NeuSIRE: Neural-Symbolic Image Representation and Enrichment for Visual Understanding and Reasoning

Muhammad 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, National University of Ireland, Galway, Ireland*
[b] *Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland, Galway, Ireland*
E-mails: m.khan12@nuigalway.ie, john.breslin@nuigalway.ie, edward.curry@nuigalway.ie

Abstract. The adoption of neural-symbolic hybrid approaches in visual intelligence is essential to progress toward seamless high-level understanding and reasoning about visual scenes. In this direction, Scene Graph Generation (SGG) is a promising and challenging task, which involves the prediction of objects, their attributes and pairwise visual relationships in a visual scene to create a structured, symbolic scene representation, known as a scene graph, which is utilized in downstream visual reasoning to perform a desired task, such as image captioning, visual question answering, image retrieval, multimedia event processing or image synthesis. The crowdsourced training datasets used for this purpose are highly imbalanced and it is nearly impossible to collect and collate training samples for every visual concept or visual relationship due to a huge number of possible combinations of objects and relationship predicates. Leveraging commonsense knowledge is a natural solution to augment the data-driven approaches with external knowledge to enhance the expressiveness and autonomy of visual understanding and reasoning frameworks. In this paper, we proposed a neural-symbolic visual understanding and reasoning framework based on commonsense knowledge enrichment. Deep neural network-based techniques are used for object detection and multi-modal pairwise relationship prediction to generate a scene graph of an image, which is followed by rule-based algorithms to refine and enrich the scene graph using commonsense knowledge. The commonsense knowledge is extracted from a heterogeneous knowledge graph in the form of related facts and background information about the scene graph elements. The enriched scene graphs are then leveraged in downstream visual reasoning pipelines. We performed comprehensive evaluation of the proposed framework using the common datasets and standard evaluation metrics. As a result of commonsense 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 (VG) dataset and similar results were observed for the COCO dataset. The proposed framework outperformed the state-of-the-art methods in terms of R@K and mR@K on the standard split of VG. We incorporated scene graph-based image captioning and image generation models as downstream tasks of SGG with knowledge enrichment. With the use of enriched scene graphs, SPICE and CIDEr scores obtained by the image captioning model increased from 20.7 and 115.3 to 23.8 and 131.4 respectively, and 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 qualitative results of image generation showed that the enriched scene graphs result in more realistic images in which the semantic concepts in the input scene graph can be more clearly observed. The encouraging results validate the effectiveness of knowledge enrichment in scene graphs using heterogeneous knowledge graphs. The source code is available at https://github.com/jaleedkhan/neusire.

Keywords: scene graph, image representation, commonsense knowledge, knowledge enrichment, visual reasoning, image captioning, image generation

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

1. Introduction

Neural-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) [1]. Such hybrid approaches benefit from the unique capabilities of each of these two classes of AI approaches to broaden their current scope and applicability and alleviate their individual limitations. For example, structured knowledge bases and symbolic reasoning help in explaining as well as improving the performance of black-box neural networks [2] and neural networks enable symbolic approaches to perform reasoning at a large scale and use machine learning for knowledge base completion [3]. The integration of knowledge graphs and deep learning is also promising for AI to emulate human-like common sense reasoning capabilities. Although the implicit nature of common sense knowledge makes it difficult to leverage it for reasoning as it is commonly recognized and used by people in everyday scenarios but it is typically overlooked when they write or speak. However, the external domain knowledge and factual information provided by heterogeneous knowledge graphs in symbolic form is a promising source of common sense knowledge for infusion in neural approaches.

Within visual intelligence, the significant advances in deep learning and multi-modal approaches during the past decade helped in solving several challenging problems in the basic vision tasks including image classification, object detection and image segmentation. Most of the vision techniques since the ImageNet-2012 breakthrough [4] are based on Deep Neural Networks (DNN). Semantic and relational information about a visual scene and its constituents, especially object interactions, is vital for high-level understanding and reasoning about the scene. Due to this reason, there is a growing trend toward neural-symbolic hybrid approaches for visual understanding and reasoning. The examples include approaches for symbolic image representation [5] and several downstream reasoning tasks including image captioning [6], image reconstruction [7], Multimodal Event Processing (MEP) for the Internet of Multimedia Things (IoMT) [8], video stream reasoning [9], image retrieval [10], visual storytelling [11] and Visual Question Answering (VQA) [12]. The performance of the downstream tasks in visual understanding and reasoning depends on the quality and expressiveness of the image representation. There have been numerous attempts to explicitly and systematically capture the visual features and object interactions. The scene graph has become a widely used symbolic image representation that captures the intricate semantics of a visual scene and models objects and their relationships in a structured and semantically grounded way. The process of Scene Graph Generation (SGG) [13], illustrated in Figure 1(a), entails the detection and contextual analysis of objects, attributes, and semantic relationships in a visual scene, followed by the construction of symbolic scene representation. The symbolic scene graphs are subsequently used for higher-level visual reasoning; the examples of image captioning and VQA are shown in Figure 1(b).

The annotation quality and long-tailed distribution of relationship predicates in the crowd-sourced datasets severely impact the accuracy and expressiveness of SGG because relationship prediction is a crucial component of SGG. In the case of Visual Genome (VG) [14] which is the most widely used dataset in SGG works, the head of the distribution comprises highly generic relationship predicates, such as 'on', 'has' and 'in' etc. as shown in Figure 1(c). These relationship predicates are insignificant for SGG because they fail to 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 Figure 1(d). This biased distribution also limits the accuracy of SGG models, especially in predicting the infrequent relationship predicates. Moreover, the difference in visual appearance of relationships in different scenes or for different object pairs due to the huge number of possible combinations of object-predicate triplets adds to the complexity of visual relationship detection. 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 Figure 1(e). To this end, several efforts have been made to address these problems by exploring new aspects of visual relationships, such as saliency [15] and heterophily [16], as well as, by utilizing cutting-edge techniques such as knowledge transfer [17], self-supervised learning [18], zero-shot learning [19], counterfactual analysis [20] and linguistic supervision [21]. However, the performance of SGG is still far from practical and it needs to significantly improve in terms of accuracy, robustness and expressiveness.

Commonsense knowledge infusion in visual understanding and reasoning is a potential solution to overcoming the existing challenges. Since the training datasets used for SGG provide limited or no explicit commonsense knowl-

