Expressive Scene Graph Generation using Commonsense Knowledge Infusion for Visual Understanding and Reasoning *

Muhammad Jaleed Khan[1 [0000−0003−4727−4722]],
John G. Breslin[1,2 [0000−0001−5790−050X]], and
Edward Curry[1,2 [0000−0001−8236−6433]]

[1]SFI Centre for Research Training in Artificial Intelligence
[2]Insight SFI Research Centre for Data Analytics
Data Science Institute, National University of Ireland, Galway.
{m.khan12, john.breslin, edward.curry}@nuigalway.ie

Abstract. Scene graph generation aims to capture the semantic elements in images by modelling objects and their relationships in a structured manner, which are essential for visual understanding and reasoning tasks including image captioning, visual question answering, multimedia event processing, visual storytelling and image retrieval. The existing scene graph generation approaches provide limited performance and expressiveness for higher-level visual understanding and reasoning. This challenge can be mitigated by leveraging commonsense knowledge, such as related facts and background knowledge, about the semantic elements in scene graphs. In this paper, we propose the infusion of diverse commonsense knowledge about the semantic elements in scene graphs to generate rich and expressive scene graphs using a heterogeneous knowledge source that contains commonsense knowledge consolidated from seven different knowledge bases. The graph embeddings of the object nodes are used to leverage their structural patterns in the knowledge source to compute similarity metrics for graph refinement and enrichment. We performed experimental and comparative analysis on the benchmark Visual Genome dataset, in which the proposed method achieved a higher recall rate ($R@K = 29.89, 35.4, 39.12$ for $K = 20, 50, 100$) as compared to the existing state-of-the-art technique ($R@K = 25.8, 33.3, 37.8$ for $K = 20, 50, 100$). The qualitative results of the proposed method in a downstream task of image generation showed that more realistic images are generated using the commonsense knowledge-based scene graphs. These results depict the effectiveness of commonsense knowledge infusion in improving the performance and expressiveness of scene graph generation for visual understanding and reasoning tasks.

Keywords: scene graph · image representation · commonsense knowledge · visual reasoning · image generation

* 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.

1 Introduction

During the past few years, recent advances in deep learning techniques and multimodal approaches have helped in solving several challenging problems in visual understanding tasks including object detection [57] and visual relationship detection [14,32,35]. Numerous efforts have been made to effectively capture and describe the image features and object relationships in a structured and explicit way. In this direction, Scene Graph Generation (SGG) [46,48,3] has attracted significant attention due to its capability to capture the detailed semantics of visual scenes by modelling objects and their relationships in a structured manner. Graph-based structured image representations like scene graphs are used in a wide range of visual understanding tasks including image reconstruction [11], image captioning [61], Visual Question Answering (VQA) [22,25], image retrieval [55], visual storytelling [54] and multimedia event processing [5,20]. The performance of SGG is compromised by challenges including bias and annotation issues in crowd-sourced datasets [23,7]. Several efforts have been made by researchers in this field to address these challenges by making use of state-of-the-art approaches, such as counterfactual analysis [48], self-supervised learning [40] and linguistic supervision [62]. However, there is still a need for significant improvement in the expressiveness, accuracy and robustness of SGG methods.

In addition to the objects and their relationships in scene graphs, higher-level visual reasoning for the downstream tasks mentioned in the last paragraph requires background information about the scene and its constituents to mimic the cognitive ability of humans to use commonsense reasoning. Leveraging and reasoning with commonsense knowledge is quite challenging because of its implicit nature; it is universally accepted and used by humans in everyday situations but generally disregarded when we speak or write. Most of the existing SGG methods use datasets that contain large collections of images along with annotations of objects, attributes, relationships, scene graphs, etc., such as, Visual Genome (VG) [23] and VRD [31]. These datasets have limited or no explicit commonsense knowledge, which limits the expressiveness of scene graphs and the higher-level reasoning capabilities in the downstream tasks unless commonsense knowledge is infused from external sources. There are several publicly available sources [50,44,43,21] that include different forms and notions of commonsense knowledge. Some consolidation efforts [17,9] have been made to unify the different sources into a global commonsense knowledge source to jointly exploit their diverse knowledge and coverage. These consolidated sources have been integrated and used in language processing methods [33,58] for improving their robustness and expressiveness. The consolidated commonsense knowledge sources have not been leveraged for visual understanding and reasoning yet, however, their capability to provide rich and diverse background information and relevant facts about the concepts in a scene can help in improving the performance of SGG and providing rich and expressive scene representations for downstream reasoning.

