

NeuSyRE: Neuro-symbolic visual understanding and reasoning framework based on scene graph enrichment

M. Jaleed Khan ^{a,*}, John G. Breslin ^{a,b} and Edward Curry ^{a,b}

^a *SFI Centre for Research Training in Artificial Intelligence, Data Science Institute, University of Galway, Ireland*

^b *Insight SFI Research Centre for Data Analytics, Data Science Institute, University of Galway, Ireland*

E-mails: m.khan12@universityofgalway.ie, john.breslin@universityofgalway.ie, edward.curry@universityofgalway.ie

Editors: Monireh Ebrahimi, IBM, USA; Pascal Hitzler, Kansas State University, USA; Md Kamruzzaman Sarker, Bowie State University, USA; Daria Stepanova, Bosch Center for AI, Germany

Solicited reviews: Ivan Donadello, Free University of Bozen-Bolzano, Italy; Luciano Serafini, Fondazione Bruno Kessler, Italy; Two anonymous reviewers

Abstract. Exploring the potential of neuro-symbolic hybrid approaches offers promising avenues for seamless high-level understanding and reasoning about visual scenes. Scene Graph Generation (SGG) is a symbolic image representation approach based on deep neural networks (DNN) that involves predicting objects, their attributes, and pairwise visual relationships in images to create scene graphs, which are utilized in downstream visual reasoning. The crowdsourced training datasets used in SGG are highly imbalanced, which results in biased SGG results. The vast number of possible triplets makes it challenging to collect sufficient training samples for every visual concept or relationship. To address these challenges, we propose augmenting the typical data-driven SGG approach with common sense knowledge to enhance the expressiveness and autonomy of visual understanding and reasoning. We present a loosely-coupled neuro-symbolic visual understanding and reasoning framework that employs a DNN-based pipeline for object detection and multi-modal pairwise relationship prediction for scene graph generation and leverages common sense knowledge in heterogeneous knowledge graphs to enrich scene graphs for improved downstream reasoning. A comprehensive evaluation is performed on multiple standard datasets, including Visual Genome and Microsoft COCO, in which the proposed approach outperformed the state-of-the-art SGG methods in terms of relationship recall scores, i.e. Recall@K and mean Recall@K, as well as the state-of-the-art scene graph-based image captioning methods in terms of SPICE and CIDEr scores with comparable BLEU, ROGUE and METEOR scores. As a result of enrichment, the qualitative results showed improved expressiveness of scene graphs, resulting in more intuitive and meaningful caption generation using scene graphs. Our results validate the effectiveness of enriching scene graphs with common sense knowledge using heterogeneous knowledge graphs. This work provides a baseline for future research in knowledge-enhanced visual understanding and reasoning. The source code is available at <https://github.com/jaleedkhan/neusire>.

Keywords: Scene graph, image representation, common sense knowledge, knowledge enrichment, visual reasoning, image captioning

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

1. Introduction

Neuro-symbolic integration is an emerging area of research that aims to jointly leverage the large-scale learning capability and generalizability of neural approaches along with the reasoning capability and explainability of symbolic approaches in Artificial Intelligence (AI) [33]. These hybrid approaches leverage the unique strengths of each class to broaden their scope and applicability while mitigating their individual limitations. For instance, structured knowledge bases and symbolic reasoning help in explaining as well as improving the performance of black-box neural networks [6]. On the other hand, neural networks and machine learning enable large-scale symbolic reasoning and knowledge base completion [15]. In addition, neuro-symbolic integration enables data and memory efficiency in deep learning [34]. These hybrid approaches either involve utilizing neural representations in symbolic reasoning, infusing symbolic knowledge into neural networks or combining both with the integration of neural and symbolic components ranging from loose to moderate and tight coupling [24,97]. To enable AI to reason with human-like common sense, it is essential to integrate knowledge graphs (KG) with deep learning, which is a crucial aspect of neuro-symbolic integration. Common sense knowledge is implicit and difficult to leverage for reasoning, as people often overlook it when they write or speak about everyday scenarios [37]. However, external domain knowledge and factual information presented in symbolic form by heterogeneous KGs offer a promising source of common sense knowledge that can be integrated with deep learning models.

The past decade witnessed significant advances in deep learning and multi-modal approaches in visual intelligence, resulting in solutions to several challenging problems in basic vision tasks, including image classification, object detection, and image segmentation [28]. However, high-level understanding and reasoning about visual scenes require semantic and relational information, particularly about object interactions. As a result, there is a growing trend toward neuro-symbolic hybrid approaches in the area of visual understanding and reasoning, such as symbolic image representation [43], image captioning [98], image reconstruction [39], multimodal event processing [16], video stream reasoning [51], Visual Question Answering (VQA) [49], and image retrieval [96]. These hybrid techniques have various applications, including visual storytelling [95], autonomous driving [83], mathematical reasoning [71], robotic control [80], and medical diagnosis [29] to name a few. The performance of downstream tasks in visual understanding and reasoning depends on the quality and expressiveness of the image representation. To this end, numerous attempts have been made to explicitly and systematically capture the visual features and object interactions. Scene graph, which models objects and their relationships in a structured and semantically grounded way, has become a widely used symbolic image representation [9]. The process of Scene Graph Generation (SGG), illustrated in Fig. 1(a), involves detection and contextual analysis of objects, attributes, and semantic relationships in a visual scene, followed by constructing symbolic scene representation. The symbolic scene graphs serve as a foundation for higher-level visual reasoning, as illustrated in Fig. 1(b) with examples of image captioning and VQA.

The annotation quality and long-tailed distribution of relationship predicates in crowd-sourced datasets severely impact the relationship prediction accuracy, especially for infrequent relationship predicates, and also limit the expressiveness of SGG. In Visual Genome [50], for instance, the head of the distribution comprises highly generic relationship predicates, such as “on”, “has” and “in” etc., as shown in Fig. 1(c). These relationship predicates have limited significance for visual understanding and reasoning because they cannot completely and clearly express the actual visual relationships in the scene. For example, the relationships *(man, riding, bike)* and *(man, wearing, helmet)* are more expressive as compared to the relationships *(man, on, bike)* and *(man, has, helmet)* as shown in Fig. 1(d). The complexity of visual relationship prediction is further increased by the high variability in the visual appearance of relationships across different scenes and as a result of a huge number of possible object-predicate triplet combinations. For example, the relationships *(man, holding, food)*, *(man, holding, bat)* and *(man, holding, umbrella)* have the same predicate but a very different appearance as shown in Fig. 1(e). To this end, several efforts have been made to address these problems by exploring new aspects of visual relationships, such as saliency [115] and heterophily [60]. In addition, cutting-edge techniques such as knowledge transfer [31], self-supervised learning [75], zero-shot learning [52], counterfactual analysis [84] and linguistic supervision [109] have been employed. However, the performance of SGG is still far from practical and needs to improve accuracy, robustness and expressiveness significantly.

Common sense knowledge infusion is a promising approach to addressing these challenges in visual understanding and reasoning. Since the training datasets used for SGG provide limited or no explicit common sense knowledge,

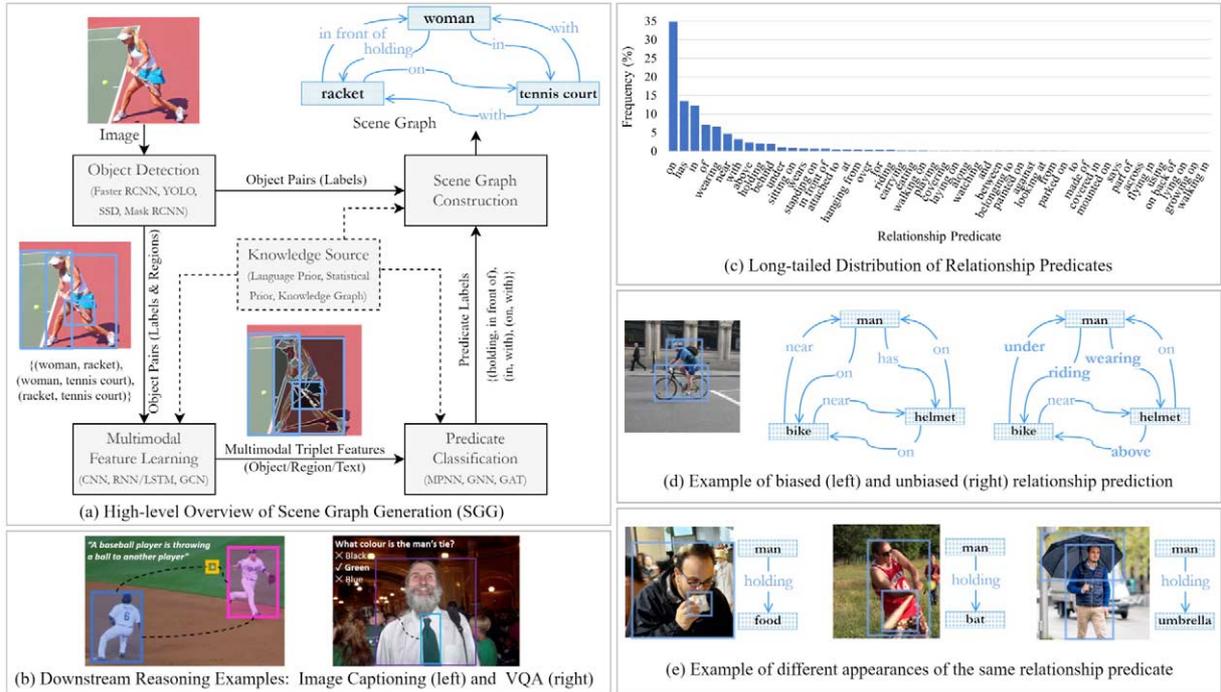


Fig. 1. (a) A high-level overview of scene graph generation (SGG), including the typical components of the SGG pipeline. (b) Downstream reasoning tasks that leverage scene graphs to generate scene descriptions or to answer questions about the scene. (c) The long-tailed distribution problem in visual genome [50] with generic relationship predicates, such as “on”, “has” and “in”, occur more frequently than the more expressive relationship predicates, such as “riding”, “looking at” and “carrying”. This uneven distribution causes (d) bias in relationship prediction. Relationships like *(man, riding, bike)* and *(man, wearing, helmet)* are more expressive than *(man, on, bike)* and *(man, has, helmet)*, but are underrepresented in the training datasets. (e) The challenge of different appearances of the same visual relationship, all of which cannot be covered in training datasets due to the huge number of possible object-predicate triplet combinations and the high variability in the visual appearance of relationships across different scenes.

the background information and related facts about the scene elements can help in improving the expressiveness of the representation, and the performance of downstream reasoning [44]. In this direction, statistical and language priors have been extensively used as sources of common sense knowledge in SGG. However, the heuristics of the statistical priors do not generalize well, and the limitations of semantic word embeddings affect the performance of language priors, especially in the case of infrequent or unseen relationships. Some KGs, such as ConceptNet [81], and WordNet [65], have been leveraged in SGG. These KGs provide text-based and lexical knowledge representing different forms and notions of common sense. Still, they do not provide broad common sense knowledge about visual concepts. Heterogeneous KGs, such as Common Sense Knowledge Graph (CSKG) [38], cover a significantly broader range of dimensions of common sense. These heterogeneous sources are essential but underappreciated sources for common sense knowledge infusion in visual understanding and reasoning. These sources provide a rich and diverse collection of facts about the semantic elements in visual scenes, such as “car is used for transport”, “street is used for parking”, and “car requires parking”. The intelligent integration of heterogeneous KGs can enrich the understanding of complex visual scenes, thus providing rich and expressive representations for effective visual reasoning.

