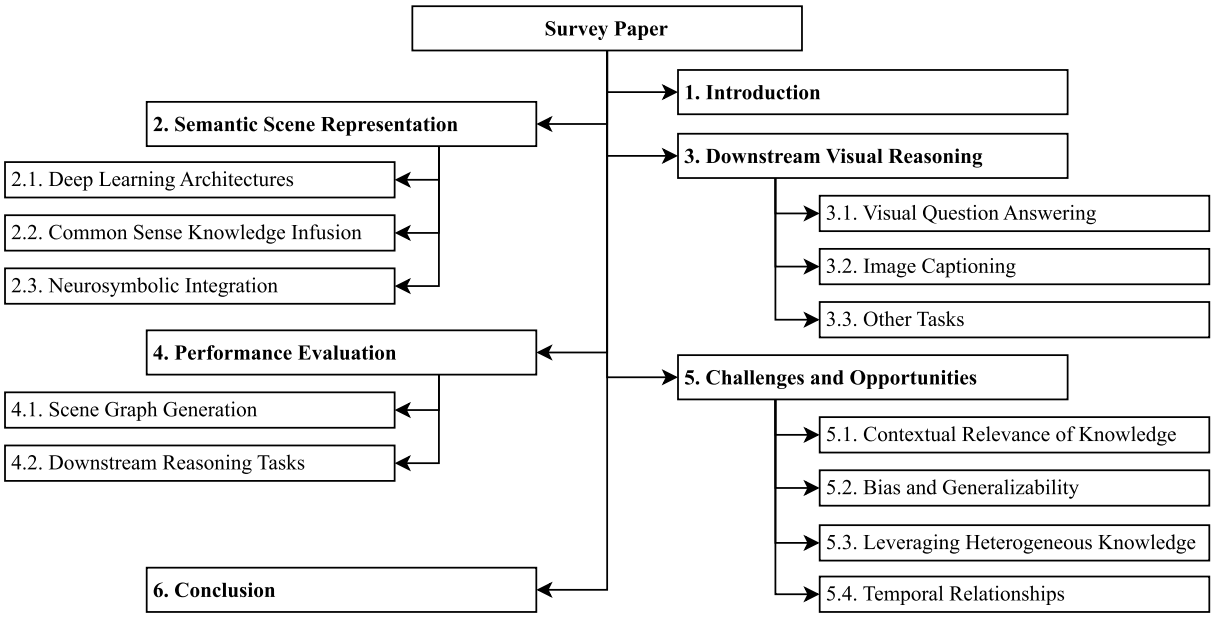

Fig. 1. Structure of the survey paper.

summary of the main challenges and promising future directions in this area of work. Finally, Section 6 summarises and concludes the paper. The structure of this survey is presented in Fig. 1.

2. Semantic scene representation

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 SGG task processes an input image to generate its scene graph, a structured representation that semantically organises objects and their relationships [13]. This process initiates with object detection within the image, then proceeds to classify object attributes. It involves a contextual analysis of these objects, coupled with multimodal feature learning to predict pairwise visual relationships. This process concludes in the creation of a symbolic representation of the scene, as depicted in Fig. 2. The scene graph has emerged as a commonly used semantic scene representation which lays the groundwork for advanced visual reasoning with examples of Visual Question Answering (VQA), image captioning, Multimedia Event Processing (MEP), image retrieval and image generation [51].

Deep learning is integral to the task of SGG. Deep learning architectures, such as Convolutional Neural Networks (CNNs) [60], Recurrent Neural Networks (RNNs) [91], and Graph Neural Networks (GNNs) [93], 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 [15,132,140], language priors [66,73,129], and KGs [28,30,49,50,52,130], 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, [30,50,52,132] the neural and symbolic components operate independently, interacting as

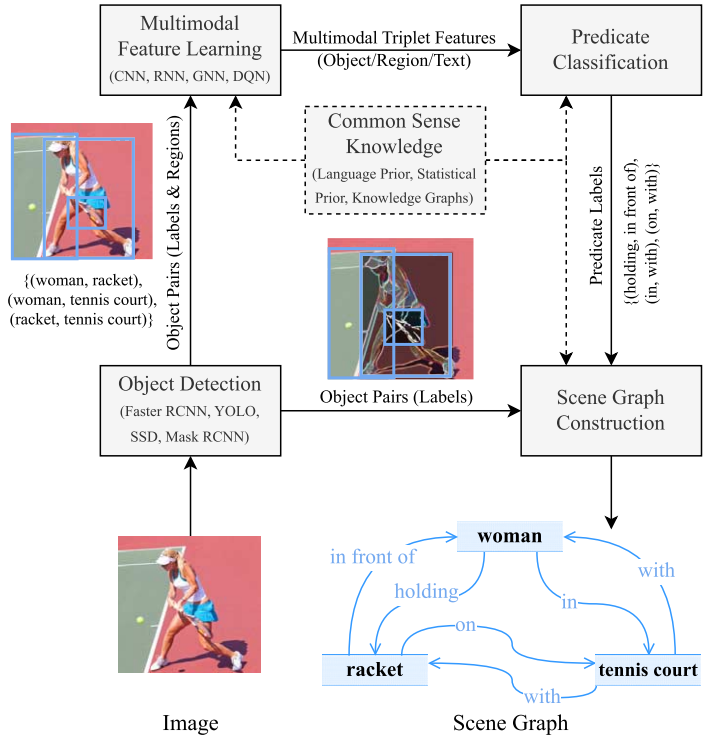

Fig. 2. A schematic representation of the typical scene graph generation process (beginning at the bottom left and concluding at the bottom right) that comprises object detection, multimodal feature learning, common sense knowledge infusion, relationship predicate classification, and scene graph construction.

needed, and focus on distinct yet complementary tasks. Meanwhile, tight coupling [11,15,28,49,66,73,129,130,140] 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 recognise 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.