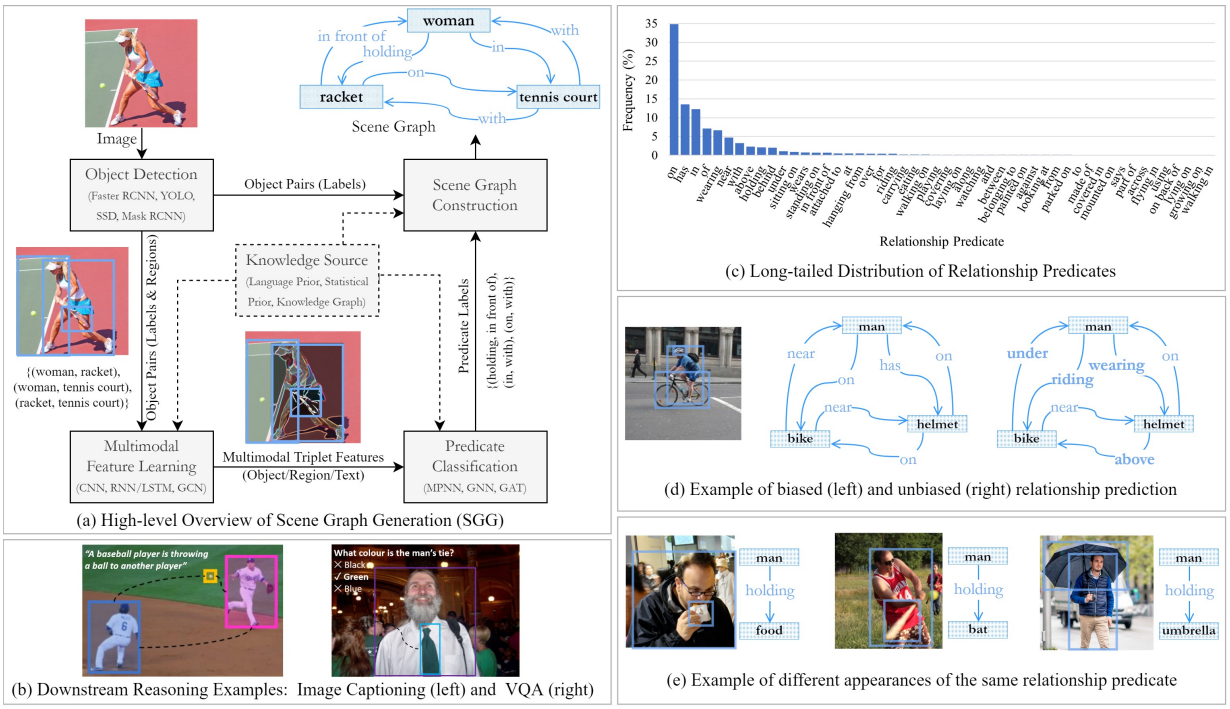


Fig. 1. (a) A high-level overview of Scene Graph Generation (SGG) including the typical components of the SGG pipeline. (b) Downstream reasoning tasks that leverage scene graphs to generate scene descriptions or to answer questions about the scene. (c) The long-tailed distribution problem in VG [14] with generic relationship predicates (on, has, in, etc.) more frequently occurring than the more expressive relationship predicates (riding, looking at, carrying etc.). 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 they are underrepresented in the training datasets. (e) Another challenge of different appearances of the same visual relationship, all of which are not covered in training datasets due to a huge number of different appearances of relationships as well as possible triplet combinations of objects and relationship predicates.

edge, 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 [22]. In this direction, statistical priors and language priors have been extensively used as sources of commonsense 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 knowledge graphs, such as ConceptNet [23] and WordNet [24], have been leveraged in SGG. These knowledge graphs provide text-based and lexical knowledge about objects that represent different forms and notions of common sense but they individually do not provide broad common sense knowledge about visual concepts. Heterogeneous knowledge graphs, such as Common Sense Knowledge Graph (CSKG) [25], cover a significantly broader range of dimensions of common sense. These heterogeneous sources are essential, but underappreciated sources for commonsense knowledge infusion in visual understanding and reasoning. Such 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' etc. Intelligent integration of heterogeneous knowledge graphs can enrich the understanding of complex visual scenes, thus providing rich and expressive representations for effective visual reasoning.

The enriched scene graph shown in Figure 2 is a motivating example of commonsense 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. Although it is simple and obvious to us that the woman is playing tennis, it is difficult for machines to deduce that without the assistance of some external common sense knowledge. Relevant edges extracted from CSKG, such as *(racket, usedFor, playing_tennis)* and *(woman, capableOf, playing_tennis)*, provide related facts and background knowledge that are essential for visual reasoning. In this paper, we systematically and substantially extend a previous ESWC work

[5] which proposed a commonsense 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 contributions made in this paper are listed below.

1. We proposed a neural-symbolic visual understanding and reasoning framework (Figure 2) with three main components:
   - Image Representation: DNN-based object detection and multi-modal pairwise relationship prediction, followed by symbolic scene graph construction
   - Knowledge Enrichment: Rule-based refinement and enrichment of the scene graph representation by leveraging commonsense knowledge about the scene entities extracted from a heterogeneous knowledge graph
   - Downstream Reasoning: DNN-based reasoning for downstream visual reasoning tasks including image captioning, image generation, etc.
2. We performed SGG evaluation on the benchmark Visual Genome [14] and Microsoft COCO [26] datasets using the standard Recall at K (R@K) [27] and mean R@K (mR@K) [28] metrics. 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 (Figure 4). Similar results were observed for the COCO dataset.
3. We performed a detailed comparative analysis with the state-of-the-art methods on the benchmark VG dataset using R@K and mR@K metrics, which showed that SGG with knowledge enrichment in the proposed framework outperformed the state-of-the-art methods by a significant margin (Table 2).
4. We incorporated a scene graph-based image captioning network [29] as a downstream task of scene graph generation and knowledge enrichment. 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 (Figure 6). 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 (Table 3), which depicts the efficacy of enriched scene graphs in image captioning.
5. We incorporated a scene graph-based image generation network [30] as a downstream task of scene graph generation and knowledge enrichment and showed that the enriched scene graphs result in more realistic images in which the semantic concepts in the input scene graph can be more clearly observed (Figure 8).

Rest of the paper is organized as follows: Section 2 presents a comprehensive review of the recent literature on this topic. The proposed neural-symbolic visual understanding and reasoning framework is presented in Section 3, which is followed by experiments and results in Section 4 and the conclusion and future prospects in Section 5.

2. Related Work

2.1. Image Representation

The scene graph is a structured image representation with detailed semantic information of a visual scene including objects, attributes and visual relationships. The SGG techniques generally follow a bottom-up approach, as shown in Figure 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, and 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 (Figure 1(c-d)), highly varying visual feature representation of the relationships in different scenes (Figure 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 a variety of visual reasoning tasks have made it a hot topic in

visual intelligence [13]. The compositional SGG approaches detect the subject, predicate, and object independently and aggregate them subsequently, for example, Li et al.[31] used the detected objects to create independent region proposals for subject, predicate, and object, which were consolidated with DNN features to predict the relationship triplets. Such approaches are scalable, however, they have very restricted performance when dealing with unseen or infrequent relationships. On the other hand, the relationship triplets are treated as single units by the visual phrase models for SGG. For instance, Sadeghi et al. [32] used DNNs to detect objects as well as predict visual phrases or triplets, which they subsequently refined by comparing to other predictions. When 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.

Recent SGG techniques integrate visual and semantic embeddings in DNNs for large-scale visual relationship prediction. Zhang et al. [33] 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. [34] 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. [20] leveraged causal inference and total direct effect in an attempt to alleviate relationship prediction bias in SGG caused by the long-tailed distribution problem. Zhang et al. [15] 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. [16] 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. The current approaches mainly focus on visual and linguistic patterns in images while ignoring the background knowledge and relevant facts about concepts in the visual scenes and the structural patterns of scene graph elements in heterogeneous knowledge graphs, which have significant potential for understanding and interpretation of visual concepts. A few recent approaches discussed in the following section explicitly used commonsense knowledge in knowledge graphs for visual understanding and reasoning.