The enriched scene graph shown in Fig. 2 is a motivating example of common sense knowledge-based scene graph representation. The conventional scene graph of the image contains visual relationships, including *(woman, on, tennis_court)* and *(woman, holding, racket)*, that represent objects and their interactions in the scene. Common sense knowledge extracted from CSKG plays a crucial role by providing related facts and background knowledge, such as the edges *(racket, usedFor, playing_tennis)* and *(woman, capableOf, playing_tennis)*, that are essential for reasoning. In this paper, we systematically and substantially extend our previous ESWC work [43], in which we proposed a

common sense knowledge-based SGG technique. It generates a scene graph of an image using a DNN-based vision-language hybrid approach, followed by graph refinement and enrichment that incorporates pertinent details and background information about the visual concepts in the scene using based on the similarity of graph embeddings. The experimental analysis was performed on the VG dataset using Recall@K (R@K) metric for evaluating the visual relationship prediction performance. The main improvements and new contributions made in this paper are listed below.

1. We present a loosely-coupled neuro-symbolic visual understanding and reasoning framework consisting of three main components (Fig. 2):
 - Image Representation: DNN-based object detection and multi-modal pairwise relationship prediction to construct symbolic scene graphs of images.
 - Knowledge Enrichment: Rule-based refinement and enrichment of the scene graphs by leveraging common sense knowledge about the scene entities extracted from a heterogeneous KG.
 - Downstream Reasoning: DNN-based visual reasoning using enriched scene graphs for caption generation.
2. We evaluated the proposed approach on the standard datasets, Visual Genome [50] and Microsoft COCO [12], using the standard evaluation metrics, Recall at K (R@K) [62] and mean R@K (mR@K) [85]. As a result of knowledge enrichment, the relationship recall scores R@100 and mR@100 increased from 36.5 and 11.7 to 39.1 and 12.6, respectively, on the Visual Genome dataset and similar results were observed for the COCO dataset (Fig. 4). We also performed a comparative analysis with the state-of-the-art methods on the benchmark VG dataset using R@K and mR@K metrics, which showed that our approach outperformed them by a significant margin (Table 3).
3. We employed the enriched scene graphs in downstream visual reasoning for image captioning. As a result of enrichment, the SPICE and CIDEr scores of the image captioning model increased from 20.7 and 115.3 to 23.8 and 131.4, respectively (Fig. 7). The qualitative analysis showed that the enriched scene graphs resulted in more intuitive and meaningful captions (Fig. 8). The proposed captioning approach outperformed the state-of-the-art scene graph-based image captioning techniques in terms of SPICE and CIDEr scores and achieved comparable performance in terms of BLEU, ROGUE and METEOR scores (Table 4).

The proposed framework combines neural and symbolic approaches to jointly leverage the efficient learning capabilities of neural networks (in SGG and downstream tasks) and the representational and reasoning power of symbolic approaches (in scene representation and knowledge enrichment). The neural and symbolic components in the proposed framework are loosely coupled as per the taxonomy of neuro-symbolic approaches in [24,97]. Contrary to tightly coupled neuro-symbolic approaches, the neural and symbolic components in our framework operate in tandem to enhance collective performance without directly affecting each other’s internal parameters. The interdependence of the neural and symbolic components is crucial for the framework’s performance, i.e. the accuracy of the predicted scene graph elements plays a crucial role in effectively enriching the representation, which ultimately impacts the performance of downstream reasoning tasks. This type of neuro-symbolic integration is weak but flexible and effective, as depicted by the state-of-the-art results. The rest of the paper is organized as follows: Section 2 presents a comprehensive review of the recent literature on this topic. The proposed neuro-symbolic visual understanding and reasoning framework is explained in Section 3. The comprehensive experiments and results are presented in Section 4. The limitations of the proposed framework and future improvements are discussed in Section 5, followed by the conclusion and prospects in Section 6.

2. Related work

2.1. Image representation

The scene graph is a structured image representation with detailed semantic information about a visual scene, including objects, attributes and visual relationships. The SGG techniques generally follow a bottom-up approach,

as shown in Fig. 1(a), in which objects in an image are detected using DNN-based object detectors, pairwise relationships between the objects are predicted using DNN-based vision-language hybrid features. The object pairs and relationship predicates are linked to construct the symbolic scene graph of the image. The most challenging task in SGG is the prediction of pairwise visual relationships between objects due to highly imbalanced training datasets in terms of relationship predicates [45] (Fig. 1(c–d)), highly varying visual feature representation of the relationships in different scenes (Fig. 1(e)) and insufficient training samples of a huge number of possible triplet combinations, which considerably limit the accuracy, robustness and expressiveness of SGG techniques. The current limitations of SGG and its promising use in various visual reasoning tasks have attracted significant attention in visual intelligence research [9]. The compositional SGG approaches detect the subject, predicate, and object independently and aggregate them subsequently. For example, Li et al. [54] used the detected objects to create independent region proposals for each subject, predicate, and object, consolidated with DNN features and used to predict the relationship triplets. Such approaches are scalable but have limited performance when dealing with unseen or infrequent relationships. On the other hand, the relationship triplets are treated as standalone units by the visual phrase models for SGG. For instance, Sadeghi et al. [77] used DNNs to detect objects and simultaneously predict visual phrases or triplets, which were refined by comparing them to other predictions. As compared to compositional models, the visual phrase models are less sensitive to the diversity of visual relationships, but they require larger training data with a large vocabulary of objects and relationship predicates. Teng et al. [86] proposed Structured Sparse R-CNN, composed of a set of learnable triplet queries that capture the general prior for object pairs and their relations and a structured triplet detection cascade that provides an initial guess of scene graphs for subsequent refinement.

Recent SGG techniques integrate visual and semantic embeddings in DNNs for large-scale visual relationship prediction. Zhang et al. [114] captured visual features in three streams, one for the subject, one for the predicate, and one for the object; the features from the subject and object streams are integrated with the predicate stream to utilize the subject-object interactions for visual relationship prediction. During the learning process, features obtained from the text space are incorporated as labelling for the visual features. Peyre et al. [74] used a visual phrase embedding space during learning to enable analogical reasoning for predicting unseen relationships and to reduce sensitivity to appearance changes of visual relationships. Tang et al. [84] leveraged causal inference with the total direct effect mechanism to alleviate relationship prediction bias in SGG. Zhang et al. [115] proposed visual Saliency-guided Message Passing (SMP) to improve relationship reasoning and generalizability of scene graphs by focusing on the most prominent visual relationships using ordinal regression. Lin et al. [60] exploited heterophily in visual relationships for refining relationship representation and improving message passing in a Graph Neural Network (GNN) along with an adaptive re-weighting transformer module for information integration across layers. Except for a few recent approaches, the existing approaches mainly focus on visual and linguistic patterns in images while ignoring the background knowledge and relevant facts about visual concepts and their structural patterns in heterogeneous KGs, which have significant potential for understanding and interpretation of visual concepts. The approaches explicitly using common sense knowledge in KGs for visual understanding and reasoning are discussed in the following section.

2.2. Knowledge enrichment

Since the 1960s, one of the major challenges in AI has been the acquisition, representation, and reasoning with common sense knowledge [64], which has led the research community to build and compile knowledge sources containing common sense knowledge in various forms and contexts [37]. For common sense knowledge enrichment, early approaches in visual understanding and reasoning relied on statistical and language priors. Deep Relational Network (DR-Net) was proposed to recognize visual relationships, with DNNs leveraging statistical interdependence between objects and predicates [17]. Chen et al. [11] and Zellers et al. [112] used pre-computed frequency priors to incorporate common sense knowledge from dataset statistics for visual relationship prediction. Recently, Zhou et al. [118] proposed a deep sparse graph attention network (DSGAN) for SGG, which uses graph attention networks to learn object and predicate features and constructs a sparse KG representation using statistical co-occurrence information. Lu et al. [62], on the other hand, used region-based CNN for object detection followed

Table 1
Knowledge graphs used for common sense knowledge infusion in SGG

Knowledge source	Knowledge type	Size	Examples	Use in SGG
CSKG [38]	Heterogeneous common sense knowledge consolidated from seven different sources	2.16M nodes, 58 relations, 6M edges	(food, located near, plate), (racket, used for, playing tennis)	Proposed, [43]
ConceptNet [81]	Text-based knowledge about everyday objects, activities, relations, etc.	8M nodes, 36 relations & 21M edges	(food, capable of, go rotten), (chair, used for, sitting)	[26,27,40,110]
VG [50]	Visual knowledge about objects, relations and attributes in images	3.8M nodes, 42k relations, 2.3M edges & 2.8M attributes	(food, on, plate), (woman, looking at, sandwich)	[27,110]
Wordnet [65]	Lexical knowledge about words, concepts, relations, etc.	0.155M words, 10 relations & 0.176M synsets	(car, has part, air bag), (eating, part meronym, chewing)	[110]

by a relationship prediction framework based on semantic word embeddings. Based on Deep Q-network and language priors, Liang et al. [57] proposed a variation-structured reinforcement learning framework for visual relationship prediction. Although SGG approaches based on statistical [11,17,112,118], and language [57,62] priors have improved relationship prediction performance in SGG, these approaches have several drawbacks that limit their expressiveness and applicability in mainstream visual reasoning methods. Statistical priors are often dependent on heuristic approaches that are not generalizable. On the other hand, language priors are vulnerable to the constraints of semantic word embeddings, particularly when generalizing to infrequent objects in training datasets.

KGs have emerged as a viable source of common sense knowledge in visual understanding and reasoning. Table 1 summarizes KGs used for common sense knowledge infusion in SGG. ConceptNet [81] is a multilingual KG with mostly lexical nodes interconnected via 34 relations. The data in ConceptNet is mostly drawn from the crowdsourced Open Mind Common Sense corpus, and it is supplemented with knowledge from other sources such as WordNet. WordNet [65] is a hand-made lexical database with ten relations. WordNet covers over 200 languages and contains terms, meanings, and taxonomical structures. Visual Genome (VG) [50] is a crowd-sourced dataset of images with entity and relationship annotations. VG contains more than 40K relationships, and the concepts are automatically linked to WordNet senses. Seven key KGs [4,47,50,65,78,81,90] containing common sense knowledge in different dimensions were systematically and formally integrated into a rich, well-connected and heterogeneous Common Sense Knowledge Graph (CSKG) [38] with 2.16 million nodes, 58 relations and 6 million edges. Some SGG approaches based on KGs extract relevant knowledge from a KG and integrate it into the model at one of the stages within the SGG pipeline [26,67,82,112]. Alternatively, some approaches employ graph-based message propagation [11,52,99,111] to embed structural information from the KG in the model representations. Wan et al. [91] proposed complementing visual features with common sense knowledge from KGs to improve relationship predicate prediction in SGG. Gu et al. [26] employed recurrent neural networks with an attention mechanism for SGG and encoded background knowledge for each object retrieved from ConceptNet into the network layers. Similarly, Kan et al. [40] leveraged background knowledge from ConceptNet in zero-shot learning for visual relationship prediction in SGG. Existing techniques primarily include triplets from knowledge sources while ignoring the substantial structural information beyond individual triplets. Buffelli et al. [8] proposed a neuro-symbolic regularization technique that uses negative integrity constraints to enforce symbolic background knowledge on a neural model. This is achieved through a logic-based loss function, which amends the neural network to minimize the maximum violation of the integrity constraints. Van et al. [88] introduced Differentiable Fuzzy Logics (DFL) that constructs differentiable loss functions based on fuzzy logic semantics in neuro-symbolic models to perform learning and reasoning simultaneously using gradient descent. To tackle the challenge of fusing deep learning representations with expert knowledge, Diaz et al. [19] proposed a compositional convolutional neural network that utilizes symbolic representations and an explainable training procedure to align deep learning processes with symbolic representations in the form of knowledge graphs. Donadello et al. [20] introduced a semantic image interpretation method based on logic tensor

networks, a statistical relational learning framework. The authors addressed the challenge of zero-shot learning by exploiting similarities with other seen relationships and background knowledge, expressed with logical constraints between subjects, relations, and objects, to predict triples not present in the training set.