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.	[15,28,30,49,50,52,66,73,129,130,132,140]
	RNN	Captures contextual information and learns dependencies between objects. Handles sequential data and maintains information over longer sequences	[28,49,50,52,132]
	GNN	Processes graph-structured data and facilitates message passing in SGG. Learns local information and captures the relationships between objects	[15,30,130,140]
	DQN	Formulates SGG as a sequential decision-making process. Handles high-dimensional continuous data and unseen inputs	[66]
	Transformer	Captures long-range inter-object dependencies in SGG. Handles global scene context to improve relation understanding.	[30]
Common sense knowledge source	Statistical prior	Captures structural regularities in visual scenes. Models statistical correlations between object pairs	[15,132,140]
	Language prior	Refines relationship predictions using semantic information. Helps recognise relationships between semantically related objects	[66,73,129]
	Knowledge graph	Provides a structured representation of common sense knowledge. Facilitates inference of unseen visual relationships	[11,28,30,49,50,52,130]
Neurosymbolic integration	Loose coupling	Independent and sequential operation of neural and symbolic components. Flexibility in handling distinct yet complementary tasks	[30,50,52,132]
	Tight coupling	Symbolic knowledge is directly encoded into the neural networks. Unified symbolic reasoning and neural learning capabilities	[11,15,28,49,66,73,129,130,140]

2.1.1. Convolutional neural networks

CNN [60] 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 [15,28,30,49,50,52,66,73,130,132,140] use Faster RCNN [90] with a CNN-based backbone network for detecting objects in images prior to visual relationship detection. Khan et al. [52] 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 [140] incorporated Faster RCNN with a VGG16 backbone in its bounding box module to generate object proposals prior to visual relationship detection. IRT-MSK [30] 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 [49] 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 [130] 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 [28] employs a Region Proposal Network (RPN), a type of CNN, for the 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 [73] 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. Yu et al. [129] utilised CNNs, specifically VGG-16, to extract visual features from the union of bounding boxes of object pairs in images to learn rich visual representations for understanding the relationships between objects, which were then integrated with linguistic knowledge for visual-linguistic feature learning for relationship prediction.

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 [91] and its variants, i.e., Long Short-Term Memory (LSTM) [38] and Gated Recurrent Unit (GRU) [17] 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. [52] used two sets of Bi-directional LSTM (BiLSTM) [95] 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 [49] utilises 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 [132] 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 utilised 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 [28] 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 [93], Graph Convolutional Networks (GCN) [55] and Graph Attention Networks (GAT) [111], 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 [140] used a GAT component in its graphical message passing module for effective contextual learning and recognising 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 [30] 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 [130] 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 [15] used a GNN to propagate messages through a graph built based on statistical object co-occurrence information, learning contextualised 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. Other

Transformer models are also used in SGG because they capture long-range dependencies and contextual information. Compared to CNNs, transformers provide a more global understanding of the scene by considering the relationships between all elements within an image. This helps SGG models understand the relational context between multiple objects to generate scene graphs. Guo et al. [30] proposed an SGG model that constructs an instance relation transformer, which applies the transformer structure to visual features, label embeddings, and position embeddings to encode contextual information and relational contexts within images. Transformers have been more extensively used in VQA [48,88,97,103] and image captioning [50,88] works, which are discussed in Section 3.

Deep Q-Networks (DQNs), also contribute to the rich variety of deep learning techniques applied in knowledge-based SGG, further expanding the possibilities for scene understanding. DeepVRL [66] approaches the task of identifying visual relationships and attributes as a sequential process, managed using a 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 systematically selects the optimal actions for identifying objects, relationships, and attributes within the visual context, directly linking the decision-making process with the visual elements in the image. Additionally, the framework incorporates a replay memory to retain information from previous episodes, which aids in stabilising the training by averaging the training distribution over past experiences and minimising the correlation among training examples, thereby enhancing its capacity to interpret and analyse visual information.

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 [42], 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 generally found in trees rather than fish, or that people generally sit on chairs rather 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 [51].

2.2.1. Statistical and language 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 generally occur based on past observations. For example, DSGAT [140] 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 [132] 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 [15] leverages statistical correlations between pairwise objects and visual relationships to regularise the semantic space and minimise 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. Causality is also used as statistical common sense knowledge to enhance visual reasoning. The CMCIR framework [71] employs causal interventions to address spurious correlations in VQA, integrating visual-linguistic reasoning, spatial-temporal transformers, and feature fusion to discern fine-grained interactions between modalities and provide deeper insights into complex events. The CIIC framework [72] targets visual and linguistic confounders in encoder–decoder models to disentangle visual features and rectify linguistic biases through structural causal modelling, enhancing the accuracy and reliability of image captions.

Language priors utilise the semantic information encapsulated in words to enhance the prediction of relationships. They aid in recognising 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 [73] 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 utilising language priors as a source of common sense knowledge. DeepVRL [66] approaches the task of identifying visual relationships and attributes as a sequential process, utilising 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. An ambiguity-aware object mining strategy is implemented to address semantic ambiguity among object categories. Yu et al. [129] proposed the extraction of symbolic knowledge from external textual data, such as Wikipedia, by parsing large-scale text data to identify and encode relationships between objects to inform the relationship prediction process. This symbolic knowledge is then incorporated into a teacher network, which distils it into an end-to-end student network for predicting relationships.