2.2. Knowledge Enrichment

Since the 1960s, one of the major challenges in AI has been the acquisition, representation, and reasoning with commonsense knowledge [35], which has led the research community to build and compile knowledge sources containing commonsense knowledge in various forms and contexts [36]. For commonsense 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 [37]. Chen et al. [38] and Zellers et al. [39] used pre-computed frequency priors to incorporate commonsense knowledge from dataset statistics for visual relationship detection. Recently, Zhou et al. [40] 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 knowledge graph representation using statistical co-occurrence information. Lu et al. [27], on the other hand, used region-based CNN for object detection followed by a relationship prediction framework based on semantic word embeddings. Based on Deep Q-network and language priors, Liang et al. [41] proposed a variation-structured reinforcement learning framework for visual relationship detection. Although SGG approaches based on statistical [37–40] and language [27, 41] priors have improved relationship prediction performance in SGG, these approaches have several significant drawbacks that limit their expressiveness and applicability in mainstream visual reasoning methods. Statistical priors are often dependent on heuristic approaches that are not generalizable, whereas language priors are vulnerable to the constraints of semantic word embeddings, particularly when generalizing to infrequent objects in training datasets.

Recently, knowledge graphs have emerged as a viable source of commonsense knowledge in visual understanding and reasoning. Table 1 summarizes knowledge graphs that have been used for commonsense knowledge infusion in SGG. ConceptNet [23] is a multilingual knowledge graph 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 [24] is a hand-made lexical database

Table 1
Knowledge Graphs used for Commonsense Knowledge Infusion in SGG

Knowledge Source	Knowledge Type	Size	Examples	Use in SGG
CSKG [25]	Consolidated commonsense knowledge from the above seven sources	2.16M nodes, 58 relations, 6M edges	(food, located near, plate), (racket, used for, playing tennis)	Proposed, [5]
ConceptNet [23]	Text-based knowledge about everyday objects, activities, relations, etc.	8M nodes, 36 relations & 21M edges	(food, capable of, go rotten), (chair, used for, sitting)	[7, 42–44]
VG [14]	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)	[43, 44]
Wordnet [24]	Lexical knowledge about words, concepts, relations, etc.	0.155M words, 10 relations & 0.176M synsets	(car, has part, air bag), (eating, part meronym, chewing)	[44]

with 10 relations. WordNet covers over 200 languages and contains terms, meanings, and taxonomical structures. Visual Genome (VG) [14] 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 knowledge graphs [14, 23, 24, 45–48] containing commonsense knowledge in different dimensions were systematically and formally integrated into a rich, well-connected and heterogeneous Common Sense Knowledge Graph (CSKG) [25] with 2.16 million nodes, 58 relations and 6 million edges. Some SGG approaches based on knowledge graphs extract relevant knowledge from a knowledge graph and integrate it into the model at one of the stages within the SGG pipeline [7, 39, 49, 50], while some approaches employ graph-based message propagation [19, 38, 51, 52] to embed structural information from the knowledge graph in the model representations. Wan et al. [53] proposed complementing visual features with commonsense knowledge from knowledge graphs to improve relationship predicate prediction in SGG. Gu et al. [7] 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. [42] leveraged background knowledge from ConceptNet in zero-shot learning for visual relationship detection in SGG. Existing techniques primarily include triplets from knowledge sources while ignoring the substantial structural information beyond individual triplets.

The richness, diversity and coverage of commonsense knowledge in knowledge graphs are merged when they are consolidated, resulting in a heterogeneous commonsense knowledge source which has a higher impact on the downstream tasks. Zareian et al. [44] proposed Graph Bridging Network (GB-Net) that generates a scene graph, connects its entities and edges to the corresponding entities and edges in a commonsense graph retrieved from VG, WordNet, and ConceptNet, and uses GNN-based message propagation to recursively refine the scene graph relationships. Guo et al. [43] extracted relational and commonsense knowledge from VG and ConceptNet and encoded it in an Instance Relation Transformer (IRT) for SGG. These SGG techniques employed multiple knowledge sources, however, they have not been employed for visual reasoning tasks, which is important to evaluate the effectiveness of incorporating commonsense knowledge from multiple knowledge graphs. Some visual reasoning techniques, specifically VQA [52, 54], have directly used a subset [55] of DBpedia, ConceptNet, and WebChild, however, the powerful structured image representation of scene graphs is not utilized in these techniques. CSKG is the most recent, largest and systematically consolidated commonsense knowledge source. Ma et al. [56] used CSKG in language models and reported state-of-the-art performance in commonsense question answering by combining diverse relevant knowledge from CSKG and aligning it with the task. CSKG was employed for knowledge infusion in SGG [5] and the scene graphs were evaluated for downstream image synthesis, however, there is a significant need for investigation of heterogeneous commonsense knowledge-based scene graphs in the mainstream visual reasoning tasks, such as image captioning, VQA and image retrieval.

Moreover, some knowledge infusion methods leveraged knowledge graph embeddings, which are widely adopted in the vector representation of entities and relationships in knowledge graphs [57]. The knowledge graph embed-

dings 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. Knowledge graph embeddings have been used in several link prediction tasks including visual relationship detection [58], recommender systems [59] and SGG [5].

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) [60] 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 [15] 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 [30]. Commonsense knowledge-based scene graphs were leveraged in a scene graph-based image synthesis network that resulted in the generation of more realistic images [5]. Gu et al. [7] 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 [61]. For example, Zhang et al. [62] proposed encoding the structural information of scene graphs in GNNs to leverage it as the foundation for VQA. Similarly, Ziaeefard et al. [63] 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 [64]. 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. [65] 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.

3. Proposed Visual Understanding and Reasoning Framework

The proposed visual understanding and reasoning framework comprises three main components: (1) scene graph generation for image representation, (2) scene graph enrichment using a commonsense knowledge graph, and (3) downstream reasoning to leverage the enriched scene graphs for specific tasks, including image captioning. The proposed framework is illustrated in Figure 2 and each component of the framework is detailed as follows.

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 [66] for object detection. The ResNeXt-101-FPN CNN architecture [67] 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.

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

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 [68], 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 [39].

$$a = RoIAlign(I[b])$$
$$v = BiLSTM(a, I[b], l)$$

The combined pairwise object features $v_{ij} | i \neq j; i, j = 1, ..., n$ 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}.

$$v_{ij} = concat(BiLSTM(v_i), BiLSTM(v_j))$$
$$p_{ij} = embed(concat(l_i, l_j))$$
$$u_{ij} = conv(RoIAlign(F[b_i \cup b_j]))$$

In order to predict the relationship predicate labels r_{ij} and their confidence values c_{ij}, the three types of features, v_{ij}, p_{ij} and u_{ij}, of the object pairs are fused using a summation function [69] and used for softmax classification. Pairwise objects and relationship predicates are connected into a structured representation to create the scene graph S.