When consolidated, the richness, diversity, and coverage of common sense knowledge are merged into a heterogeneous knowledge source, which can have a greater impact on downstream tasks. Zareian et al. [110] proposed Graph Bridging Network (GB-Net) that generates a scene graph, connects its entities and edges to the corresponding entities and edges in a common sense graph retrieved from VG, WordNet, and ConceptNet, and uses GNN-based message propagation to refine the scene graph relationships recursively. Guo et al. [27] extracted relational and common sense knowledge from VG and ConceptNet and encoded it in an Instance Relation Transformer (IRT) for SGG. These SGG techniques employed multiple knowledge sources but have not been employed for visual reasoning tasks, which is important to evaluate the effectiveness of incorporating common sense knowledge from multiple KGs. Some visual reasoning techniques for VQA [61,99] have directly used a subset [93] of DBpedia, ConceptNet, and WebChild; however, these techniques did not use scene graphs, ignoring the structural information about visual concepts. CSKG is the most recent, largest and systematically consolidated common sense knowledge source. Ma et al. [63] used CSKG in language models and reported state-of-the-art performance in common sense question answering by combining diverse, relevant knowledge from CSKG and aligning it with the task. CSKG was employed for knowledge infusion in SGG [43], and the resulting scene graphs were employed for downstream image synthesis; however, there is a significant need for investigation of heterogeneous common sense knowledge-based scene graphs in the mainstream visual reasoning tasks, such as image captioning, VQA and image retrieval. Moreover, some knowledge infusion methods leveraged KG embeddings, widely adopted in the vector representation of entities and relationships in KGs [72]. The KG embeddings capture the latent properties of the semantics in the KG, due to which similar entities are represented with similar vectors. The similarity of entities in the vector space is interpreted using vector similarity measures, such as cosine similarity. KG embeddings have been used in several link prediction tasks, including visual relationship prediction [3], recommender systems [92] and SGG [43].

2.3. Visual reasoning

Scene graphs are widely utilized in image captioning, VQA, MEP, image retrieval, and image synthesis, among the common visual reasoning tasks. The expressiveness and quality of the scene graphs determine the efficacy of these downstream tasks. Image captioning techniques use scene graphs to leverage the pairwise semantic relationships between objects to effectively generate scene descriptions, as it is more challenging to achieve it solely based on vision-language features. Abstract Scene Graph (ASG) [10] representation recognizes and encodes users' intentions in scene graphs along with the semantics information that aids the generation of desired and diverse text descriptions of scenes. The scene graphs generated by SMP [115] based on the saliency of visual relationships were leveraged for improved caption generation. Scene graphs have been found to be more efficient and adaptive than textual scene descriptions for image generation while text-based techniques struggle to sustain performance when the number of objects and their interactions increases [39]. Common sense knowledge-based scene graphs were leveraged in a scene graph-based image synthesis network that resulted in the generation of more realistic images [43]. Gu et al. [26] employed ConceptNet for object and phrase refinement based on common sense knowledge in an attention-based RNN for image reconstruction from scene graphs.

VQA models determine the best answers to questions about visual scenes using multi-modal features and semantic relationships in scene graphs [14]. For example, Zhang et al. [113] proposed encoding the structural information of scene graphs in GNNs to leverage it as the foundation for VQA. Similarly, Ziaeeafard et al. [119] proposed a Graph Attention Networks-based VQA method for encoding scene graphs along with background knowledge from ConceptNet. Graph-based visual semantic models are also used for multimedia stream representation for real-time multimedia event processing in IoMT [46]. Objects and attributes are detected using DNNs, and symbolic rules are employed to identify spatial-temporal interactions between the objects, which are required for matching high-level events questioned by users. In image retrieval, scene graphs are used to explicitly define the semantics and structured information of images, allowing images to be efficiently retrieved from large-scale databases depending on their content. Schroeder et al. [79] presented Structured Query-based Image Retrieval (SQIR) that represents visual interactions in scene graphs as directed sub-graphs for the graph matching task in image retrieval using scene graph

embeddings and structured queries. Donadello et al. [21] presented a novel approach for Semantic Content-Based Image Retrieval (SCBIR) that leverages ontological constraints and low-level image features to generate semantically rich descriptions of image content. The authors propose an unsupervised method where the interpretation of an image is considered a logical model of an ontology describing the image domain.

3. Proposed visual understanding and reasoning framework

The proposed visual understanding and reasoning framework comprises three main components: (1) scene graph generation (neural) for (symbolic) image representation, (2) scene graph enrichment (symbolic) using a common sense KG, and (3) downstream reasoning (neural) to leverage the enriched scene graphs for image captioning. The proposed framework is illustrated in Fig. 2, and each component of the framework is detailed in the following sections. The neural and symbolic components of the framework are loosely coupled but interdependent, i.e. the accuracy of object detection is required for effective enrichment, and the performance of downstream reasoning depends on the quality of scene graphs. This design enables independent operation and modification of each component without affecting the others, thus ensuring flexibility.

3.1. Scene graph generation

The SGG method in the proposed framework uses a multi-modal DNNs cascade for object detection followed by pairwise visual relationship prediction to generate a scene graph of an image. We used Faster RCNN [76] for object detection. The ResNeXt-101-FPN CNN architecture [59] serves as the base feature extractor for the Faster RCNN. For an input image I , the Faster RCNN outputs the object bounding boxes b and object class labels l of each object that is detected in the image. The feature maps F are also taken from the underlying CNN in Faster RCNN, which are used for extracting and encoding region features in a subsequent step.

$$\{b, l, F\} = \text{FasterRCNN}(I)$$

The relationships between object pairs are predicted after object detection and feature map extraction. The region features a of each detected object are computed using RoIAlign [30], which is applied to the image regions $I[b]$ cropped using the object bounding boxes. a , $I[b]$ and l are encoded as the individual visual context features v for each object using Bi-directional Long Short Term Memory (Bi-LSTM) layers [112]. The choice of Bi-LSTM is based on its ability to capture long-term dependencies in data and handle variable-length input sequences; its bi-directional architecture considers the past and future context of object in the image, making it suitable for predicting relationships between objects.

$$a = \text{RoIAlign}(I[b])$$

$$v = \text{BiLSTM}(a, I[b], l)$$

The combined pairwise object features $v_{ij} | i \neq j; i, j = 1, \dots, n$ are obtained by encoding the individual visual context features (v_i, v_j) of objects using Bi-LSTM and concatenating them. n represents the number of detected objects. The language prior p_{ij} is computed by encoding the pairwise object labels (l_i, l_j) through an embedding layer. Applying RoIAlign to the union regions of pairwise objects in the feature maps F allows for the extraction of the contextual union features u_{ij} .

$$v_{ij} = \text{concat}(\text{BiLSTM}(v_i), \text{BiLSTM}(v_j))$$

$$p_{ij} = \text{embed}(\text{concat}(l_i, l_j))$$

$$u_{ij} = \text{conv}(\text{RoIAlign}(F[b_i \cup b_j]))$$

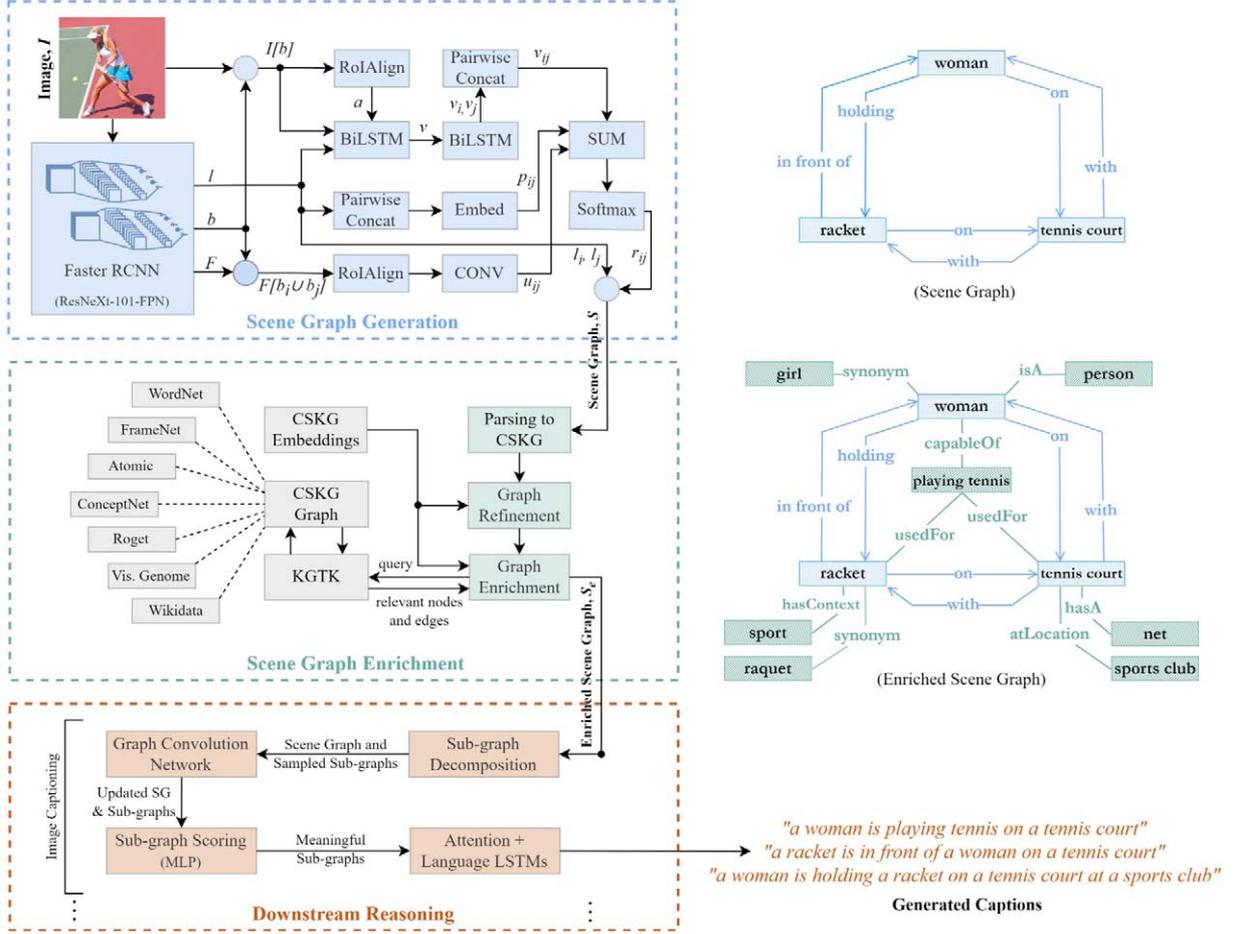


Fig. 2. Proposed neural-symbolic visual understanding and reasoning framework based on enriched scene graph representation.