2.2.2. Knowledge graphs

KGs are structured databases that formally represent real-world entities and their interrelations, effectively encoding the structure of the world and its complex relationships. As outlined in [45], KGs are pivotal in providing structured and interpretable information. In the context of SGG, KGs are instrumental as they supply essential common sense knowledge, aiding in the creation of more accurate and comprehensive scene graphs. The application of KGs in SGG is detailed in Table 3. As an example, a KG could contain explicit relationships such as "birds are often found in trees" or "cars are typically on roads," thereby enabling the SGG system to infer these relationships in images, even if such specific instances are absent from the training dataset. This integration of KGs into SGG systems enriches them with a level of contextual understanding that significantly improves their performance in scene interpretation.

Khan et al. [52] 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 [30], 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 [49] 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,

Table 3
Utilisation of knowledge graphs for infusing common sense knowledge in SGG

Knowledge graph	Nature of knowledge	Dimensions	Examples	Methods employed
ConceptNet [100]	Information about common objects, activities, relations, etc., in text format	8M nodes, 36 relations & 21M edges	(book, used for, reading), (pen, capable of, writing)	[11,28,30,49,130]
Visual genome [57]	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)	[30,130]
Wordnet [79]	Lexical information about words, concepts, relations, etc.	0.155M words, 10 relations & 0.176M synsets	(dog, has part, tail), (reading, part meronym, scanning)	[130]
CSKG [43]	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)	[50,52]

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 initialising the path-based knowledge with the neighbour-based knowledge.

GB-Net [130] 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 [28] 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. Buffelli et al. [11] encoded external knowledge from ConceptNet and training facts and injected it into the regularisation process of training SGG models. Herron et al. [34] demonstrated the application of Web Ontology Language (OWL) in visual reasoning by enabling the representation of source and ranges of predicates, hierarchies and inverse relationships within datasets, which can aid in more sophisticated reasoning with knowledge graphs and facilitate deeper infusion of common sense knowledge in SGG.

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 [59]. However, these techniques are data-intensive and operate as black boxes, making their decision-making processes difficult to interpret [92]. 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 [23]. However, they offer less trainability and can be brittle when faced with out-of-domain data [75]. In scene understanding and visual reasoning, NeSy integration can enable systems that not only recognise complex visual scenes but also provide logical, traceable explanations for their interpretations, aligning with the principles of explainable AI [29]. The ability to extract and utilise compact, interpretable knowledge representations from neural models enhances local explanations, offering clarity on decision-making processes [80]. NeSy systems thus stand at the forefront of advancing AI towards being more semantically sound, explainable, and ultimately more trustworthy, particularly in sophisticated tasks involving vision and language [37]. A fine-grained classification of NeSy approaches with six different types is provided in [24,115]. 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 categorise NeSy approaches into two types, loosely coupled and tightly coupled [24], based on the degree of integration between the symbolic and neural components.

2.3.1. Loose coupling

Loosely coupled NeSy approaches feature a relative independence between their symbolic and neural components. Each adheres to its own processes and methodologies, interacting as required without deep intertwinement. This allows each component to leverage its strengths while benefiting from the capabilities of the other. In some loosely coupled NeSy approaches, neural and symbolic elements focus on distinct, complementary tasks within a larger pipeline. They collaborate to achieve the overall task, maintaining the ability to function independently. This arrangement combines the advantages of both components while preserving their autonomy. It is particularly effective in complex tasks requiring a blend of symbolic reasoning and neural learning. An example is the NeSy Concept Learner (NS-CL) by Mao et al. [74], comprising a neural network for learning visual concepts and a symbolic module for processing symbolic programs. The symbolic module provides feedback signals aiding the neural module's gradient-based optimisation.

In several loosely coupled NeSy approaches, the components function sequentially. The neural component first transforms raw input into a format processable by the symbolic component, which then processes this input and sends its output back to the neural component for further processing. This sequential operation allows the fusion of symbolic reasoning and neural processing while maintaining operational independence. For instance, IRT-MSK [30] extracts knowledge-embedded semantic features from KGs to explore visual features in SGG, with symbolic and neural components operating independently. MotifNet [132] uses a global context (symbolic) to sequentially inform its predictions (neural). Here, the global context, represented via LSTMs, infuses symbolic knowledge into the neural component, enhancing prediction capabilities. Khan et al. [50] proposed a loosely coupled NeSy framework for SGG, knowledge enrichment, and downstream visual reasoning. This approach employs symbolic approaches for scene representation and common sense knowledge enrichment. Neural modules predict semantic elements in images and process enriched scene graphs for downstream tasks. In this setup, the interdependence of modules is evident: the accuracy of scene graph elements predicted by the neural module directly impacts the enrichment process, which in turn directly influences the performance of downstream reasoning tasks.

2.3.2. Tight coupling

The symbolic and neural components in tightly coupled NeSy approaches are deeply integrated, leveraging the strengths of both domains in a more unified way for enhanced performance and capabilities. This integration often involves incorporating symbolic rules or knowledge into the architecture or training of neural networks. The structure of these networks or their training methodologies is influenced by symbolic rules, resulting in a deep fusion of symbolic and neural elements. Such integration allows neural networks to utilise the reasoning capabilities of symbolic rules while benefiting from the learning capabilities of neural components. Some recent works like Gu et al. [28] and Marino et al. [76] employ GNNs to embed entities and relations from external knowledge bases to enhance performance in scene understanding and visual reasoning tasks. Several VQA methods [3,40,47,98,109] generate and execute symbolic programs, implemented as neural networks or fully differentiable operations, to answer questions.