Figure 1 shows a motivating example of an image and its commonsense knowledge-based scene graph representation. The scene graph of the image contains the relationship triplets *(woman, holding, racket)* and *(woman, on, ten-*

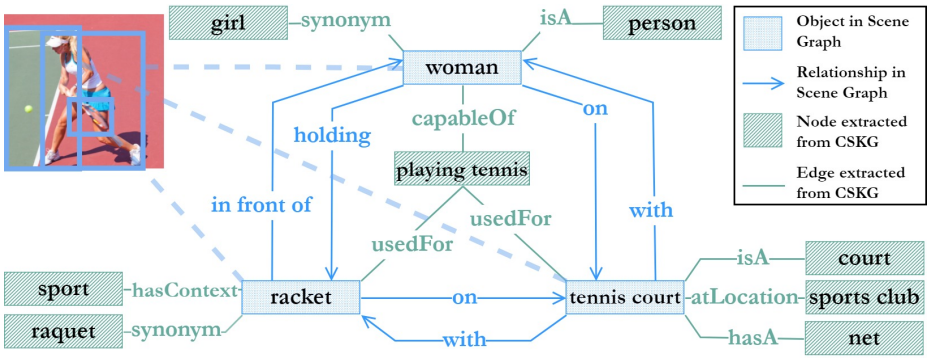

Fig. 1. A motivating example of a scene graph of an image with commonsense knowledge infusion using CommonSense Knowledge Graph (CSKG). The scene graph (blue) provides information about the objects and their pairwise relationships in the scene. The relevant nodes and edges extracted from CSKG (green) complement and enrich the scene graph by providing the necessary information about the possible spatial proximity of objects relative to each other and any possible interactions between objects, i.e. *(woman, at, tennis court)* and *(woman, holding, racket)*, and more importantly the background information and related facts, i.e. *(woman, capableOf, playing_tennis)* and *(racket, usedFor, playing_tennis)*, which allows higher-level reasoning to deduce *"the woman is playing tennis"*.

nis_court) representing the objects and their pairwise interactions. Though it is easy and straightforward for us to infer that the woman is playing tennis, it is challenging for machines to infer that without some external commonsense knowledge. The relevant nodes and edges extracted from the CommonSense Knowledge Graph (CSKG) [17] including *(woman, capableOf, playing_tennis)* and *(racket, usedFor, playing_tennis)* provide the necessary background information and facts for higher level reasoning. In this paper, we propose a commonsense knowledge-based SGG method that generates scene graph of an image and infuses the background knowledge and relevant facts about the concepts in the scene graph from CSKG [17], which is a large consolidated commonsense knowledge source. Graph embeddings were leveraged to compute the similarity of object nodes in the graph refinement and enrichment steps because similar entities tend to have similar vector representations in the embedding space [38]. The commonsense knowledge complements and enriches the scene graph relationships, which improves the performance of SGG and the expressiveness of scene graph representations. We evaluated the proposed method on the benchmark VG dataset and noted improvement of relationship prediction results for SGG. The encouraging experimental results depict the potential of commonsense knowledge in scene graph generation and its promising applications in visual understanding and reasoning. The main contributions of this paper include:

1. We propose a commonsense knowledge-based scene graph generation approach, which extracts background knowledge and relevant facts from com-

monsense knowledge sources based on graph embeddings and integrates them in the scene graphs to generate rich and expressive scene graph representations of images. We employed a heterogeneous knowledge graph [17], containing rich commonsense knowledge consolidated from seven diverse sources, which has not been investigated for visual understanding and reasoning yet.
2. We performed experimental and comparative analysis (shown in Figure 4, Figure 5 and Table 2) on the benchmark Visual Genome dataset using the standard metric, and showed that the proposed method achieved a higher recall rate ($R@K = 29.89, 35.4, 39.12$ for $K = 20, 50, 100$) as compared to the existing state-of-the-art technique ($R@K = 25.8, 33.3, 37.8$ for $K = 20, 50, 100$).
3. We employed image generation as a downstream task of scene graph generation and showed improved results of image generation from scene graphs after commonsense knowledge infusion as shown in Figure 6.