These three types of extracted features of the object pairs, i.e. v_{ij} , p_{ij} and u_{ij} , are fused using a summation function [23] and used for softmax classification to predict the relationship predicate labels r_{ij} along with their confidence values c_{ij} . Pairwise objects and relationship predicates are finally connected into a structured representation to create the scene graph S .

$$\{r_{ij}, c_{ij}\} = \text{softmax}(\text{SUM}(v_{ij}, p_{ij}, u_{ij}))$$

$$S = \{l_i, r_{ij}, l_j\}$$

3.2. Scene graph enrichment

The scene graph representation captures the objects in an image and their visual relationships. Still, it may not fully convey the meanings and inter-relatedness of the objects, which are crucial for complete visual understanding. To address this limitation, we enhance the scene graphs by infusing common sense knowledge to improve the expressiveness of the representation and correct any inaccurately predicted visual relationships. We used CSKG [38] for the enrichment of scene graphs with background knowledge and related facts in the form of triplets. Graph embeddings were used to compute the similarity of nodes in the graph refinement and enrichment stages, as similar entities often have similar vector representations in the embedding space. Algorithm 1 was used to refine the scene graph predictions by eliminating potentially irrelevant or redundant predictions, indicated by the similarity of labels,

Algorithm 1: Graph refinement

Input: S, b
Output: S_r

- 1 $S_r = []$
- 2 **for** each triplet $\in S$ **do**
- 3 $e_1 = \text{csvg_emb}(\text{triplet}[\text{node1}])$
- 4 $e_2 = \text{csvg_emb}(\text{triplet}[\text{node2}])$
- 5 $b_1 = b[\text{triplet}[\text{node1}]]$
- 6 $b_2 = b[\text{triplet}[\text{node2}]]$
- 7 $\text{metric}_{\text{sim}} = \text{cosine_sim}(e_1, e_2)$
- 8 $\text{metric}_{\text{IoU}} = \text{compute_IoU}(b_1, b_2)$
- 9 **if** $\text{metric}_{\text{sim}} \leq \tau_{\text{sim}} \wedge \text{metric}_{\text{IoU}} \leq \tau_{\text{iou}}$ **then**
- 10 $S_r.\text{append}(\text{triplet})$

Algorithm 2: Graph enrichment

Input: S, G_{csvg}
Output: S_e

- 1 $S_e = S$
- 2 **for** each node $\in S$ **do**
- 3 $e_1 = \text{csvg_emb}(\text{node})$
- 4 $\text{triplets}_{\text{csvg}} = \text{query}(G_{\text{csvg}}, \text{node})$
- 5 $\text{triplets}_{\text{csvg}} = \text{preprocess}(\text{triplets}_{\text{csvg}})$
- 6 **for** each triplet $\in \text{triplets}_{\text{csvg}}$ **do**
- 7 **if** node == triplet[node1] **then**
- 8 $e_2 = \text{csvg_emb}(\text{triplet}[\text{node2}])$
- 9 **else**
- 10 $e_2 = \text{csvg_emb}(\text{triplet}[\text{node1}])$
- 11 $s = \text{cosine_sim}(e_1, e_2)$
- 12 **if** $s \geq \tau \wedge \text{triplet} \notin S_e$ **then**
- 13 $S_e.\text{append}(\text{triplet})$
- 14 $S_e = \text{postprocess}(S_e)$

overlapping of bounding boxes, and the structural patterns of their corresponding nodes in CSKG. At this stage, the prediction errors have been reduced by eliminating objects with high intersection over union (IoU) of its bounding boxes or high CSKG embedding similarity scores with another object in the scene graph.

We utilized Knowledge Graph Toolkit (KGTK) [36] to query CSKG and pull new triplets that contain a subject or object node in the predicted scene graph. Duplicate triplets and triplets with the same node on both ends, such as (person, synonym, person) and (chair, similarTo, chair), are not useful and thus eliminated in the pre-processing stage prior to complementing the scene graph with the new triplets. Next, we connect the nodes of the new triplets that have reasonable structural similarities with the corresponding object nodes in the scene graph to those object nodes. This enables the addition of new triplets to the scene graph. If a node from a new triplet already exists in the scene graph, we link the edge of that triplet to the existing object node instead of creating a redundant node. After enriching the scene graphs with common sense knowledge, the enriched scene graphs need to be employed for downstream reasoning tasks. This requires the enriched scene graphs to be post-processed to match the original representation model of scene graphs for consistent integration with downstream reasoning models. In the VG subset

of CSKG, all predicates of the triplets are expressed as “LocatedNear” edge type. We substitute the predicates of the new triplets having VG as their source in CSKG with the most common predicate in the original VG dataset between the nodes of those triplets. This step relies on the statistical prior knowledge about relationships in VG for consistency and better interpretation of visual relationships in downstream reasoning. The process of complementing scene graphs with common sense knowledge from CSKG is outlined in Algorithm 2. During the experiments, a threshold of 0.5 was set in both algorithms to achieve a balance between the quantity and precision of visual relationships detected through SGG and those added via knowledge enrichment.

3.3. Downstream caption generation

The proposed framework includes scene graph-based image captioning [117] as a downstream task of scene graph generation and knowledge enrichment, which enables it to generate precise and meaningful captions using the enriched scene graphs. The proposed framework can be extended to include more scene graph-based visual reasoning tasks, such as VQA, multimodal event processing, and image retrieval.

The enriched scene graphs are first decomposed into sub-graphs and the sub-graphs are sampled using neighbour sampling [48]. This involves selecting a random set of seed nodes on the graph and extracting the immediate neighbours of these nodes, along with the edges connecting them, to get a sampled sub-graph. Similar sub-graphs are removed to avoid redundancy, and greedy non-maximal suppression is used to filter out sub-graphs with high IoU scores of their nodes. The nodes and edges of the scene graph are then augmented with their visual features and text embeddings using a Graph Convolutional Network (GCN). The GCN aggregates information from the neighbouring nodes via multiple graph convolutions and ReLU layers to integrate contextual information within the scene graph. The relationship information has been integrated using GCN; thus, only the features of nodes obtained from the GCN are used for *sub-graph scoring* to select the most meaningful sub-graphs. A two-layer Multi-Layer Perceptron (MLP) with a sub-graph readout function [102] is used to rank sub-graphs, followed by sigmoid normalization. Finally, an attention-based LSTM is used to assign importance scores to the sub-graph nodes. These scores are used by a language LSTM to generate sentences corresponding to each sub-graph. Sub-graph scoring and ranking ensure that the network prioritizes the most relevant and meaningful triplets for generating captions. The enriched scene graph-based caption generation pipeline is illustrated in Fig. 2.

4. Experiments and results

4.1. Experimental setup

4.1.1. Platform specifications and tools

We used a machine with AMD Ryzen 7 1700 Eight-Core Processor, 16 GB RAM, NVIDIA TITAN Xp GPU (with 12 GB memory) and Ubuntu 18.04 LTS (64-bit) operating system for implementation and experiments. We used the PyTorch deep learning library¹ for implementing the scene graph generation and image captioning methods and KGTK² for implementing the graph refinement and enrichment algorithms.

4.1.2. Datasets and knowledge source

We used the Visual Genome (VG) [50] and Microsoft COCO [12] datasets for experimental analysis and benchmark comparison of SGG. VG contains 108K labelled images and annotations for objects and visual relationships. COCO contains 132K labelled images with annotations for objects and captions. The standard subset [101] of VG contains the most frequent 50 predicate classes and 150 object classes, which was used for training Faster RCNN and SGG pipeline. 70% of the training samples were used for training, out of which 5000 samples were used for validation during training. The remaining 30% of samples were used for testing. Following the state-of-the-art methods, we used the standard split [41] of COCO for the evaluation of enriched scene graph-based image captioning. The

¹<https://pytorch.org/>

²<https://kgtk.readthedocs.io/>

standard split comprises 5K images each for validation and testing and the rest for training. We used the pre-trained CSKG embeddings [38] for computing the similarity of nodes in the graph refinement and enrichment steps of the scene graph enrichment part of the proposed framework.

4.1.3. Evaluation metrics

We used cross-entropy **loss** to evaluate the training performance of the Faster RCNN and SGG models. Cross-entropy loss determines how well the probability distribution output P by the softmax layer in the model matches the one hot encoded ground truth label L of the object or relationship.

We used mean average precision (mAP) [25] to evaluate the object detection performance of the Faster RCNN model. **mAP** is defined as the arithmetic mean of the average precision values for detection of N object categories, where the average precision (AP) for an object category is calculated as the area under the precision-recall curve. For evaluation of object detection, *precision* is the percentage of detected objects that are correct, while *recall* is the percentage of all objects present in the videos that are correctly detected. Precision and recall are computed based on the number of true positives (TP), false positives (FP) and false negatives (FN) among the detected objects. TP , FP and FN are determined by comparing the predicted label and the ground truth label for each detected object and by applying a threshold to the Intersection over Union (IoU), which is the overlap ratio between the predicted bounding box and the ground truth bounding box of an object.

For evaluating the performance of SGG, we used the commonly used evaluation metrics for relationship prediction, i.e. Recall@K ($R@K$) and mean Recall@K ($mR@K$). **R@K** is defined as the fraction of times the correct relationship is predicted in the top K confident relationship predictions [62]. The confidence score is taken into account in $R@K$, requiring the relationship labels to be correctly predicted as well as to hold a higher score. **mR@K** is the arithmetic mean of $R@K$ values that are independently calculated for each relationship category in order to minimize the bias towards dominant relationships during the evaluation [11,85]. The standard evaluation protocol for SGG was followed. The predicted scene graphs were enriched as a part of the proposed approach during the experiments, and the ground truth scene graphs in the test set were left unchanged to ensure a fair comparison with the existing approaches.

BLEU score [73] and **METEOR** score [5] were primarily introduced for machine translation. BLEU score is based on n-gram precision between sentences taking into account n-grams up to length four. METEOR favours the recall of matching unigrams from the candidate and reference sentences, i.e. alignment between words, in their exact form, stemmed form, and meaning. BLEU and METEOR are typically more effective for corpus-level comparisons as compared to sentence-level comparisons. **ROUGE** score [58] was initially intended for text summarization, and its variant **ROGUE-L** is widely used for caption generation. ROGUE-L considers the longest subsequence of tokens in the same relative order, potentially with other tokens in between, that exists in both candidate and reference caption. These evaluation metrics for image captioning are adapted from those used for Natural Language Processing (NLP) tasks including text summarization and machine translation. The reference **CIDEr** score [89] was designed for the evaluation of caption generation and it is based on 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, taking precision and recall into consideration. TF gives more weight to n-grams that appear frequently in reference sentences describing an image, whereas IDF gives less weight to n-grams that appear frequently in all descriptions, thus, IDF calculates word saliency by discounting popular terms that are less visually informative. The **SPICE** score [2] is the latest, best correlated to human judgements and most relevant to scene graph-based image captioning evaluation. SPICE score takes into account matching tuples retrieved from the candidate and reference scene graphs, due to which, it prefers semantic information over fluency in text and better simulates human judgment.