Symbolic knowledge in tightly coupled NeSy systems is often encoded into the distributed representations of neural networks. This configuration allows neural components to leverage reasoning capabilities inherent in symbolic knowledge alongside their learning capabilities. This approach is beneficial in tasks requiring a deep understanding of the data, as it exploits symbolic knowledge and neural learning. For instance, Li et al. [64] developed a hierarchical semantic segmentation network using compositional relations across semantic hierarchies as additional training targets. Similarly, Zhou et al. [140] designed a three-module system for SGG, infusing statistical probabilities into the modules for a tightly coupled NeSy approach. COACHER [49] integrated common sense knowledge for zero-shot relation prediction in SGG, embedding symbolic knowledge into the distributed representation of the neural component. GB-Net [130] uses a graph-based neural network to refine connections between a scene graph and a common sense graph, exploiting the heterogeneous structure of the interconnected graphs. KERN [15] incorporates statistical correlations between pairwise objects and visual relationships in DNNs, using a structured KG to propagate messages and explore object interactions. KBGAN [28] integrates symbolic knowledge from an external knowledge base into neural components for SGG, refining object and phrase features.

The VRD model [73] integrates symbolic knowledge into CNNs to detect visual relationships from images. DeepVRL [66] employs a directed semantic action graph to capture semantic relationships. It utilises 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. Buffelli et al. [11] introduce a regularisation technique for SGG models that embeds symbolic background knowledge encoded in first-order logic into the learning process of neural SGG models without increasing the computational overhead. Yu et al. [129] use a neural model, specifically CNN, to extract detailed visual features from images, particularly focusing on the relationships between object pairs. This neural processing is then synergistically combined with symbolic knowledge extracted from extensive text data sources like Wikipedia. The integration occurs within another neural model, where the visual features and the distilled symbolic knowledge inform each other to predict visual relationships for the generation of symbolic scene graphs. Herron et al. [35] highlight the importance of combining explicit symbolic knowledge representation and reasoning machinery of OWL-based KGs with deep learning in NeSy visual reasoning. OWL-based KGs support knowledge embedding into vectors, which can also guide neural learning via KG completion, support knowledge

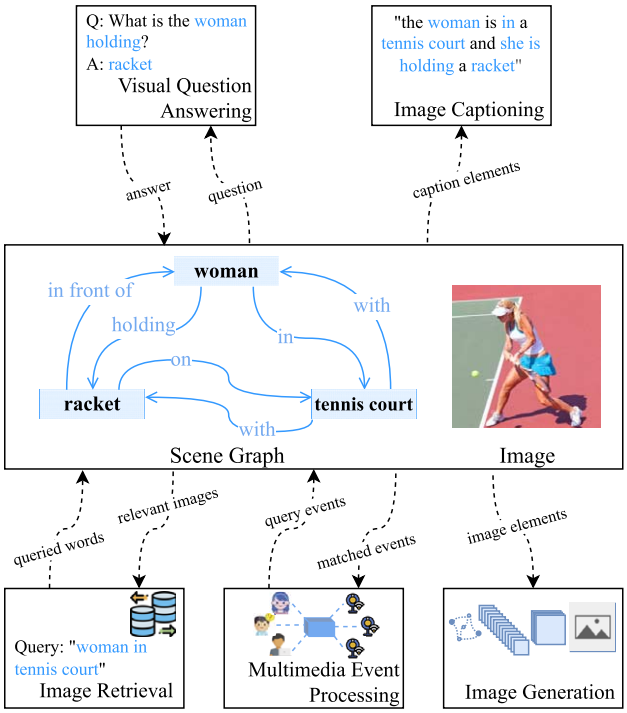

Fig. 3. An overview of the downstream visual reasoning tasks of scene graph generation, including VQA, image captioning, MEP, image retrieval, and image generation.

infusion into DNNs for improved learning, and enable symbolic reasoning and formalised semantics in scene graph-based visual reasoning. Hyperbolic methods enhance visual reasoning in a tightly-coupled NeSy approach by utilising the Poincaré ball model to encode hierarchical structures with minimal distortion [78]. Atigh et al. [6] leverage hyperbolic space for pixel-level image segmentation, reformulating multinomial logistic regression to optimise segmentation in a lower-dimensional, hierarchically structured embedding space. In structured multi-label prediction, Xiong et al. [120] employ hyperbolic Poincaré hyperplanes as linear decision boundaries, encoding logical relationships like implication and exclusion. This geometric approach ensures logical consistency in tasks like image annotation and text categorisation.

3. Downstream visual reasoning

Scene graphs are widely utilised in downstream visual reasoning tasks, including VQA, image captioning, MEP, image retrieval, and image generation, as shown in Fig. 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

VQA models aim to generate accurate responses to queries about visual scenes by harnessing multimodal features, semantic relationships within scene graphs, and factual knowledge. For example, Zhang et al. [133] enhanced Graph Neural Networks (GNNs) by integrating structural data from scene graphs for VQA tasks. Similarly, Ziaeefard et al. [144] proposed a VQA approach using Graph Attention Networks (GATs) that synergises scene graph encoding with knowledge sourced from ConceptNet. Wu et al. [119] combined image content with a common knowledge base for image-based question answering, though their method does not explicitly focus on reasoning. On the other

hand, Narasimhan et al. [82] created a learning-based retrieval method, embedding questions, images, and facts together, linking visual concepts with related facts. ConceptBert, introduced by Garderes et al. [25], merges pre-trained image and language features with knowledge graph embeddings, bypassing the need for external knowledge annotations. Shevchenko et al. [97] employed a transformer that integrates knowledge base information for aligning visual and language data. Zhu et al. [143] crafted Mucko, a model adept at multilayer cross-modal knowledge reasoning, forming a multimodal heterogeneous graph for evidence capture. Yu et al. [127] devised a model using a multimodal knowledge graph and a memory-based recurrent network for cross-modal reasoning, adept at handling knowledge across various modalities. Anderson et al. [5] utilised Faster R-CNN for proposing image regions and incorporated attention mechanisms for improved VQA interpretability. The LXMERT framework by Tan et al. [103] employs a large-scale Transformer model with triple encoders, focusing on understanding both visual concepts and language semantics. The Meta Module Network (MMN) [16] brings scalability and generalisability to VQA, using a metamorphic meta module that adapts to diverse instances without increasing model complexity. MDETR [48] is an end-to-end modulated detector, leveraging a transformer-based architecture for the early-stage fusion of image and text modalities, efficiently extracting visual concepts from the text in multi-modal reasoning systems. Lastly, Zhang et al. [136] applied their object detection model to visual reasoning in VQA, achieving richer visual representations of objects and concepts.