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

3.2. Scene Graph Enrichment

The scene graph representation is enriched via commonsense knowledge infusion for improved expressiveness and accuracy of the visual relationships in the scene graph. We used CSKG [25] for the enrichment of the scene graph with background knowledge and relevant facts in the form of triplets. To perform that, we utilized the graph embeddings to compute the similarity of nodes in the graph refinement and enrichment stages because similar entities often have similar vector representations in the embedding space. Algorithm 1 is used to refine the scene graph predictions by eliminating the predictions that are likely to be irrelevant or redundant. Multiple redundant predictions of the same object are indicated by the similarity of labels, overlapping of bounding boxes, and the structural patterns of their corresponding nodes in CSKG. At this stage, such prediction errors are 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 query CSKG with Knowledge Graph Toolkit (KGTK) [70] and pull new triplets that contain a subject or object node in the predicted scene graph. Any 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. The nodes of the new triplets that have reasonable structural similarities with the corresponding object nodes in the scene graph are connected to the object nodes by the edges of the new triplets. This way, the new triplets are added to the scene graph. In case the node of a new triplet exists in the scene graph already, the edge of that triplet is linked to the existing object node in the scene graph rather than creating a redundant node. For the enriched scene graphs to be evaluated for performance comparison or to be employed in downstream reasoning, the format of the enriched scene graph is fine-tuned in the post-processing stage in accordance with the original representation model of the scene graph. All predicates of the triplets in the VG subset of CSKG are expressed as "LocatedNear" edge type in CSKG, therefore, we substitute the predicates of the new triplets with VG as their source in CSKG with the most common predicate in the original VG dataset between the nodes of those triplets. For correct interpretation of the relationship predicates, this post-processing step makes use of some statistical prior knowledge from the relationships in VG. The process of complementing scene graphs

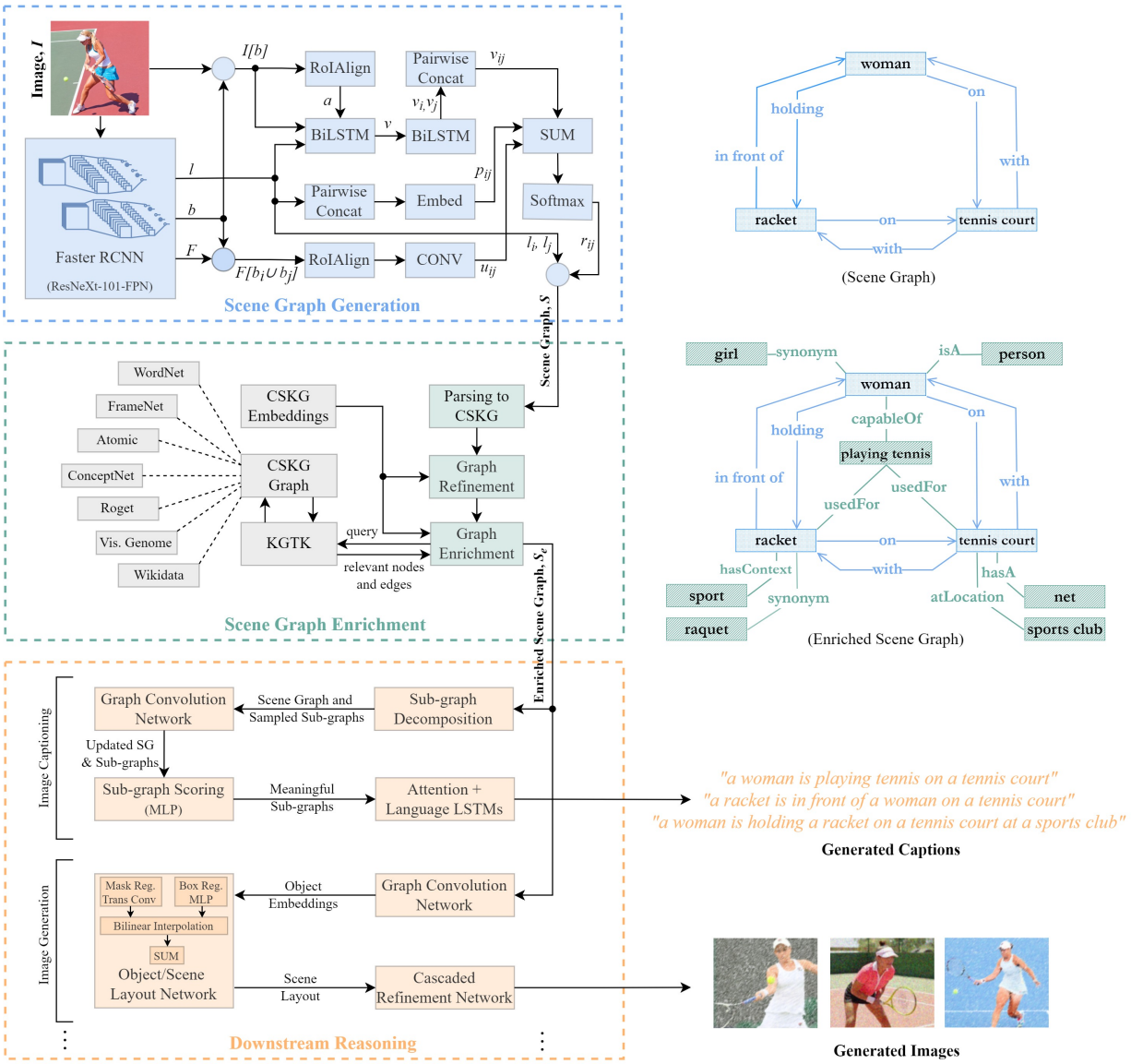


Fig. 2. Proposed Neural-Symbolic Visual Understanding and Reasoning Framework based on Enriched Scene Graph Representation

with common sense knowledge from CSKG is outlined in Algorithm 2. In both algorithms, the thresholds were set to 0.5 for the experimental analysis. These thresholds set the balance between the quantity and precision of visual relationships that are detected via SGG and infused via knowledge enrichment.

3.3. Downstream Reasoning

3.3.1. Caption Generation

We incorporated a scene graph-based image captioning network [29] in the proposed framework as a downstream task of scene graph generation and knowledge enrichment to utilize the enriched scene graphs for precise and expressive caption generation. The enriched scene graphs are decomposed and sub-graphs are sampled using neighbour *sampling* [71], i.e. a random set of seed nodes on the graph is selected and the immediate neighbours of the seed nodes with the edges in-between are taken as a sampled sub-graph. Similar sub-graphs are removed to avoid

Algorithm 1: Graph Refinement

Input: S, b
Output: S_r

```
S_r = []
for each triplet ∈ S do
    e_1 = cskg_emb(triplet[node1])
    e_2 = cskg_emb(triplet[node2])
    b_1 = b[triplet[node1]]
    b_2 = b[triplet[node2]]
    metric_sim = cosine_sim(e_1, e_2)
    metric_IoU = compute_IoU(b_1, b_2)
    if metric_sim ⩽ τ_sim ∧ metric_IoU ⩽ τ_iou then
        S_r.append(triplet)
```

Algorithm 2: Graph Enrichment

Input: S, G_{cskg}
Output: S_e

```
S_e = S
for each node ∈ S do
    e_1 = cskg_emb(node)
    triplets_cskg = query(G_cskg, node)
    triplets_cskg = preprocess(triplets_cskg)
    for each triplet ∈ triplets_cskg do
        if node == triplet[node1] then
            e_2 = cskg_emb(triplet[node2])
        else
            e_2 = cskg_emb(triplet[node1])
        s = cosine_sim(e_1, e_2)
        if s ⩾ τ ∧ triplet ∉ S_e then
            S_e.append(triplet)
    S_e = postprocess(S_e)
```

redundancy. Greedy non-maximal suppression is used to filter out sub-graphs with higher IoU scores of their nodes. The nodes and edges of the scene graph are augmented with their visual features and text embeddings using a *Graph Convolutional Network (GCN)*. The GCN updates the node and edge features by aggregating information from the neighbours via multiple graph convolutions and ReLU layers to integrate contextual information within the scene graph. The relationship information has been integrated using GCN, thus features of only 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 [72] followed by sigmoid normalization is used to rank sub-graphs. An attention-based *LSTM* is used to assign importance scores to the sub-graph nodes which is used by a language LSTM to generate sentences corresponding to each sub-graph. The enriched scene graph-based caption generation pipeline is illustrated in Figure 2.