4.2. Results

4.2.1. Training and evaluation of models

The two main components of the SGG pipeline in the proposed framework, i.e. Faster RCNN for object detection and the LSTM-based deep learning cascade for relationship prediction, are separately trained and evaluated. The Faster RCNN model was trained on images and ground truth annotations of objects in the dataset using Stochastic

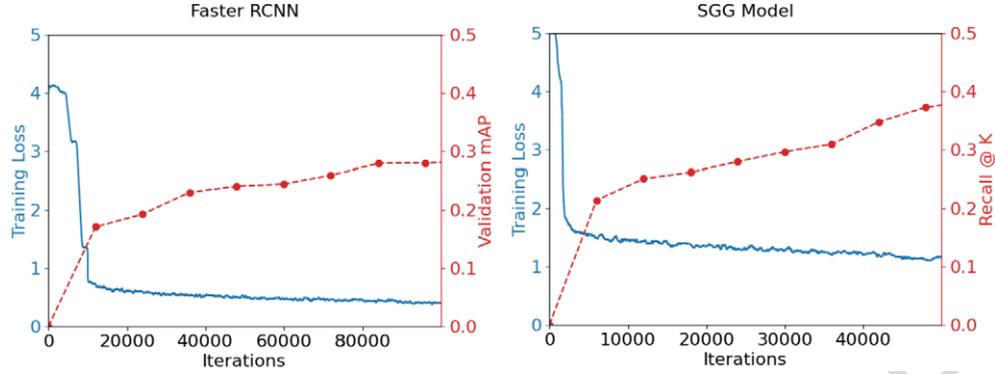


Fig. 3. Training progress plots along with periodic validation checks of the faster RCNN and SGG models.

Gradient Descent (SGD) as an optimizer, a batch size of 2, and an initial learning rate of 0.002, which was reduced by a factor of 10 after 60,000 and 80,000 iterations. We froze the trained Faster RCNN model and trained the entire SGG pipeline on images and ground truth annotations of visual relationships in the dataset using SGD as an optimizer, batch size of 4, and an initial learning rate of 0.04 that was reduced by a factor of 10 twice during training when validation performance stopped improving considerably. The Scene Graph Detection (SGDet) configuration was used for training and evaluation of the SGG pipeline. Figure 3 depicts the plots of training loss and validation mAP for object detection and training loss and $R@100$ for scene graph detection, which demonstrate a smooth convergence of the models during the training phase. The Faster RCNN model achieved 29.19 mAP (with a 0.5 IoU threshold) on the test set, while the SGG model achieved 36.5 $R@100$.

4.2.2. Post-enrichment evaluation

We repeated the SGG evaluation after integrating the proposed knowledge enrichment steps after the typical scene graph generation. We obtained $R@K = 29.9, 35.5, 39.1$ on the VG test set for $K = 20, 50, 100$, which is significantly higher than the $R@K$ values obtained using the conventional scene graphs (without knowledge enrichment), i.e. $R@K = 26.1, 32.7, 36.5$, as shown in Fig. 4. A similar trend is observed in the case of COCO test set, in which $R@K$ increased from 24.0, 32.9, 36.2 to 27.9, 36.3, 38.5 for $K = 20, 50, 100$. Visual cues regarding the spatial placement of objects in the scene relative to each other and physical interactions between the objects provided by CSKG helps reduce the number of missed or incorrect predictions made during scene graph construction and increase the recall rate for relationship prediction. The recall rate for each relationship predicate category in the VG dataset is shown in Fig. 5. Knowledge enrichment helps improve the recall rate for almost all relationship predicates, including the ones towards the tail of the distribution. This depicts the potential of knowledge enrichment in alleviating bias towards frequent relationships in datasets.

4.2.3. Benchmark comparison

Table 2 and Table 3 present a detailed comparison of the proposed enriched scene graph-based approach with the state-of-the-art SGG techniques evaluated on the benchmark VG dataset. The recall scores for each method are reported for the same experimental setting of SGG, i.e. SGDet, for the sake of fair comparison. The performance of the proposed SGG method without knowledge enrichment is comparable to the existing data-centric SGG techniques, as shown in Table 2. On the other hand, the proposed SGG method combined with knowledge enrichment obtained considerably higher recall scores, outperforming the state-of-the-art knowledge-based SGG techniques, as shown in Table 3. The proposed method achieved $R@K = 29.9, 35.5, 39.1$ for $K = 20, 50, 100$, while the latest technique [118] among the common sense knowledge-based SGG techniques has a recall score of $R@K = 23.2, 28.8, 32.9$ and the latest technique with no external knowledge infusion [60] has a recall score of $R@K = 26.0, 33.7, 38.1$. The superior performance of the proposed method depicts the effectiveness of incorporating the most recent, largest, and most diverse heterogeneous KG as a common sense knowledge source for scene graph enrichment.

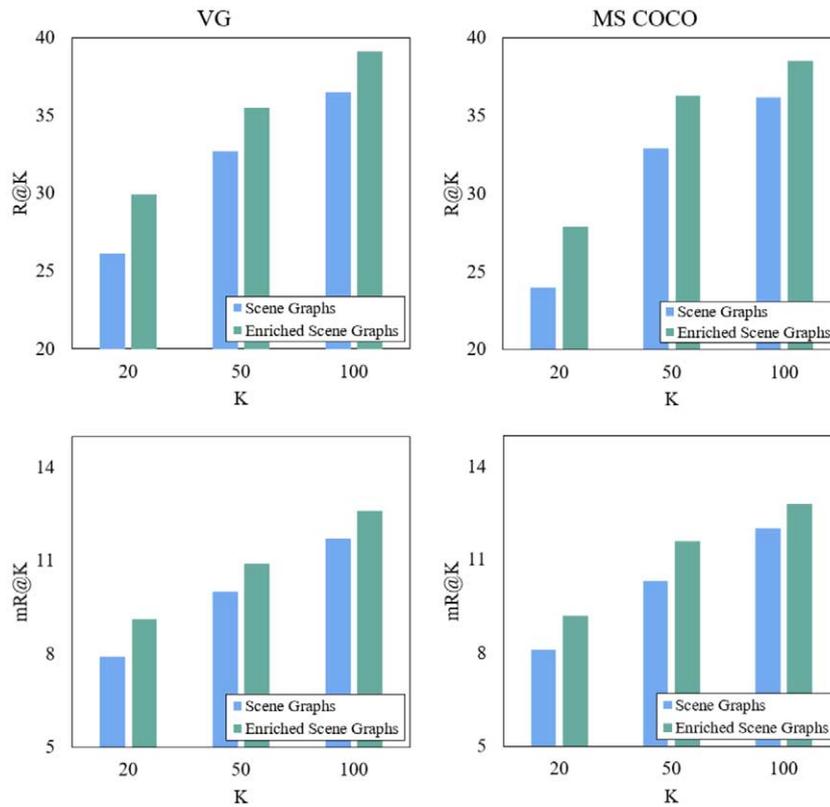


Fig. 4. Comparison of conventional and enriched scene graphs on VG and COCO datasets using recall rates $R@K$ and $mR@K$.

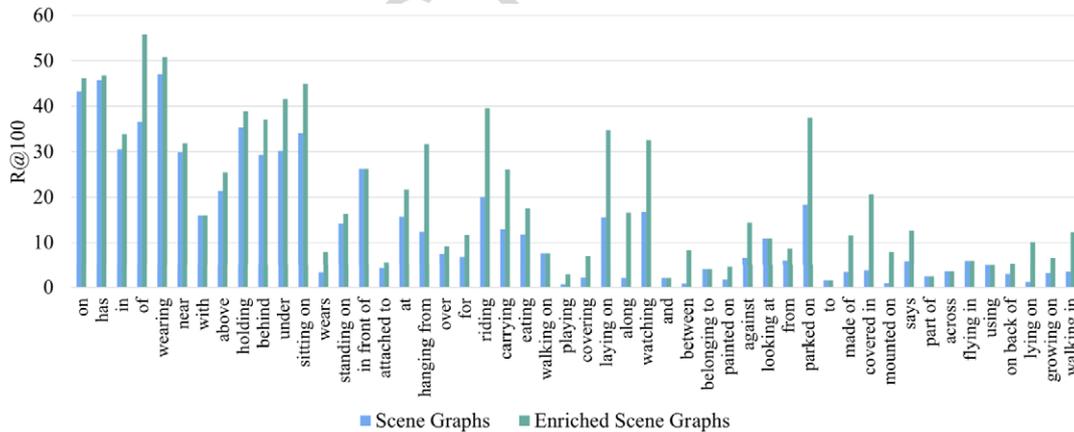


Fig. 5. Detailed comparison of conventional and enriched scene graphs on top 50 relationship predicates in VG dataset using $R@100$.

4.2.4. Qualitative results

Figure 6 shows some qualitative results of the proposed enriched scene graph-based SGG approach. The enriched scene graphs contain background facts about the underlying concepts, additional knowledge about the spatial placement of objects in the scene relative to each other, and possible physical interactions between the objects, in addition to the objects and their pairwise visual relationships. The scene graph representations are supplemented by common sense relationships about object interactions, such as (*person, holding, surfboard*) in the first row in Fig. 6, and

Table 2

Detailed comparison of the proposed method with the state-of-the-art data-centric SGG methods using common evaluation metrics (R@K and mR@K) and standard split of the VG dataset

SGG method	Approach	SGG performance		Downstream task
		R@20/50/100 (%)	mR@20/50/100 (%)	
HL-Net [60]	Transformer & MP based Heterophily Learning	26.0/ 33.7 / 38.1	-/-/9.2	-
TDE [84]	Causal Inference and Total Direct Effect	25.8/33.3/37.8	6.9/9.3/11.1	-
SS-RCNN [86]	Structured Sparse R-CNN	25.8/32.7/36.9	6.1/8.4/10.0	-
SMP [115]	Visual Relation Saliency-guided Message Passing	-/32.6/36.9	-/-/-	Image Captioning
Proposed (SGG)	SGG based on Fusion of Visual-Textual Features	26.1 /32.7/36.5	7.9 /10.0/11.7	Image Cap. & Gen.
NICEST [53]	Noisy label correction & training for Robust SGG	-/29.0/32.7	-/ 10.4 / 12.4	-
VCTree [85]	Dynamic tree structures & Bi-dir TreeLSTM	22.0/27.9/31.3	5.2/6.9/8.0	Visual Q/A
IMP+ [101]	Object and relationship feature refinement via MP	14.6/20.7/24.5	-/3.8/4.8	-
FactorizableNet [55]	Clustering-based graph factorization	-/13.1/16.5	-/-/-	-
MSDN [56]	Scene description at object, phrase & caption levels	-/10.7/14.2	-/-/-	-
Graph RCNN [105]	RPN followed by Attention GCN	-/11.4/13.7	-/-/-	-

Table 3

Detailed comparison of the proposed method with the state-of-the-art common sense knowledge-based methods using common evaluation metrics (R@K and mR@K) and standard split of the VG dataset