Over the past few years, several scene graph-based VQA methods have emerged. Hudson et al. [40] introduced a visual reasoning method using Neural State Machines (NSM), which blends visual and linguistic inputs into semantic concepts through a probabilistic scene graph, enabling sequential reasoning and inference. Zhang et al. [133] took a different route by embedding the structural features of scene graphs into a GNN for downstream VQA. Building on this, Yang et al. [125] developed the Scene Graph Convolutional Network (SceneGCN), integrating object properties and semantic relationships for a structured scene representation that boosts VQA through visual context and language priors. Exploring further, Graphhopper [56] tackles the complexity of multi-hop reasoning over complex visual scenes to deduce reasoning paths leading to answers in VQA. The Dual Message-passing enhanced GNN (DM-GNN) [63] encodes multi-scale scene graph information into two distinct graphs focusing on objects and relations. This dual structure achieves a balanced representation of object, relation, and attribute features in VQA. Finally, the Scene Graph Refinement network (SGR) [88] presents a transformer-based network to refine object and relation feature learning in VQA. It leverages question semantics to learn multimodal representations and selects the most relevant relations for improved question answering.

3.2. Image captioning

Scene graphs have become a pivotal tool in image captioning, enhancing scene descriptions beyond the conventional vision-language features. The Abstract Scene Graph (ASG) approach by Chen et al. [14] notably integrates user intentions and semantic information into scene graphs, paving the way for diverse and accurate text descriptions of scenes. Similarly, the SMP method by Zhang et al. [137] leverages the saliency of visual relationships in scene graphs to enrich caption generation. Goel et al. [26] proposed integrating prior knowledge in image captioning models through conditional latent topic attention and used the semantic and syntactic structure of captions for regularisation. This approach has not only produced more human-like captions but also marked significant advancements on the COCO dataset, especially in scenarios with limited data availability. The R-SCAN model by Lee et al. [62] stands out for its focus on learning visual relationship features and proposing pre-training SGG models with relevant visual relationship data. Liao et al. [67] ventured into the realm of 3D Scene Graph-based Change Captioning (SGCC), aiming to boost object location accuracy in change captioning tasks. Wu et al. [118] introduced a method that integrates high-level visual concepts and external knowledge into a deep learning cascade of CNN and RNN, resulting in significant improvements in image captioning and VQA performance. Yu et al. [128] took a step further with the 3D-SceneCaptioner, a point clouds-based image captioning technique that leverages the rich semantic information in point clouds for generating more precise captions. Zhang et al. [138] enhanced the performance of transformers in image captioning by incorporating a knowledge graph and augmenting maximum likelihood estimation with a Kullback-Leibler divergence term. Some recent works [50,141] highlight the superior performance of image captioning methods that utilise information from knowledge graphs over those relying solely

on image data. The versatility of these techniques extends to other tasks that suffer from data scarcity, showcasing their potential to enhance generalisation and expressiveness.

3.3. Other tasks

NeSy visual semantic models have found useful applications in the representation of multimedia streams for real-time multimodal event processing in the Internet of Multimedia Things (IoMT) [19,54]. These models blend DNNs for object and attribute detection with symbolic rules to understand spatiotemporal relations among objects. This integration is pivotal for correlating high-level events in response to user queries. In image retrieval, scene graphs transform the way we articulate image semantics and structure, enabling more efficient searches in vast databases based on image content. Schroeder et al. [94] developed Structured Query-based Image Retrieval (SQIR), which employs scene graphs as directed subgraphs to streamline graph matching for image retrieval through structured queries and scene graph embeddings. Meanwhile, Ward et al. [117] have taken a NeSy approach, combining deep learning with knowledge graphs, to ingeniously guide the colourisation of black-and-white images. Compared to the conventional colourisation methods, this approach not only classifies objects but also taps into contextual knowledge for accurately coloring both simple and complex scenes. Unlike textual scene descriptions, scene graphs have emerged as a more dynamic and scalable solution for image generation, excelling as the complexity of objects and relationships grows [46]. Further enhancing their utility, scene graphs infused with common sense knowledge have paved the way for more lifelike image generation, as evidenced by Khan et al. [52]. Lastly, Gu et al. [28] harnessed ConceptNet within an attention-based RNN framework to refine objects and phrases with common sense knowledge for reconstructing 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 summarised in Table 4 and Table 5, respectively. Table 4 details each SGG technique, including the DNN architecture, knowledge source, and the type of NeSy integration, as discussed in Section 2. Additionally, it summarises their performance metrics as reported in existing literature. The benchmark dataset frequently employed for SGG evaluation is Visual Genome [57]. 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 [73]. 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 [15,107].
- $zR@K$ is similar to $R@K$, but it is only computed for relationships that do not appear in the training dataset [73,106].