3.3.2. Image Generation

We incorporated a scene graph-based image generation network [30] as a downstream task of scene graph generation and knowledge enrichment to utilize the enriched scene graphs for synthesizing realistic images from scene graphs. The labels of nodes and edges in the enriched scene graph are encoded into a vector representation using an embedding layer. A *GCN* processes the feature vectors and uses graph convolutions to aggregate nodes with their neighbourhood/structural information by propagating information along the edges. Since the relationship information has been integrated by the GCN, only the object encodings are further processed by the *layout network*. Within the layout network, a mask regression network uses transpose convolutions to predict a soft binary mask of each object, while a box regression network uses an MLP to predict a bounding box of each object. The object encoding is multiplied with the soft binary mask to get a masked embedding of the object, which is warped to the spatial location of the bounding box using bilinear interpolation, which generates an object layout. All the object layouts are then merged into a scene layout, which is processed by several cascaded refinement modules *Cascaded Refinement Network (CRN)* with increasing spatial resolution across the modules resulting in coarse to fine generation. The image generation network is adversarially trained against a patch-based discriminator that ensures that the generated images look realistic overall and an object discriminator that ensures that objects in the generated image look realistic. The enriched scene graph-based image generation pipeline is illustrated in Figure 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 library[1] for implementing the scene graph generation, image captioning and image generation methods and KGTK[2] for implementing the graph refinement and enrichment algorithms.

4.1.2. Datasets and Knowledge Source

We used the Visual Genome (VG) [14] and Microsoft COCO [26] 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 [73] of VG contains the most frequent 50 predicate classes and 150 object classes, which was used for training Faster RCNN, SGG pipeline and the image generation network. 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 [74] of COCO for the evaluation of enriched scene graph-based image captioning. The standard split comprises 5K images each for validation and testing, and the rest for training.

We used the pre-trained CSKG embeddings [25] 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*) [75] 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

[1] https://pytorch.org/
[2] https://kgtk.readthedocs.io/

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 [27]. The confidence score is taken into account, therefore $R@K$ requires 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 [28, 38].

BLEU score [76] and **METEOR** score [77] 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 [78] 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 [79] 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 [80] 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 Faster RCNN model was trained on images and ground truth annotations of objects in the dataset using Stochastic 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. On the test set, the Faster RCNN model achieved 29.19 mAP (with a 0.5 IoU threshold), 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 and 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 Figure 4. Similar trend is observed in 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

in reducing the number of missed or incorrect predictions made during scene graph construction and increases the recall rate for relationship prediction.

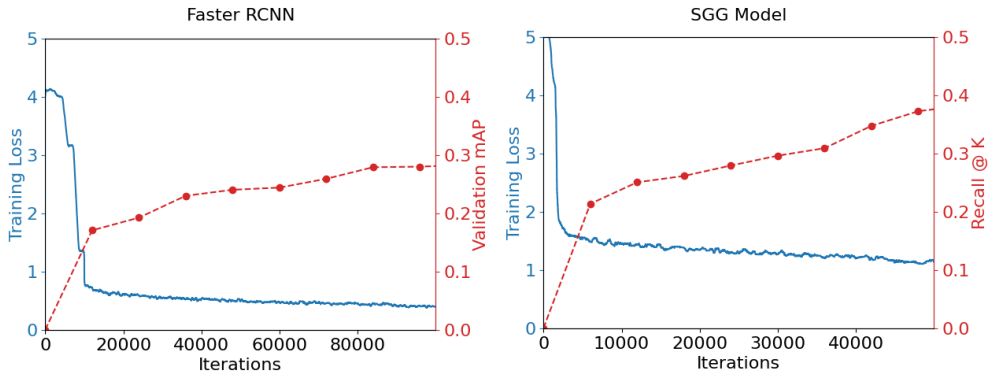


Fig. 3. Training progress plots along with periodic validation checks of the Faster RCNN and SGG models.

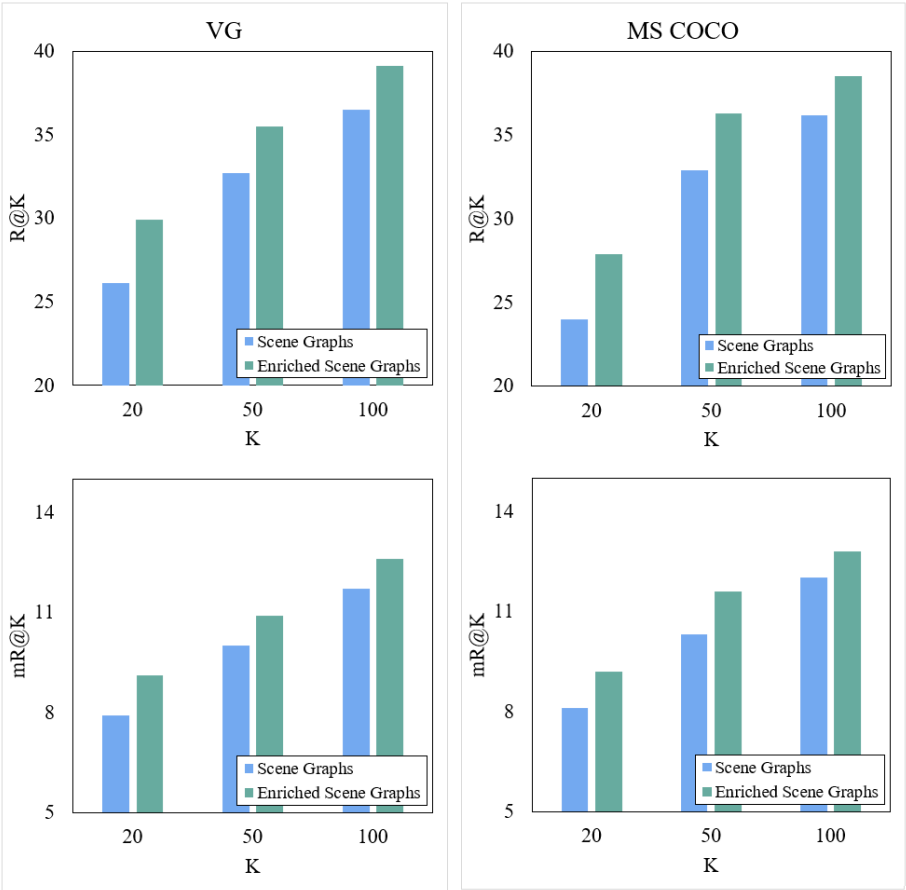


Fig. 4. Comparison of the recall metrics computed using the conventional scene graphs and proposed enriched scene graphs.