2 Related Work

2.1 Scene Graph Generation

Scene graph generation (SGG) is a challenging research problem and is actively investigated by researchers in computer vision. In the compositional methods, the subject, predicate and object are separately detected and aggregated later. Li et al. [26] used detected objects in an image to generate separate region proposals for subject, predicate and object; these region proposals are aggregated with features from a deep neural network (DNN) to reach a triplet prediction. Such methods are scalable, but they have very limited performance in the case of rare or unseen relations. The visual phrase models for visual relation detection treat relation triplets as a single entity. Sadeghi et al. [42] employed DNNs to predict objects as well as visual phrase or triplets and then refined those predictions by comparing them to other predictions in the image. Deep relational networks are also used for visual relation detection, in which the DNN also leverages the statistical dependency among objects and predicates [6]. The visual phrase models are less sensitive to the diversity of visual relations as compared to the compositional models, but they require a greater number of training examples in datasets with a large vocabulary of objects and predicates.

The more recent scene graph generation and visual relationship detection methods fuse visual and semantic embeddings in DNNs to detect visual relations on a large scale. Zhang et al. [67] extract visual features in three branches each for the subject, predicate and object, with the predicate branch fusing its features with the subject and object features at a later stage to leverage the interactions between subject and object for relation detection. During learning, features extracted from the text space are also embedded as labelling for the visual features. In a similar approach with improved precision, Peyre et al. [39] add a visual phrase embedding space during learning to enable analogical reasoning for predicting unseen relations and to improve robustness to appearance variations of visual relations. Tang et al. [48] attempted to address the problem

of bias in SGG models due to the unbalanced distribution of relationships in datasets by leveraging causal inference and total direct effect.

Most of the existing works focus on visual and linguistic patterns in images while neglecting the background information and related facts about concepts in images and the structural patterns of scene graph elements in commonsense knowledge graphs, which have significant potential in understanding and interpretation of visual concepts. Only a few recent works mentioned in the next subsection explicitly leverage commonsense knowledge graphs for visual understanding and reasoning.

Table 1. Commonsense Knowledge Sources

Knowledge Source	**Knowledge Type**	**Size**	**Example**
ConceptNet [44]	Text-based knowledge about everyday objects, activities, relations, etc.	8M nodes, 36 relations & 21M edges	(chair, used for, sitting)
Wikidata [50]	General taxonomic knowledge about instances, concepts, relations etc.	75M objects, 1200 relations & 900M edges	(eating, subclass of, ingestion)
ATOMIC [43]	Procedural knowledge about pre/post conditions of events	0.3M nodes, 9 relations & 0.877M edges	(PersonX eating dinner, xEffect, satisfies hunger)
Roget [21]	Lexical knowledge about words, relations, etc.	72k words, 2 relations & 1.4M edges	(motorcycle, synonym, bike)
FrameNet [2]	Lexical knowledge about frames, roles, relations, etc.	1.2k frames, 12k roles, 1.9k edges & 13k lexical units	(cooking creation, has frame element, produced food)
Wordnet [36]	Lexical knowledge about words, concepts, relations, etc.	0.155M words, 10 relations & 0.176M synsets	(car, has part, air bag)
Visual Genome [23]	Visual knowledge about objects, relations and attributes in images	108k images, 3.8M nodes, 42k relations, 2.3M edges & 2.8M attributes	(food, on, plate), (woman, looking at, sandwich)
CSKG [17]	Consolidated commonsense knowledge from the above seven sources	2.16M nodes, 58 relations, 6M edges	(racket, used for, playing tennis)