4.2. Downstream reasoning tasks

4.2.1. Visual question answering

The commonly used datasets for scene graph-based VQA include GQA [41], MS COCO [69], and Visual Genome [57]. The GQA dataset [41] 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

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 [52]	CNN and LSTM	CSKG	Loose coupling	**35.5/39.1**	**10.9/12.6**	-/-
DSGAT [140]	CNN and GAT	Statistical prior	Tight coupling	28.8/32.9	8.9/11.8	-/-
IRT-MSK [30]	CNN and GCN	ConceptNet and Visual Genome	Loose coupling	27.8/31.0	-/-	-/-
COACHER [49]	CNN and LSTM	ConceptNet	Tight coupling	-/-	-/-	**19.3/22.2**
MotifNet [132]	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 [130]	CNN and GNN	ConceptNet, WordNet and Visual Genome	Tight coupling	26.4/30.0	6.1/7.3	-/-
KERN [15]	CNN and GNN	Statistical prior	Tight coupling	27.1/29.8	6.4/7.3	-/-
KB-GAN [28]	CNN and GRU	ConceptNet	Tight coupling	13.6/17.6	-/-	18.1/21.1
DeepVRL [66]	CNN and DQN	Language prior	Tight coupling	13.3/12.6	-/-	6.3/7.1
VRD [73]	CNN	Language prior	Tight coupling	0.3/0.5	-/-	-/-

*on GQA dataset [41]

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 [57]	108K images	33.8K	42K	✓	✓	✗
VG150 [121]	88K images	150	50	✓	✓	✗
VG200 [134]	99K images	200	100	✓	✓	✗
VG80k [135]	100K images	53K	29K	✓	✓	✗
VG-MSDN [65]	95K images	150	50	✓	✓	✗
MS COCO [69]	330K images	80	–	✓	✓	✗
Flickr30K [87]	30K images	–	–	✓	✗	✗
GQA [41]	113K images	1.7K	310	✗	✓	✗
VQA-v2 [27]	204K images	–	–	✗	✓	✗
VCR [131]	110K images	–	–	✗	✓	✗
KB-VQA [113]	700 images	–	–	✗	✓	✓
FVQA dataset [112]	2190 images	–	–	✗	✓	✓
OK-VQA [77]	14K images	–	–	✗	✓	✓
KRVQA [12]	33K images	–	–	✗	✓	✓
VRD [73]	5K images	100	70	✗	✗	✗
NeSy4VRD [34]	5K images	109	71	✗	✗	✗

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, 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 [41] 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 [113] 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 [112] pairs questions and answers with supporting facts in a structured triplet format, using a knowledge base built from DBpedia [7], WebChild [104,105] and ConceptNet [100]. It includes 2,190 images, 5,286 questions, and 193,449 facts, with questions categorised by visual concept type, answer source, and supporting knowledge base. The OK-VQA dataset [77], 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 categorised into one of 10 knowledge categories, or "Other" if they don't fit into any specific category. The KRVQA dataset [12], 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 categorised by reasoning steps and knowledge involvement, and the dataset is built on the scene graph annotations of the Visual Genome dataset [57] and the knowledge base of the FVQA dataset [112]. Although these datasets contain external knowledge to some extent, KB-VQA [113] and FVQA [112] 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 [69], Flickr30k [87], and Visual Genome [57] datasets. Various metrics are used to assess the quality of the generated image captions, each focusing on different aspects.

- The BLEU score [85], 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 [8], another metric from the machine translation field, emphasises 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 [68], initially designed for text summarisation, 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 [110], 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 [4], 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 [51]. 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 either due to their inherent contextual limitations [21] or lack of formal semantics [35]. 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 [51]. 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 directly assess 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 [33,81] 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 [102], adaptive thresholds [89], and domain-specific knowledge [1]. 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. KGs with formal semantics [35] have the potential to enhance contextual reasoning using relationships and hierarchies between objects and predicates. Furthermore, integrating domain-specific knowledge can address specialised 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 generalisability

A main cause of the limited performance of existing SGG methods is the long-tailed distribution of crowd-sourced datasets [13], restricting the SGG methods from generalising 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. [53] Moreover, visual feature representations of relationships can significantly differ across various scenes, adding another layer of complexity [142]. 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 generalisability in addition to solving the long-tailed distribution problem.

Approaches such as zero-shot [49,116] and few-shot learning [30] have been investigated to address these challenges in SGG. Zero-shot learning leverages previously learned relationships to recognise visual relationships that have not been seen before. Conversely, few-shot learning utilises 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. The NeSy4VRD dataset [34] extended the original VRD dataset with more meaningful and non-ambiguous predicates and highlighted the potential of formal semantics in addressing

these challenges. Additionally, knowledge transfer and distillation techniques [86,124] present another promising direction. Previously learned visual relationships can be leveraged by employing models trained on diverse common sense knowledge bases, enhancing the generalisation 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 generalisability, 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 [79] and ConceptNet [100], 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 [43], 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 [52]. However, their application in tightly-coupled SGG approaches and mainstream visual reasoning tasks, especially VQA, remains unexplored. These heterogeneous sources are essential but underutilised 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 DNNs for SGG can be an effective approach [2,20]. This strategy can empower DNNs 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 [28,130], further exploration is needed to understand how the utilisation 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 DNNs [39,83] for scene understanding and visual reasoning. The rich class and property hierarchies in the VRD-World ontology [34] offer opportunities for meaningful OWL reasoning capabilities, which can help understand relationships and hierarchies between objects and predicates, enhancing the depth and accuracy of visual reasoning. OWL-based KGs can also guide neural learning and provide structured semantics in the context of scene graph-based NeSy visual reasoning [35]. 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 [22,123] and corresponding datasets [44,96], there is an opportunity to integrate external common sense knowledge into these methods.