Table 2

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

	SGG Method	Approach	Commonsense Knowledge Source	SGG Performance		Downstream Reasoning Task
				R@20/50/100 (%)	mR@20/50/100 (%)	
SGG using Knowledge Infusion	**Proposed (SGG with Enrichment)**	Multi-modal DNN-based SGG and enrichment using CSKG	Heterogeneous KG: CSKG [25]	**29.9/35.5/39.1**	**9.1/10.9/12.6**	Image Captioning, Image Generation
	DSGAN [40]	Deep sparse graph attention network	Sparse KG & Statistical Prior	23.2/28.8/32.9	7.8/8.9/11.8	-
	IRT-MSK [43]	Instance Relation Transformer with Multiple Structured Knowledge	Multiple KGs: ConceptNet [23], VG [14]	21.9/27.8/31.0	-/-/-	-
	MOTIFS [39]	RNN-LSTM based Stacked Motif Networks	Statistical Prior	21.4/27.2/30.3	4.2/5.7/6.6	-
	GB-Net [44]	Message passing between scene graphs and commonsense graph	Multiple KGs: ConceptNet [23], WordNet [24], VG [14]	-/26.4/30.0	-/6.1/7.3	-
	KERN [38]	Knowledge-embedded routing network	Statistical Prior	22.3/27.1/29.8	-/6.4/7.3	-
	COACHER [42]	Zero-shot relationship prediction via commonsense infusion	KG: ConceptNet [23]	13.4/19.3/22.2	-/-/-	-
	KB-GAN [7]	Commonsense and reconstruction-based object and phrase refinement	KG: ConceptNet [23]	-/13.6/17.6	-/-/-	Image Generation
	DeepVRL [41]	Deep Q-network for variation-structured reinforcement learning	Language Prior	-/13.3/12.6	-/-/-	-
	VRD Model [27]	Relationship prediction using semantic word embeddings	Language Prior	-/0.3/0.5	-/-/-	Image Retrieval
Conventional SGG	HL-Net [16]	Transformer & MP based heterophily learning network	-	26.0/**33.7/38.1**	-/-/9.2	-
	Unbiased SGG [20]	Causal inference and total direct effect	-	25.8/33.3/37.8	6.9/9.3/11.1	-
	SMP [15]	Visual Relation Saliency-guided Message Passing	-	-/32.6/36.9	-/-/-	Image Captioning
	Proposed (SGG Only)	Scene graph generation based on fusion of visual (region and object) and text features	-	**26.1**/32.7/36.5	**7.9**/10.0/11.7	Image Captioning, Image Generation
	NICEST [81]	Noisy label correction and training for Robust SGG	-	-/29.0/32.7	-/**10.4/12.4**	-
	VCTree [28]	Dynamic tree structures and Bi-directional TreeLSTM	-	22.0/27.9/31.3	5.2/6.9/8.0	Visual Question Answering
	IMP+ [73]	Object and relationship feature refinement via message passing	-	14.6/20.7/24.5	-/3.8/4.8	-
	Factorizable Net [82]	Clustering-based graph factorization	-	-/13.1/16.5	-/-/-	-
	MSDN [83]	Scene description at object, phrase and caption levels	-	-/10.7/14.2	-/-/-	-
	Graph RCNN [84]	RPN followed by Attention GCN	-	-/11.4/13.7	-/-/-	-

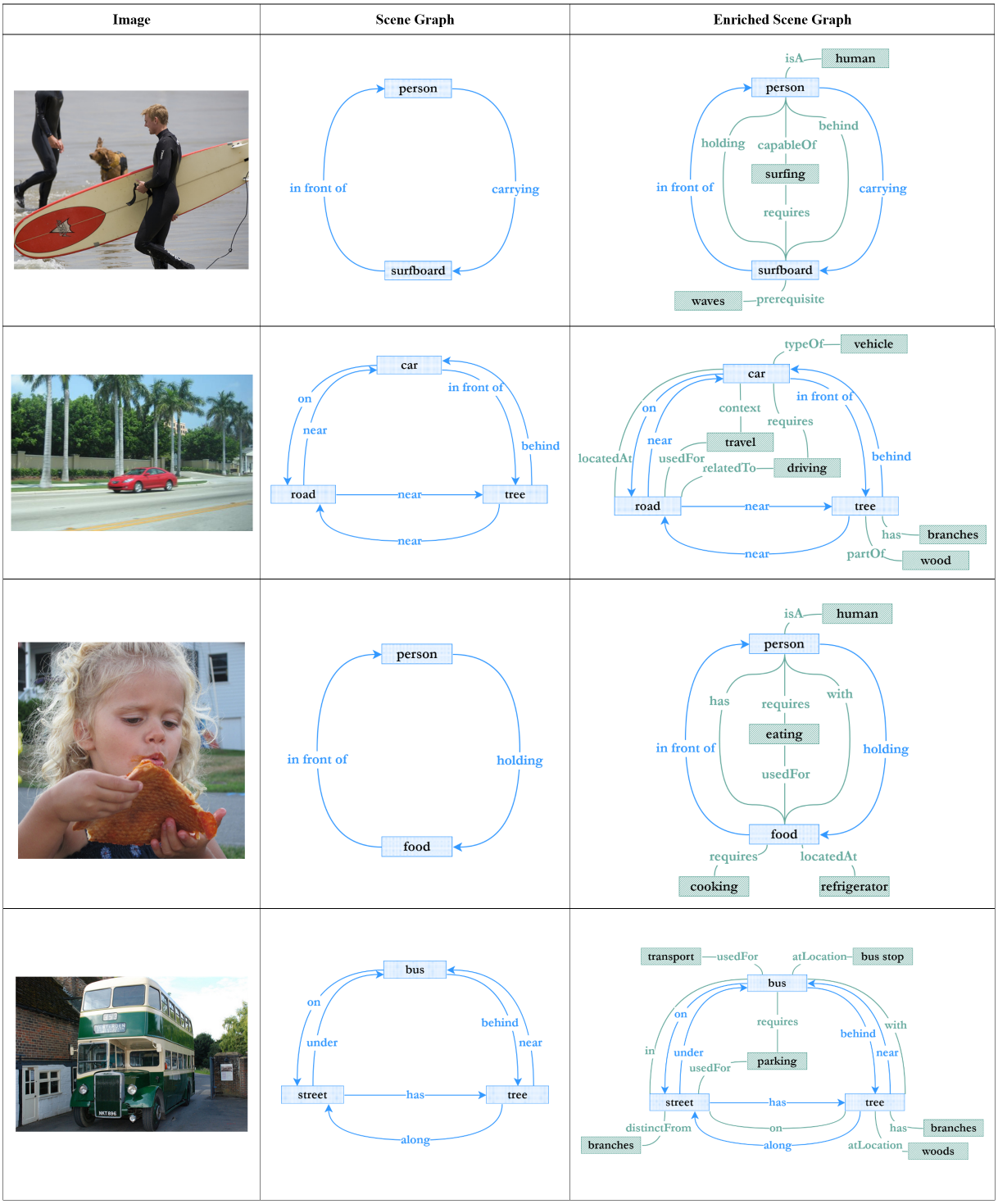


Fig. 5. Examples of the proposed enriched scene graphs for visual understanding and reasoning (VG images).

4.2.3. Benchmark Comparison

Table 2 presents a detailed comparison of the proposed enriched scene graph-based approach with the state-of-the-art SGG techniques. The proposed method obtained a considerable higher recall score on the benchmark VG dataset, outperforming the state-of-the-art techniques. The proposed method achieved $R@K = 29.9, 35.5, 39.1$ for $K = 20, 50, 100$, while the latest technique [40] among the commonsense 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 [16] has a recall score of $R@K = 26.0, 33.7, 38.1$. The superior performance of the proposed approach depicts the effectiveness of incorporating the most recent, largest, and diverse heterogenous knowledge graph as a commonsense knowledge source in SGG.