SGG method	Approach	Knowledge source	SGG performance		Downstream task
			R@20/50/100 (%)	mR@20/50/100 (%)	
Proposed	Multi-modal DNN-based SGG and enrichment using CSKG	Heterogeneous KG: CSKG [38]	29.9 / 35.5 / 39.1	9.1 / 10.9 / 12.6	Image Captioning & Generation
DSGAN [118]	Deep sparse graph attention network	Sparse KG & Statistical Prior	23.2/28.8/32.9	7.8/8.9/11.8	-
IRT-MSK [27]	Instance Relation Transformer with Mult. Struc. Knowledge	KGs: CN [81], VG [50]	21.9/27.8/31.0	-/-/-	-
MOTIFS [112]	RNN-LSTM based Stacked Motif Networks	Statistical Prior	21.4/27.2/30.3	4.2/5.7/6.6	-
GB-Net [110]	Message passing between scene graphs and common sense graph	KGs: CN [81], WN [65], VG [50]	-/26.4/30.0	-/6.1/7.3	-
KERN [11]	Knowledge-embedded routing network	Statistical Prior	22.3/27.1/29.8	-/6.4/7.3	-
COACHER [40]	Zero-shot relationship prediction via common sense infusion	KG: ConceptNet [81]	13.4/19.3/22.2	-/-/-	-
KB-GAN [26]	Common sense, reconstr. based object & phrase refinement	KG: ConceptNet [81]	-/13.6/17.6	-/-/-	Image Generation
DeepVRL [57]	Deep Q-network for variation-structured reinf. learning	Language Prior	-/13.3/12.6	-/-/-	-
VRD [62]	Relationship prediction using semantic word embeddings	Language Prior	-/0.3/0.5	-/-/-	Image Retrieval

spatial placement, such as (*tree, on, street*) in the last row of Fig. 6. In the third row in Fig. 6, (*person, requires, eating*) and (*food, usedFor, eating*) represent useful background facts extracted from CSKG.

4.2.5. Downstream caption generation

We trained the image captioning network on the COCO dataset that was used to train the SGG pipeline. The trained network was used to generate captions using conventional scene graphs as well as enriched scene graphs. The performance of the image captioning network using both types of scene graphs is evaluated in terms of the

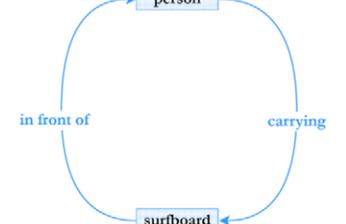
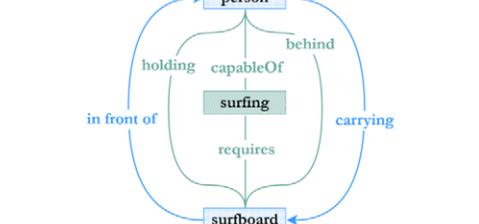
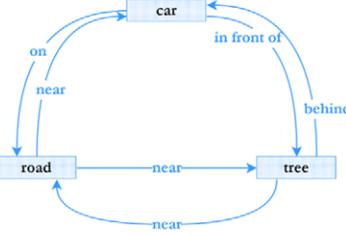
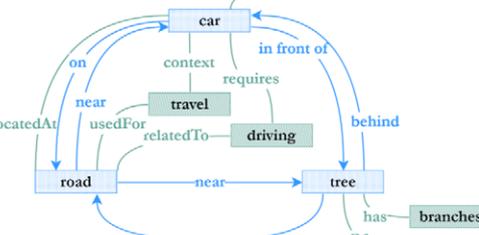
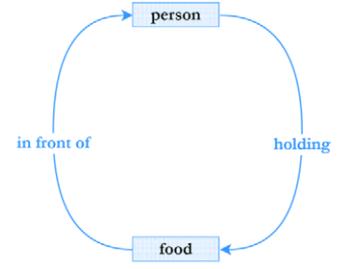
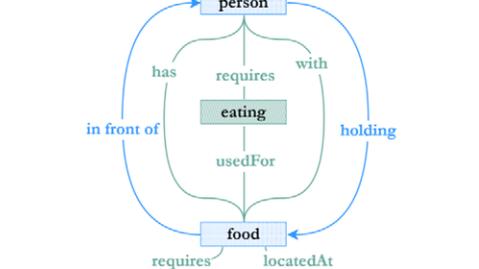
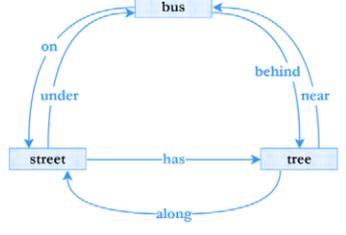
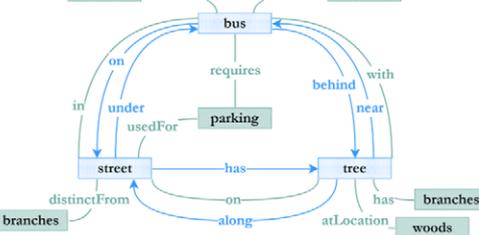
Image	Scene Graph	Enriched Scene Graph
		
		
		
		

Fig. 6. Examples of the proposed enriched scene graphs for visual understanding and reasoning (VG images).

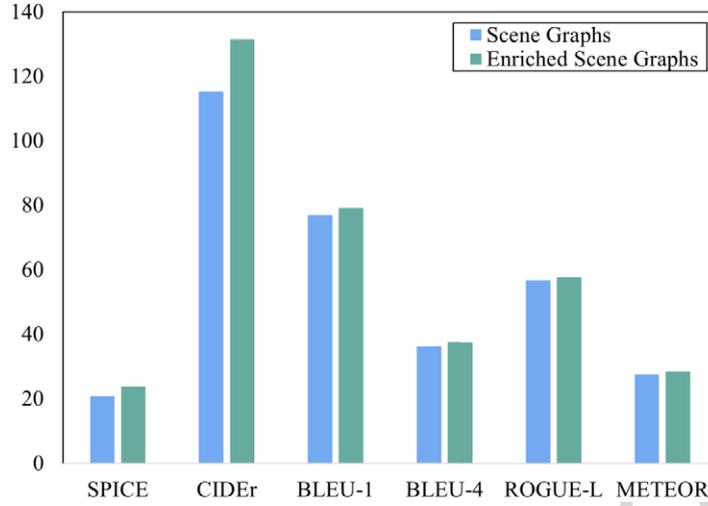


Fig. 7. Comparison of image captioning using conventional scene graphs and proposed enriched scene graphs in terms of the standard evaluation metrics. Enriched scene graphs resulted in higher SPICE and CIDEr scores and comparable performance in terms of BLEU, ROGUE and METEOR scores.

Table 4

Comparison of the proposed enriched scene graph-based image captioning method with the state-of-the-art conventional scene graph-based image captioning methods using the common evaluation metrics and standard split [41] of the COCO dataset [12]. The proposed approach outperformed the state-of-the-art methods in terms of SPICE and CIDEr scores and achieved comparable performance in terms of BLEU, ROGUE and METEOR scores

Method	SPICE	CIDEr	BLEU-1	BLEU-4	ROGUE-L	METEOR
Proposed	23.8	131.4	79.1	37.6	57.7	28.5
Yang et al. [107]	22.4	129.6	81.0	38.8	58.8	28.8
Yang et al. [106]	20.9	116.3	77.3	36.8	57.0	27.9
Yao et al. [108]	20.9	116.7	77.6	36.9	57.2	27.7
Zhong et al. [117]	20.7	115.3	76.8	36.2	56.6	27.7
Ke et al. [42]	20.5	115.3	77.5	36.8	56.8	27.2
Anderson et al. [2]	20.3	113.5	77.2	36.2	56.4	27.0
Nguyen et al. [69]	19.8	106.6	–	32.6	55.0	26.4

standard evaluation metrics, including SPICE, BLEU, CIDEr, ROGUE and METEOR, which is presented in Fig. 7. The SPICE and CIDEr scores obtained by the image captioning model increased from 20.7 and 115.3 to 23.8 and 131.4, respectively with the use of enriched scene graphs, which depicts the efficacy of enriched scene graphs for image captioning. The performance of both types of scene graphs is comparable in terms of BLEU, ROGUE and METEOR scores. Table 4 shows the performance comparison of the proposed enriched scene graph-based image captioning approach with the state-of-the-art scene graph-based image captioning techniques. The proposed approach outperforms the state-of-the-art techniques in terms of SPICE and CIDEr scores and achieves comparable performance in terms of BLEU, ROGUE and METEOR scores.

SPICE and CIDEr are the most reliable among these metrics because SPICE best simulates human judgment in the evaluation by leveraging semantic and structural information, while CIDEr was originally designed for scene graph-based image captioning. Since the network divides the enriched scene graph into subgraphs and prioritizes the most relevant and meaningful triplets for creating captions, it generally produces more meaningful text, leading to higher evaluation scores. Some qualitative results of caption generation using conventional and enriched scene graphs are shown in Fig. 8. The promising results show that enriched scene graphs result in more expressive and meaningful captions as compared to conventional scene graphs.

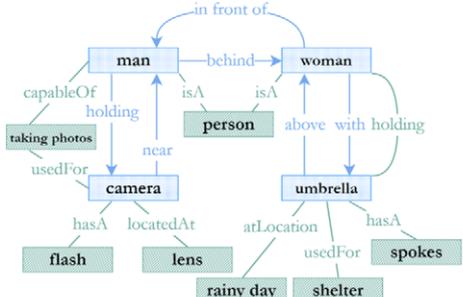
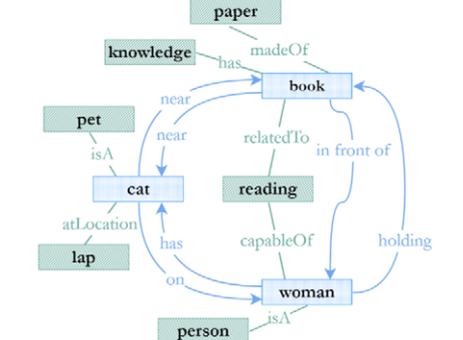
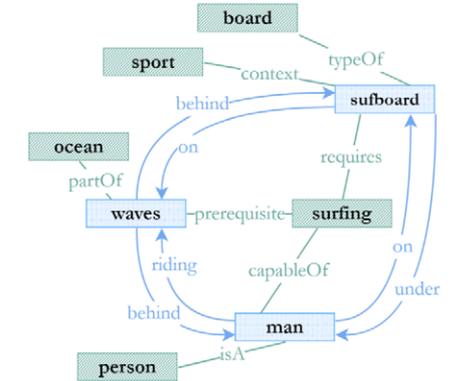
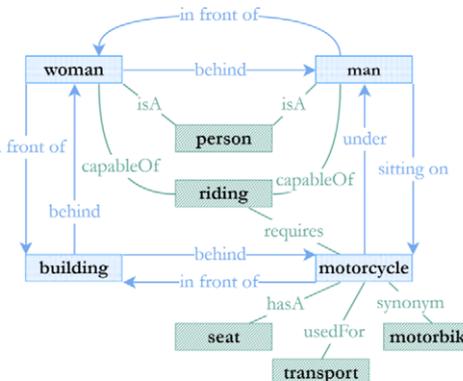
Image	Enriched Scene Graph	Caption
		<p>a man is holding a camera standing next to a woman with an umbrella</p> <p>a man taking a picture of a woman holding an umbrella in the rain</p>
		<p>a woman is holding a book near a cat</p> <p>a woman is reading a book while a cat sits on her lap</p>
		<p>a man is riding waves on a surfboard</p> <p>a man is surfing in the ocean on a surfboard</p>
		<p>a woman is sitting behind a man on a motorcycle in front of a building</p> <p>a man and a woman are riding on a motorcycle outside a building</p>

Fig. 8. Qualitative results of caption generation using conventional scene graphs (blue) and enriched scene graphs (green). Enriched scene graphs result in more expressive and meaningful image captions. (COCO images.)