Addressing these gaps requires a multi-faceted approach. Firstly, object tracking [10] can be integrated to maintain continuity in recognising 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 [114] 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 [122,139] 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 [108]. 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.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6223 and 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] B. Abu-Salih, Domain-specific knowledge graphs: A survey, *Journal of Network and Computer Applications* **185** (2021), 103076. doi:10.1016/j.jnca.2021.103076.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] M.G. Atigh, J. Schoep, E. Acar, N. Van Noord and P. Mettes, Hyperbolic image segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4453–4462.

[7] 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. doi:10.1007/978-3-540-76298-0_52.

[8] 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.

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

[10] 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.

[11] D. Buffelli and E. Tsamoura, Scalable theory-driven regularization of scene graph generation models, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, 2023, pp. 6850–6859.

[12] 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. doi:10.1109/TNNLS.2020.3045034.

[13] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen and A. Hauptmann, Scene graphs: A survey of generations and applications, 2021, arXiv preprint arXiv:2104.01111.
[14] 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.
[15] 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.
[16] 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.
[17] K. Cho, B. Van Merriënboer, D. Bahdanau and Y. Bengio, On the properties of neural machine translation: Encoder–decoder approaches, 2014, arXiv preprint arXiv:1409.1259.
[18] W.W. Cohen, H. Sun, R.A. Hofer and M. Siegler, Scalable neural methods for reasoning with a symbolic knowledge base, 2020, arXiv preprint arXiv:2002.06115.
[19] 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).
[20] H. Dai, Y. Tian, B. Dai, S. Skiena and L. Song, Syntax-directed variational autoencoder for structured data, 2018, arXiv preprint arXiv:1802.08786.
[21] 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.
[22] 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.
[23] A.D. Garcez, M. Gori, L.C. Lamb, L. Serafini, M. Spranger and S.N. Tran, Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning, 2019, arXiv preprint arXiv:1905.06088.
[24] A.D. Garcez and L.C. Lamb, Neurosymbolic AI: The 3rd wave, *Artificial Intelligence Review* (2023), 1–20.
[25] 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. doi:10.18653/v1/2020.findings-emnlp.44.
[26] 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.
[27] 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.
[28] 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.
[29] D. Gunning, Explainable artificial intelligence (xai), *Defense advanced research projects agency (DARPA) and Web* **2**(2) (2017), 1.
[30] 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. doi:10.1145/3394171.3414025.
[31] M. Hassan, H. Guan, A. Melliou, Y. Wang, Q. Sun, S. Zeng, W. Liang, Y. Zhang, Z. Zhang, Q. Hu et al., 2022, Neuro-symbolic learning: Principles and applications in ophthalmology, arXiv preprint arXiv:2208.00374.
[32] 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, 2020, arXiv preprint arXiv:2006.07585.
[33] N. Heist, Towards knowledge graph construction from entity co-occurrence, EKAW (Doctoral consortium), 2018.
[34] D. Herron, E. Jiménez-Ruiz, G. Tarroni and T. Weyde, NeSy4VRD: A multifaceted resource for neurosymbolic AI research using knowledge graphs in visual relationship detection, 2023, arXiv preprint arXiv:2305.13258.
[35] D. Herron, E. Jiménez-Ruiz and T. Weyde, On the benefits of OWL-based knowledge graphs for neural-symbolic systems, in: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning*, Vol. 3432, CEUR Workshop Proceedings, 2023, pp. 327–335.
[36] P. Hitzler, F. Bianchi, M. Ebrahimi and M.K. Sarker, Neural-symbolic integration and the semantic web, *Semantic Web* **11**(1) (2020), 3–11. doi:10.3233/SW-190368.
[37] 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. doi:10.1093/nsr/nwac035.
[38] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural computation* **9**(8) (1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735.
[39] 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.
[40] D. Hudson and C.D. Manning, Learning by abstraction: The neural state machine, *Advances in Neural Information Processing Systems* **32** (2019).
[41] 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.
[42] F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D.L. McGuinness and P. Szekely, Dimensions of commonsense knowledge, 2021, arXiv preprint arXiv:2101.04640.
[43] F. Ilievski, P. Szekely and B. Zhang, Cskg: The commonsense knowledge graph, in: *European Semantic Web Conference*, Springer, 2021, pp. 680–696.

[44] 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.
[45] S. Ji, S. Pan, E. Cambria, P. Marttinen and S.Y. Philip, A survey on knowledge graphs: Representation, acquisition, and applications, *IEEE Transactions on Neural networks and Learning Systems* **33**(2) (2021), 494–514. doi:10.1109/TNNLS.2021.3070843.
[46] 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.
[47] J. Johnson, B. Hariharan, L. Van Der Maaten, J. Hoffman, L. Fei-Fei, C.L. 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.
[48] 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.
[49] 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.
[50] M.J. Khan, J. Breslin and E. Curry, NeuSyRE: Neuro-symbolic visual understanding and reasoning framework based on scene graph enrichment, *Semantic Web* (2023).
[51] 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. doi:10.1109/MIC.2022.3176500.
[52] 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. doi:10.1007/978-3-031-06981-9_6.
[53] M.J. Khan, J.G. Breslin and E. Curry, Towards fairness in multimodal scene graph generation: Mitigating biases in datasets, knowledge sources and models, in: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM'23) Workshops*, 2023.
[54] 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.
[55] T.N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.
[56] 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.
[57] 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. doi:10.1007/s11263-016-0981-7.
[58] U. Kursuncu, M. Gaur and A. Sheth, Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning, 2019, arXiv preprint arXiv:1912.00512.
[59] Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *nature* **521**(7553) (2015), 436–444. doi:10.1038/nature14539.
[60] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE* **86**(11) (1998), 2278–2324. doi:10.1109/5.726791.
[61] 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.
[62] 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, 2019, arXiv preprint arXiv:1909.09953.
[63] 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. 1–6.
[64] 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.
[65] 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.
[66] 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.
[67] 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. doi:10.1145/3474085.3475712.
[68] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text Summarization Branches Out*, 2004, pp. 74–81.
[69] 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.
[70] 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.
[71] B. Liu, D. Wang, X. Yang, Y. Zhou, R. Yao, Z. Shao and J. Zhao, Show, deconfound and tell: Image captioning with causal inference, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18041–18050.
[72] Y. Liu, G. Li and L. Lin, Cross-modal causal relational reasoning for event-level visual question answering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
[73] 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.
[74] 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, 2019, arXiv preprint arXiv:1904.12584.