4.2.4. Qualitative Results

Figure 5 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 commonsense relationships about object interactions, such as *(person, holding, surfboard)* in the first row in Figure 5, and spatial placement, such as *(tree, on, street)* in the last row of Figure 5. In the third row in Figure 5, *(person, requires, eating)* and *(food, usedFor, eating)* represent useful background facts extracted from CSKG.

4.2.5. Downstream Reasoning: Image Captioning

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 the 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 standard evaluation metrics, including SPICE, BLEU, CIDEr, ROGUE and METEOR, which is presented in Figure 6. 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 3 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. Some qualitative results of caption generation using conventional and enriched scene graphs are shown in 7. The promising results show that enriched scene graphs result in more expressive and meaningful captions as compared to conventional scene graphs.

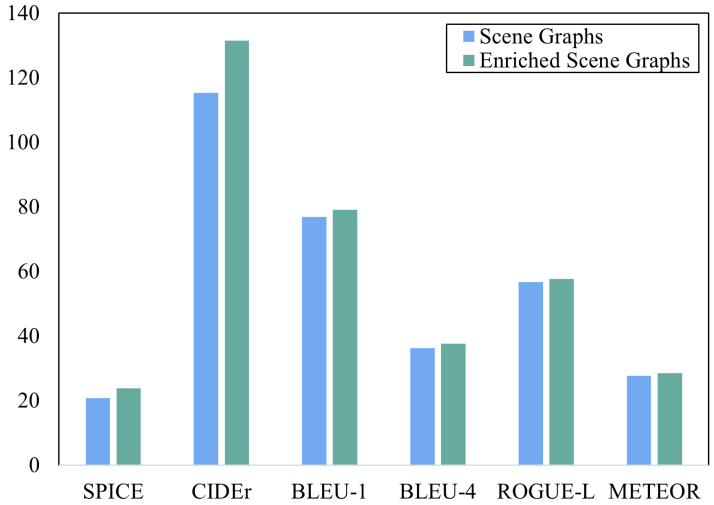


Fig. 6. Comparison of image captioning using conventional scene graphs and proposed enriched scene graphs in terms of the standard evaluation metrics. Enriched scene graphs resulted in higher SPICE and CIDEr scores and comparable performance in terms of BLEU, ROGUE and METEOR scores.

Table 3

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 [74] of the COCO dataset [26]. 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. [85]	22.4	129.6	**81.0**	**38.8**	**58.8**	**28.8**
Yang et al. [86]	20.9	116.3	77.3	36.8	57.0	27.9
Yao et al. [87]	20.9	116.7	77.6	36.9	57.2	27.7
Zhong et al. [29]	20.7	115.3	76.8	36.2	56.6	27.7
Ke et al. [88]	20.5	115.3	77.5	36.8	56.8	27.2
Anderson et al. [80]	20.3	113.5	77.2	36.2	56.4	27.0
Nguyen et al. [89]	19.8	106.6	–	32.6	55.0	26.4

4.2.6. Downstream Reasoning: Image Generation

We trained the scene graph to image generation network on the Visual Genome subset that was used to train the SGG pipeline. The trained network was used to generate images using conventional scene graphs as well as enriched scene graphs. The results of image generation from scene graphs are presented in Figure 8. The semantic concepts in the input scene graph can be more clearly observed in the images generated using enriched scene graphs. In the first example in Figure 8, the scene graph objects person, helmet and bike are more clearly visible in the image generated using enriched scene graphs. Similarly, the objects {house, door}, {person, beach, sky}, and {person, surfboard} can be more clearly seen in the subsequent examples in Figure 8. The enriched scene graphs resulted in more realistic images of higher quality as compared to the conventional scene graphs, which depicts the efficacy of the proposed scene graph enrichment for image generation. Our evaluation of image generation is based on qualitative analysis, as the automatic quantitative evaluation metrics do not provide a good measure of image quality and human judgement according to the baseline technique [30].

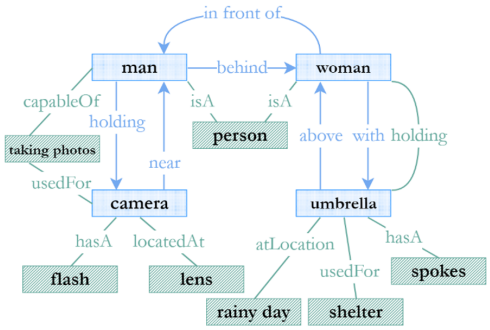
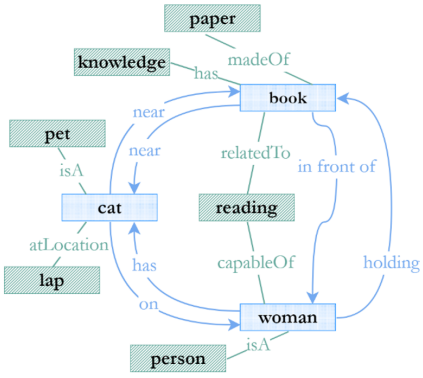
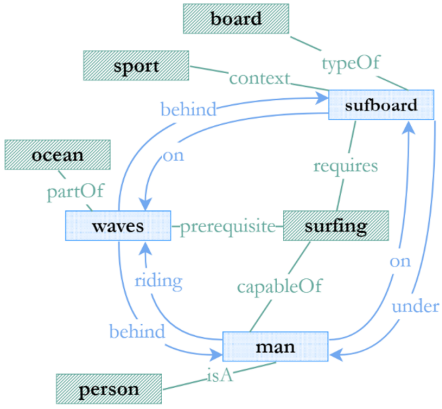
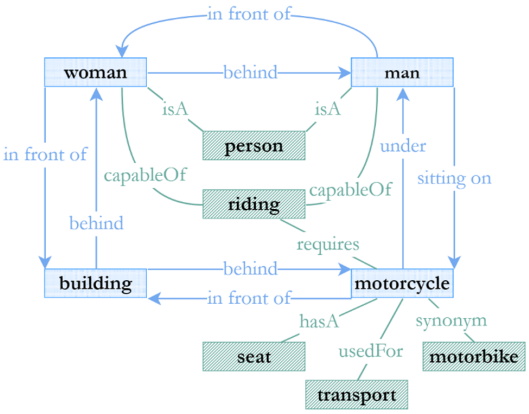
Image	Enriched Scene Graph	Caption
		a man is holding a camera standing next to a woman with an umbrella a man taking a picture of a woman holding an umbrella in the rain
		a woman is holding a book near a cat a woman is reading a book while a cat sits on her lap
		a man is riding waves on a surfboard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

Fig. 7. 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)