5. Limitations and future improvements

5.1. Limited contextual relevance

Heterogenous KGs are currently the richest and most diverse sources of common sense knowledge as they capture detailed structural and semantic features of general concepts in the world. The enrichment of scene graphs using heterogeneous KGs has demonstrated its potential to enhance the overall performance of SGG, as depicted by our results. Due to limited contextual knowledge [22], heterogenous KGs cannot always provide contextually valid information about visual concepts in a specific scene. Although a cosine similarity threshold is employed to effectively remove irrelevant relationships, it does not account for the contextual relevance of the new relationships, which can potentially lead to the addition of out-of-context relationships, limiting contextual reasoning ability in downstream tasks.

This highlights the need to evaluate the quality of enrichment based on the ratio of accurate and contextually valid relationships among the newly added relationships and use context-aware approaches [32,66] to ensure that only relevant as well as contextually valid relationships are added during enrichment, thus refining the enriched scene graph further and leading to improved downstream reasoning. Future work can also investigate approaches with feedback mechanisms, adaptive thresholds and domain-specific knowledge for incorporating new relationships based on contextual similarity in addition to structural and semantic similarity for more accurate and reliable scene graph enrichment. Future work in this direction would enable contextually accurate image generation [43] and historical image colourization [100].

5.2. Lack of knowledge-infused learning in DNNs

We have demonstrated that incorporating common sense knowledge to enrich scene representation improves the accuracy and expressiveness of SGG and enhances downstream reasoning. However, in our current approach, we only use common sense knowledge to enrich the scene graphs and do not integrate it into the neural learning process. Alternatively, common sense knowledge can be directly infused into neural learning: within the layers or feedback mechanism of DNNs [1,18] for SGG. This can enable DNNs to learn the patterns of visual relationships more deeply and effectively, leading to more accurate SGG that will no more require scene graph refinement. Enrichment can only add external contextual knowledge to the scene graphs, leading to more meaningful visual understanding and reasoning.

While this research direction has been explored to some extent [26,110], it would be useful to investigate how leveraging heterogeneous common sense knowledge can alleviate the challenges of such approaches. CSKG offers a diverse knowledge base that consolidates multiple sources of knowledge, making it interesting to explore and identify the common relation types and sources within CSKG that are most useful for visual understanding and reasoning. Heterogenous KGs can also help extract rules about visual concepts and encode them into DNNs [35,68] for visual understanding and reasoning. In addition, leveraging hard constraints from ontologies as prior knowledge [19,21], i.e. ontological priors, could help minimize errors in scene graph generation. Future work along these lines can pave the way for leveraging the maximum potential of common sense knowledge enrichment and infusion in visual understanding and reasoning.

5.3. Temporal patterns in visual relationships

The proposed framework can effectively process image data to extract semantic elements in scenes, predict visual relationships between them and enrich the representation for improved downstream reasoning. Visual understanding and reasoning also requires processing video data, where visual relationships can vary temporally, and common sense knowledge and rules about the varying patterns can be essential for reasoning. The proposed framework, in its current form, can only process each video frame individually, which can be computationally inefficient and might overlook the temporal aspects of visual relationships.

To address this limitation, incorporating object tracking [7], temporal dimensions of visual relationships [95], and graph aggregation [103,116] can help achieve semantically-rich and knowledge-enhanced summarization, providing

a more concise and meaningful representation of video content. Integrating video data processing capabilities can expand the framework’s capabilities beyond static image understanding and reasoning. For example, it can use temporal relationship variations and common sense knowledge to identify congestion patterns and detect suspicious behaviour in smart city surveillance applications [87].

5.4. Limited generalizability to new concepts

Scene graphs rely on visual relationship prediction that struggles with the long-tailed distribution problem of crowdsourced datasets, limiting their generalizability to unseen or rare visual relationships. Many meaningful relationship predicates have only a few instances, making it difficult for SGG methods to learn their feature representations. On the other hand, frequent predicates are generic and do not express visual relationships as clearly as underrepresented predicates. In addition, visual feature representations of relationships can vary greatly across different scenes. Collecting and annotating sufficient training samples for all possible triplet combinations is nearly impossible, indicating the need for zero-shot approaches in addition to augmenting data-centric SGG methods with common sense knowledge.

Zero-shot [40,99] and few-shot [27] learning approaches for SGG have been explored to address this limitation. Zero-shot learning approaches utilize prior knowledge about seen relationships to recognize unseen visual relationships. In contrast, few-shot learning approaches learn from a small set of labelled examples for new relationships, which is useful where collecting large amounts of labelled data is costly and time-consuming. These approaches can leverage heterogeneous KGs to incorporate common sense knowledge and retrieve relevant triplets for improved prediction of unseen and rare visual relationships. Furthermore, knowledge transfer and distillation approaches [74,107] for SGG can employ models trained on heterogeneous KGs to use prior knowledge about visual relationships for enhanced generalization in SGG, making it more practical and applicable in real-world scenarios.

5.5. Weak neuro-symbolic integration

Loose coupling between neural and symbolic components in the proposed framework provides greater flexibility and ease of modification. However, the independent operation of neural and symbolic components in the framework can limit its ability to fully exploit the complementary strengths of neural learning and symbolic reasoning. The neural components may not be able to effectively leverage the knowledge encoded in the symbolic component, and the symbolic component may not be able to effectively incorporate the rich visual features learned by the neural component. While a tighter neuro-symbolic integration can lead to the increased complexity of the framework, it also has the potential to effectively combine the strengths of the neural and symbolic components to reason about complex relationships and improve its overall performance. As an example, instead of a sequential process, a feedback mechanism can be incorporated for error correction in scene graph generation, which could complement and further improve upon corrections carried out in the enrichment method.

To this end, few SGG techniques [26,27,110] have explored neuro-symbolic integration with neural learning diluted with external knowledge. Still, their applicability to downstream reasoning and incorporation of rich and diverse common sense knowledge is limited. A fully integrated neuro-symbolic visual understanding and reasoning framework can have symbolic structures integrated and grounded within neural learning, such as using tensor calculus to imitate logical reasoning in DNNs [94]. It can also be achieved with program synthesis approaches [13,70] or structured knowledge augmentation in reinforcement learning models [104]. An end-to-end approach will make the framework capable of inductive learning for logical reasoning in visual understanding and reasoning. Incorporating heterogeneous common sense knowledge at the same time can further strengthen the neuro-symbolic integration in visual understanding and reasoning, paving the way for groundbreaking progress in the field.

6. Conclusion

Scene graph is a semantically rich, symbolic image representation generated using DNNs, which is used for several visual reasoning tasks, including image captioning, VQA, image retrieval, multimedia event processing and

image synthesis. Scene graph enrichment using heterogeneous KGs is a promising approach toward alleviating the existing challenges in SGG and improving the expressiveness of visual understanding and reasoning frameworks. We proposed a loosely-coupled neuro-symbolic visual understanding and reasoning framework based on enriched scene graphs. A DNNs cascade is used to generate symbolic scene graphs, followed by rule-based graph refinement and enrichment using common sense knowledge extracted from a heterogeneous KG in the form of related facts and background information about the scene graph elements. We integrated an image captioning model in the proposed framework as a downstream task of scene graph enrichment. The evaluation results showed that common sense knowledge enrichment resulted in a significant increase in the relationship recall scores $R@100$ and $mR@100$ from 36.5 and 11.7 to 39.1 and 12.6, respectively, on the VG test set. The proposed framework outperformed the state-of-the-art methods in terms of $R@K$ and $mR@K$ on the standard split of VG in the comparative analysis. These encouraging results depict the efficacy of scene graph enrichment using heterogeneous KGs. Moreover, the enriched scene graphs resulted in an increase in SPICE and CIDEr scores obtained by the downstream image captioning model from 20.7 and 115.3 to 23.8 and 131.4, respectively. The proposed approach outperformed the state-of-the-art scene graph-based image captioning techniques in terms of SPICE and CIDEr scores and achieved comparable performance in terms of BLEU, ROGUE and METEOR scores. The future work will focus on multi-hop KG reasoning to further refine visual relationship detection, zero- and few-shot methods with knowledge transfer for improved generalization and scalability, including more downstream reasoning tasks and strengthening the neuro-symbolic integration in visual understanding and reasoning.

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] 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.
- [2] 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.
- [3] S. Baier, Y. Ma and V. Tresp, Improving visual relationship detection using semantic modeling of scene descriptions, in: *International Semantic Web Conference*, Springer, 2017, pp. 53–68.
- [4] C.F. Baker, C.J. Fillmore and J.B. Lowe, The Berkeley framenet project, in: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Vol. 1, 1998, pp. 86–90.
- [5] 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.
- [6] A. Bennetot, J.-L. Laurent, R. Chatila and N. Díaz-Rodríguez, Towards explainable neural-symbolic visual reasoning, arXiv preprint, 2019. [arXiv:1909.09065](https://arxiv.org/abs/1909.09065).
- [7] 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. 23–28.
- [8] D. Buffelli and E. Tsamoura, Scalable regularization of scene graph generation models using symbolic theories, arXiv preprint, 2022. [arXiv:2209.02749](https://arxiv.org/abs/2209.02749).
- [9] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen and A. Hauptmann, Scene graphs: A survey of generations and applications, arXiv preprint, 2021. [arXiv:2104.01111](https://arxiv.org/abs/2104.01111).
- [10] 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.
- [11] 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.
- [12] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár and C.L. Zitnick, Microsoft coco captions: Data collection and evaluation server, arXiv preprint, 2015. [arXiv:1504.00325](https://arxiv.org/abs/1504.00325).