[75] G. Marcus, Deep learning: A critical appraisal, 2018, arXiv preprint arXiv:1801.00631.
[76] 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.
[77] 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.
[78] P. Mettes, M.G. Atigh, M. Keller-Ressel, J. Gu and S. Yeung, 2023, Hyperbolic deep learning in computer vision: A survey, arXiv preprint arXiv:2305.06611.
[79] G.A. Miller, WordNet: A lexical database for English, *Communications of the ACM* **38**(11) (1995), 39–41. doi:10.1145/219717.219748.
[80] G. Montavon, W. Samek and K.-R. Müller, Methods for interpreting and understanding deep neural networks, *Digital signal processing* **73** (2018), 1–15. doi:10.1016/j.dsp.2017.10.011.
[81] 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. doi:10.18653/v1/P19-1081.
[82] 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.
[83] 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).
[84] 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.
[85] 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.
[86] 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.
[87] 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.
[88] 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).
[89] 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.
[90] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6) (2016), 1137–1149. doi:10.1109/TPAMI.2016.2577031.
[91] D.E. Rumelhart, G.E. Hinton and R.J. Williams, Learning representations by back-propagating errors, *nature* **323**(6088) (1986), 533–536. doi:10.1038/323533a0.
[92] W. Samek, T. Wiegand and K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017, arXiv preprint arXiv:1708.08296.
[93] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner and G. Monfardini, The graph neural network model, *IEEE Transactions on Neural Networks* **20**(1) (2008), 61–80. doi:10.1109/TNN.2008.2005605.
[94] 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.
[95] M. Schuster and K.K. Paliwal, Bidirectional recurrent neural networks, *IEEE Transactions on Signal Processing* **45**(11) (1997), 2673–2681. doi:10.1109/78.650093.
[96] 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.
[97] V. Shevchenko, D. Teney, A. Dick and A.V.D. Hengel, Reasoning over vision and language: Exploring the benefits of supplemental knowledge, 2021, arXiv preprint arXiv:2101.06013.
[98] 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.
[99] T. Silver, A. Athalye, J.B. Tenenbaum, T. Lozano-Perez and L.P. Kaelbling, Learning neuro-symbolic skills for bilevel planning, 2022, arXiv preprint arXiv:2206.10680.
[100] 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.
[101] 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.
[102] 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. doi:10.1145/3522586.
[103] H. Tan and M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, 2019, arXiv preprint arXiv:1908.07490.

[104] 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. doi:10.1145/2556195.2556245.
[105] 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.
[106] 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.
[107] 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.
[108] 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. doi:10.1145/3543873.3587665.
[109] 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.
[110] 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.
[111] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio and Y. Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.
[112] 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. doi:10.1109/TPAMI.2017.2754246.
[113] P. Wang, Q. Wu, C. Shen, A.V.D. Hengel and A. Dick, Explicit knowledge-based reasoning for visual question answering, 2015, arXiv preprint arXiv:1511.02570.
[114] 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.
[115] W. Wang and Y. Yang, 2022, Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing, arXiv preprint arXiv:2210.15889.
[116] 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.
[117] 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.
[118] 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. doi:10.1109/TPAMI.2017.2708709.
[119] 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.
[120] B. Xiong, M. Cochez, M. Nayyeri and S. Staab, Hyperbolic embedding inference for structured multi-label prediction, *Advances in Neural Information Processing Systems* **35** (2022), 33016–33028.
[121] 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.
[122] 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. doi:10.1109/GC46384.2019.00011.
[123] 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.
[124] 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).
[125] 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. doi:10.1109/ICIP40778.2020.9190771.
[126] 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.
[127] 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. doi:10.1016/j.patcog.2020.107563.
[128] 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, Proceedings, Part II, Springer, 2021, pp. 275–286.
[129] R. Yu, A. Li, V.I. Morariu and L.S. Davis, Visual relationship detection with internal and external linguistic knowledge distillation, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1974–1982.
[130] 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.
[131] 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.
[132] 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.

[133] C. Zhang, W.-L. Chao and D. Xuan, An empirical study on leveraging scene graphs for visual question answering, 2019, arXiv preprint arXiv:1907.12133.
[134] 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.
[135] 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.
[136] 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.
[137] 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).
[138] Y. Zhang, X. Shi, S. Mi and X. Yang, Image captioning with transformer and knowledge graph, *Pattern Recognition Letters* **143** (2021), 43–49. doi:10.1016/j.patrec.2020.12.020.
[139] 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.
[140] 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. doi:10.1016/j.patcog.2021.108367.
[141] 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.
[142] 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, 2022, arXiv preprint arXiv:2201.00443.
[143] 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, 2020, arXiv preprint arXiv:2006.09073.
[144] 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. doi:10.18653/v1/2020.coling-main.169.