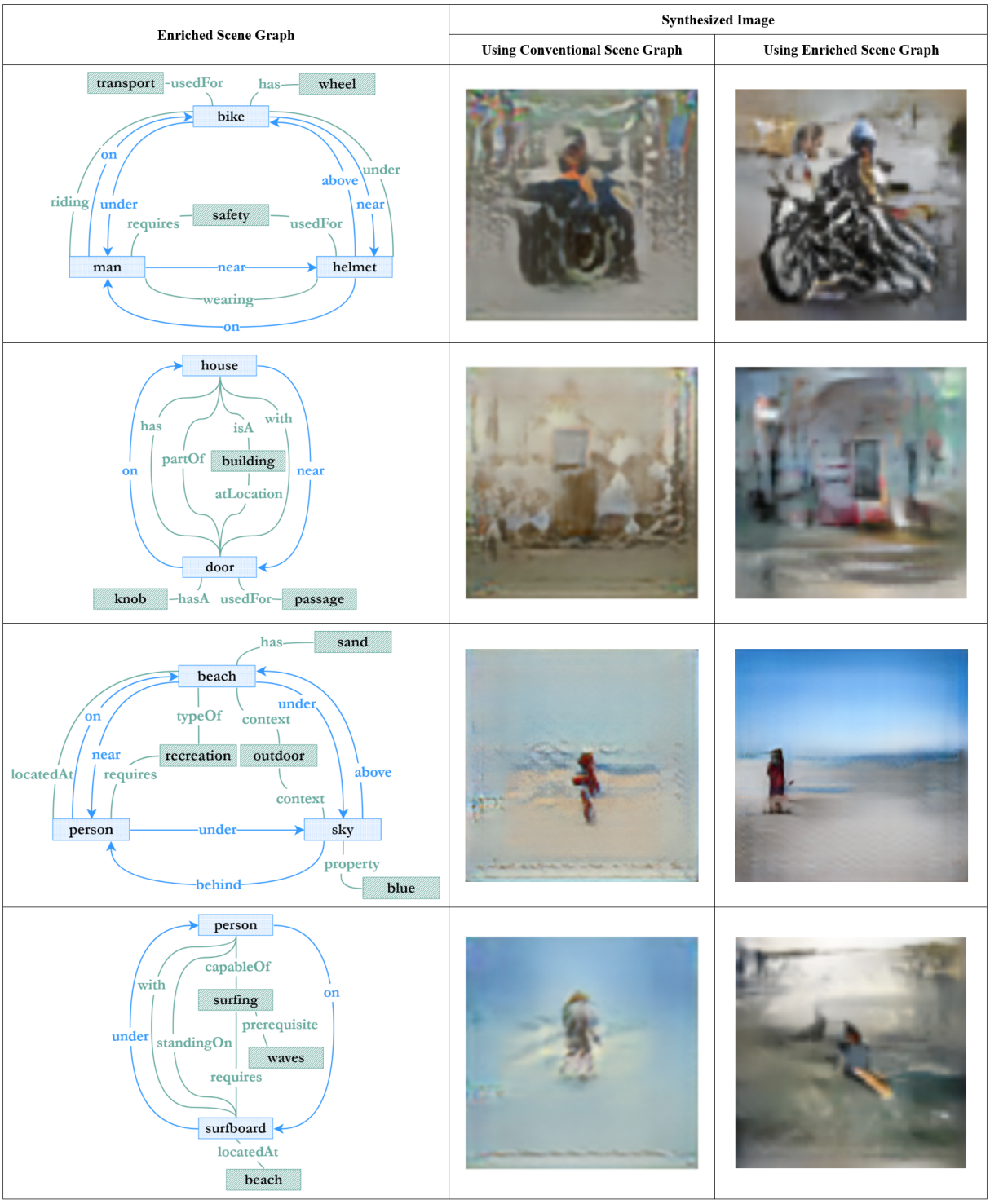


Fig. 8. Results of scene graph to image generation using conventional scene graphs (left) and enriched scene graphs (right). Enriched scene graphs resulted in more realistic images, in which the objects in scene graphs can be more clearly observed.

5. Conclusion

The 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 knowledge graphs is a promising approach toward alleviating the existing challenges in SGG and improving the expressiveness of visual understanding and reasoning frameworks. We proposed a neural-symbolic visual understanding and reasoning framework based on enriched scene graphs. A DNNs cascade is used to generate symbolic scene graphs, which is followed by rule-based graph refinement and enrichment using commonsense knowledge extracted from a heterogeneous knowledge graph in the form of related facts and background information about the scene graph elements. We integrated image captioning and image generation models in the proposed framework as downstream tasks of scene graph enrichment. The evaluation results showed that commonsense 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 dataset. 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 knowledge graphs. Moreover, the enriched scene graphs. The use of enriched scene graphs resulted in an increase in SPICE and CIDEr scores obtained by the image captioning model from 20.7 and 115.3 to 23.8 and 131.4 respectively, and 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 qualitative results of image generation showed that the enriched scene graphs result in generating more realistic images in which the semantic concepts in the input scene graph can be more clearly observed. The future work will focus on the integration of more downstream reasoning tasks including image retrieval, VQA and visual storytelling etc. as well as the use of graph neural networks, message passing and multi-hop relationship reasoning approaches to further improve visual relationship prediction performance in SGG.

References

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

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

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

[4] A. Krizhevsky, I. Sutskever and G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems* **25** (2012).

[5] 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.

[6] 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.

[7] 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.

[8] 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).

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen and A. Hauptmann, Scene graphs: A survey of generations and applications, *arXiv preprint arXiv:2104.01111* (2021).

[14] 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.

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

[16] 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.

[17] T. He, L. Gao, J. Song, J. Cai and Y.-F. Li, Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation, *arXiv preprint arXiv:2006.07585* (2020).

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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 arXiv:1504.00325* (2015).

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] M.A. Sadeghi and A. Farhadi, Recognition using visual phrases, in: *CVPR 2011*, IEEE, 2011, pp. 1745–1752.

[33] 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.

[34] 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.

[35] J. McCarthy et al., *Programs with common sense*, RLE and MIT computation center, 1960.

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85.

[46] 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.

[47] B. Kipfer, Roget's 21st Century Thesaurus in Dictionary Form (éd. 3). New York: The Philip Lief Group, Inc, 2005.

[48] 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, Volume 1*, 1998, pp. 86–90.
[49] 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.
[50] 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.
[51] 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.
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] 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.
[57] M. Palmonari and P. Minervini, Knowledge graph embeddings and explainable AI, *Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges, IOS Press,, Amsterdam* (2020), 49–72.
[58] 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.
[59] 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.
[60] 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.
[61] 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.
[62] C. Zhang, W.-L. Chao and D. Xuan, An empirical study on leveraging scene graphs for visual question answering, *arXiv preprint arXiv:1907.12133* (2019).
[63] M. Ziaeefard and F. Lécué, Towards Knowledge-Augmented Visual Question Answering, in: *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1863–1873.
[64] M.J. Khan and E. Curry, Neuro-symbolic Visual Reasoning for Multimedia Event Processing: Overview, Prospects and Challenges., in: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM'2020) Workshops*, 2020.
[65] 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.
[66] 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.
[67] 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.
[68] 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.
[69] 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.
[70] 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.
[71] J.M. Klusowski and Y. Wu, Counting motifs with graph sampling, in: *Conference On Learning Theory*, PMLR, 2018, pp. 1966–2011.
[72] K. Xu, W. Hu, J. Leskovec and S. Jegelka, How powerful are graph neural networks?, *arXiv preprint arXiv:1810.00826* (2018).
[73] 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.
[74] 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.
[75] 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.
[76] 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.
[77] 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.
[78] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: *Text summarization branches out*, 2004, pp. 74–81.

[79] 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.

[80] 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.

[81] 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 arXiv:2207.13316* (2022).

[82] 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.

[83] 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.

[84] 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.

[85] 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).

[86] 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.

[87] 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.

[88] 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.

[89] 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.