- [13] X. Chen, C. Liang, A.W. Yu, D. Zhou, D. Song and Q.V. Le, Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension, in: *International Conference on Learning Representations*, 2020.
- [14] Z. Chen, J. Chen, Y. Geng, J.Z. Pan, Z. Yuan and H. Chen, Zero-shot visual question answering using knowledge graph, in: *International Semantic Web Conference*, Springer, 2021, pp. 146–162.
- [15] W.W. Cohen, H. Sun, R.A. Hofer and M. Sieglar, Scalable neural methods for reasoning with a symbolic knowledge base, arXiv preprint, 2020. [arXiv:2002.06115](https://arxiv.org/abs/2002.06115).
- [16] 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).
- [17] B. Dai, Y. Zhang and D. Lin, Detecting visual relationships with deep relational networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3076–3086.
- [18] H. Dai, Y. Tian, B. Dai, S. Skiena and L. Song, Syntax-directed variational autoencoder for structured data, arXiv preprint, 2018. [arXiv:1802.08786](https://arxiv.org/abs/1802.08786).
- [19] N. Díaz-Rodríguez, A. Lamas, J. Sanchez, G. Franchi, I. Donadello, S. Tabik, D. Filliat, P. Cruz, R. Montes and F. Herrera, EXplainable neural-symbolic learning (X-NeSyL) methodology to fuse deep learning representations with expert knowledge graphs: The MonuMAI cultural heritage use case, *Information Fusion* **79** (2022), 58–83. doi:[10.1016/j.inffus.2021.09.022](https://doi.org/10.1016/j.inffus.2021.09.022).
- [20] I. Donadello and L. Serafini, Mixing low-level and semantic features for image interpretation: A framework and a simple case study, in: *Computer Vision – ECCV 2014 Workshops*, Zurich, Switzerland, September 6–7 and 12, 2014, Proceedings, Part II 13, Springer, 2015, pp. 283–298.
- [21] I. Donadello and L. Serafini, Compensating supervision incompleteness with prior knowledge in semantic image interpretation, in: *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019, pp. 1–8.
- [22] A. Ettore, 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.
- [23] C. Feichtenhofer, A. Pinz and A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1933–1941.
- [24] A.D. Garcez and L.C. Lamb, Neurosymbolic AI: The 3rd wave, *Artificial Intelligence Review* (2023), 1–20.
- [25] R. Girshick, J. Donahue, T. Darrell and J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.
- [26] 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.
- [27] 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](https://doi.org/10.1145/3394171.3414025).
- [28] M. Hassaballah and A.I. Awad, *Deep Learning in Computer Vision: Principles and Applications*, CRC Press, 2020.
- [29] M. Hassan, H. Guan, A. Melliou, Y. Wang, Q. Sun, S. Zeng, W. Liang, Y. Zhang, Z. Zhang, Q. Hu et al., Neuro-symbolic learning: Principles and applications in ophthalmology, arXiv preprint, 2022. [arXiv:2208.00374](https://arxiv.org/abs/2208.00374).
- [30] K. He, G. Gkioxari, P. Dollár and R. Girshick, Mask R-CNN, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961–2969.
- [31] T. He, L. Gao, J. Song, J. Cai and Y.-F. Li, Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation, arXiv preprint, 2020. [arXiv:2006.07585](https://arxiv.org/abs/2006.07585).
- [32] N. Heist, Towards knowledge graph construction from entity co-occurrence, in: *EKAW (Doctoral Consortium)*, 2018.
- [33] 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](https://doi.org/10.3233/SW-190368).
- [34] 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](https://doi.org/10.1093/nsr/nwac035).
- [35] 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.
- [36] F. Ilievski, D. Garijo, H. Chalupsky, N.T. Divvala, Y. Yao, C. Rogers, R. Li, J. Liu, A. Singh, D. Schwabe et al., KGTK: A toolkit for large knowledge graph manipulation and analysis, in: *International Semantic Web Conference*, Springer, 2020, pp. 278–293.
- [37] F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D.L. McGuinness and P. Szekely, Dimensions of commonsense knowledge, arXiv preprint, 2021. [arXiv:2101.04640](https://arxiv.org/abs/2101.04640).
- [38] F. Ilievski, P. Szekely and B. Zhang, CSKG: The commonsense knowledge graph, in: *European Semantic Web Conference*, Springer, 2021, pp. 680–696.
- [39] 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.
- [40] 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.
- [41] A. Karpathy and L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
- [42] L. Ke, W. Pei, R. Li, X. Shen and Y.-W. Tai, Reflective decoding network for image captioning, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8888–8897.
- [43] 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](https://doi.org/10.1007/978-3-031-06981-9_6).

- [44] 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](https://doi.org/10.1109/MIC.2022.3176500).
- [45] 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.
- [46] 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 '20) Workshops*, 2020.
- [47] B. Kipfer, *Roget's 21st Century Thesaurus in Dictionary Form*, 3rd edn. The Philip Lief Group, Inc, New York, 2005.
- [48] J.M. Klusowski and Y. Wu, Counting motifs with graph sampling, in: *Conference on Learning Theory*, PMLR, 2018, pp. 1966–2011.
- [49] 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.
- [50] 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](https://doi.org/10.1007/s11263-016-0981-7).
- [51] D. Le-Phuoc, T. Eiter and A. Le-Tuan, A scalable reasoning and learning approach for neural-symbolic stream fusion, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, 2021, pp. 4996–5005.
- [52] 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.
- [53] L. Li, L. Chen, H. Shi, H. Zhang, Y. Yang, W. Liu and J. Xiao, NICEST: Noisy label correction and training for robust scene graph generation, arXiv preprint, 2022. [arXiv:2207.13316](https://arxiv.org/abs/2207.13316).
- [54] Y. Li, W. Ouyang, X. Wang and X. Tang, Vip-CNN: Visual phrase guided convolutional neural network, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1347–1356.
- [55] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang and X. Wang, Factorizable net: An efficient subgraph-based framework for scene graph generation, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 335–351.
- [56] 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.
- [57] 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.
- [58] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text Summarization Branches Out*, 2004, pp. 74–81.
- [59] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, Feature pyramid networks for object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125.
- [60] 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.
- [61] L. Liu, M. Wang, X. He, L. Qing and H. Chen, Fact-based visual question answering via dual-process system, *Knowledge-Based Systems* (2021), 107650.
- [62] 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.
- [63] K. Ma, F. Ilievski, J. Francis, Y. Bisk, E. Nyberg and A. Oltramari, Knowledge-driven data construction for zero-shot evaluation in commonsense question answering, in: *35th AAAI Conference on Artificial Intelligence*, 2021.
- [64] J. McCarthy et al., Programs with common sense, RLE and MIT computation center, 1960.
- [65] G.A. Miller, WordNet: A lexical database for English, *Communications of the ACM* **38**(11) (1995), 39–41. doi:[10.1145/219717.219748](https://doi.org/10.1145/219717.219748).
- [66] 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](https://doi.org/10.18653/v1/P19-1081).
- [67] 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.
- [68] 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).
- [69] K. Nguyen, S. Tripathi, B. Du, T. Guha and T.Q. Nguyen, In defense of scene graphs for image captioning, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1407–1416.
- [70] M. Nye, A. Solar-Lezama, J. Tenenbaum and B.M. Lake, Learning compositional rules via neural program synthesis, *Advances in Neural Information Processing Systems* **33** (2020), 10832–10842.
- [71] 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.
- [72] M. Palmonari and P. Minervini, Knowledge graph embeddings and explainable AI, in: *Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges*, IOS Press, Amsterdam, 2020, pp. 49–72.
- [73] 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.
- [74] J. Peyre, I. Laptev, C. Schmid and J. Sivic, Detecting unseen visual relations using analogies, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1981–1990.

- [75] A. Prakash, S. Debnath, J.-F. Lafleche, E. Cameracci, S. Birchfield, M.T. Law et al., Self-supervised real-to-sim scene generation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16044–16054.
- [76] 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](https://doi.org/10.1109/TPAMI.2016.2577031).
- [77] M.A. Sadeghi and A. Farhadi, Recognition using visual phrases, in: *CVPR 2011*, IEEE, 2011, pp. 1745–1752. doi:[10.1109/CVPR.2011.5995711](https://doi.org/10.1109/CVPR.2011.5995711).
- [78] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N.A. Smith and Y. Choi, Atomic: An atlas of machine commonsense for if-then reasoning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019, pp. 3027–3035.
- [79] 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.
- [80] T. Silver, A. Athalye, J.B. Tenenbaum, T. Lozano-Perez and L.P. Kaelbling, Learning neuro-symbolic skills for bilevel planning, arXiv preprint, 2022. [arXiv:2206.10680](https://arxiv.org/abs/2206.10680).
- [81] 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.
- [82] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen and J. Li, Learning visual knowledge memory networks for visual question answering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7736–7745.
- [83] 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.
- [84] 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.
- [85] 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.
- [86] Y. Teng and L. Wang, Structured sparse R-CNN for direct scene graph generation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19437–19446.
- [87] 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](https://doi.org/10.1145/3543873.3587665).
- [88] E. van Krieken, E. Acar and F. van Harmelen, Analyzing differentiable fuzzy logic operators, *Artificial Intelligence* **302** (2022), 103602. doi:[10.1016/j.artint.2021.103602](https://doi.org/10.1016/j.artint.2021.103602).
- [89] 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.
- [90] D. Vrandečić and M. Krötzsch, Wikidata: A free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85. doi:[10.1145/2629489](https://doi.org/10.1145/2629489).
- [91] H. Wan, J. Ou, B. Wang, J. Du, J.Z. Pan and J. Zeng, Iterative visual relationship detection via commonsense knowledge graph, in: *Joint International Semantic Technology Conference*, Springer, 2019, pp. 210–225.
- [92] H. Wang, F. Zhang, X. Xie and M. Guo, DKN: Deep knowledge-aware network for news recommendation, in: *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1835–1844.
- [93] 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](https://doi.org/10.1109/TPAMI.2017.2754246).
- [94] P.-W. Wang, P. Donti, B. Wilder and Z. Kolter, Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 6545–6554.
- [95] 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.
- [96] S. Wang, R. Wang, Z. Yao, S. Shan and X. Chen, Cross-modal scene graph matching for relationship-aware image-text retrieval, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1508–1517.
- [97] W. Wang and Y. Yang, Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing, arXiv preprint, 2022. [arXiv:2210.15889](https://arxiv.org/abs/2210.15889).
- [98] X. Wang, L. Ma, Y. Fu and X. Xue, Neural symbolic representation learning for image captioning, in: *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 312–321. doi:[10.1145/3460426.3463637](https://doi.org/10.1145/3460426.3463637).
- [99] 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.
- [100] 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.
- [101] 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.
- [102] K. Xu, W. Hu, J. Leskovec and S. Jegelka, How powerful are graph neural networks? arXiv preprint, 2018. [arXiv:1810.00826](https://arxiv.org/abs/1810.00826).
- [103] 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](https://doi.org/10.1109/GC46384.2019.00011).
- [104] F. Yang, D. Lyu, B. Liu and S. Gustafson, Pearl: Integrating symbolic planning and hierarchical reinforcement learning for robust decision-making, arXiv preprint, 2018. [arXiv:1804.07779](https://arxiv.org/abs/1804.07779).
- [105] J. Yang, J. Lu, S. Lee, D. Batra and D. Parikh, Graph R-CNN for scene graph generation, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 670–685.

- [106] X. Yang, K. Tang, H. Zhang and J. Cai, Auto-encoding scene graphs for image captioning, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10685–10694.
- [107] 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).
- [108] T. Yao, Y. Pan, Y. Li and T. Mei, Exploring visual relationship for image captioning, in: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 684–699.
- [109] 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.
- [110] 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.
- [111] A. Zareian, S. Karaman and S.-F. Chang, Weakly supervised visual semantic parsing, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3736–3745.
- [112] 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.
- [113] C. Zhang, W.-L. Chao and D. Xuan, An empirical study on leveraging scene graphs for visual question answering, arXiv preprint, 2019. [arXiv:1907.12133](https://arxiv.org/abs/1907.12133).
- [114] 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.
- [115] 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).
- [116] 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.
- [117] Y. Zhong, L. Wang, J. Chen, D. Yu and Y. Li, Comprehensive image captioning via scene graph decomposition, in: *European Conference on Computer Vision*, Springer, 2020, pp. 211–229.
- [118] 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](https://doi.org/10.1016/j.patcog.2021.108367).
- [119] M. Ziaefard 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](https://doi.org/10.18653/v1/2020.coling-main.169).