2.2 Commonsense Knowledge Sources and Infusion

The acquisition and representation of commonsense knowledge and reasoning with it have been one of the major challenges in artificial intelligence since the 1960s [34], which has led the research community to develop and curate several knowledge sources containing commonsense knowledge in different forms and contexts [16]. Some of the popular sources of commonsense knowledge along with their details are presented in Table 1. Some of these sources, especially ConceptNet [44], have been used in a few visual understanding and reasoning techniques. These techniques either extract relevant facts from a source and embed them in the model at a certain stage [11,37,45,66], or use graph-based message passing to embed the structural information from the source in the representations of the model [64,4,56,24]. Chen et al. [4] and Zellers et al. [66] incorporated commonsense knowledge from dataset statistics by employing pre-computed frequency priors in their predicate classification models to improve the performance of SGG. Wan et al. [51] proposed the use of a commonsense knowledge graph along with the visual features to enhance predicate detection for detected objects in visual relation detection. Gu et al. [11] retrieve relevant facts from a single source, i.e. ConceptNet [44] for each object, encode the facts into its features using recurrent neural networks and an attention mechanism in SGG. Kan et al. [19] infused commonsense knowledge from ConceptNet for zero-shot relationship prediction in SGG. The existing approaches mostly infuse triplets from the knowledge sources and ignore the rich structural information beyond individual triplets.

The knowledge sources are rich and diverse and cover different domains and contexts of commonsense knowledge, which can be consolidated to provide a rich and heterogeneous source of commonsense knowledge and to increase its impact in the downstream reasoning tasks. Zareian et al. [63] proposed GB-Net, which links the entities and edges in a scene graph to the corresponding entities and edges in a commonsense graph extracted from VG, WordNet and ConceptNet, and iteratively refine the scene graph using graph neural network-based message passing. Guo et al. [12] employed an instance relation transformer to extract relational and commonsense knowledge from VG and ConceptNet for SGG. These are the only SGG approaches that leverage multiple knowledge sources, while a subset [53] of DBpedia, ConceptNet and WebChild containing knowledge about visual concepts has been used in VQA [56,30]. The CommonSense Knowledge Graph (CSKG) [17] is currently the latest and largest consolidated source that integrates commonsense knowledge from the seven diverse and disjoint sources, including ConceptNet [44], Wikidata [50], ATOMIC [43], VG [23], Wordnet [36], Roget [21] and FrameNet [2]. Ma et al. [33] employed CSKG in language models and achieved the best performance in commonsense question answering by utilizing the diverse relevant knowledge from CSKG and aligning the knowledge with the task. To the best of our knowledge, the use and potential of CSKG have not yet been explored for visual understanding and reasoning tasks.

The knowledge-infusion methods also leverage knowledge graph embeddings, which are widely adopted in the vector representation of entities and relation-

ships in knowledge graphs [38]. The knowledge graph 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. Knowledge graph embeddings have been used in several link prediction tasks including visual relationship detection [1] and recommender systems [52].

3 Proposed Method

The proposed commonsense knowledge-based scene graph generation method employs a DNN-based approach for detecting objects and their pairwise relationships in an image to generate its scene graph, which is followed by commonsense knowledge infusion using CSKG [17] for the enrichment of scene graph with background knowledge and relevant facts in the form of triplets. Figure 2 provides a detailed overview of the proposed method. The proposed method is built on the SGG toolkit [47].

Following the trend in recent SGG methods [59,66,49,48], we use Faster RCNN [41] for detecting objects in the images. We use ResNeXt-101-FPN architecture [29] as the backbone CNN for Faster RCNN. The Faster RCNN takes an image I as input and provides the object bounding boxes b and object class labels l of the n detected objects. The feature maps F are also extracted from the underlying CNN in the Faster RCNN.

$$\{b, l, F\} = FasterRCNN(I) \tag{1}$$

After detecting the objects and extracting the feature maps, the relationships between object pairs are predicted. RoIAlign [13] is applied to the image regions $I[b]$, which provides the region features a of each detected object.

$$a = RoIAlign(I[b]) \tag{2}$$

For all n objects, Bi-directional Long Short Term Memory (Bi-LSTM) layers [66] are used to encode a, $I[b]$ and l as the individual visual context features v_i.

$$v = BiLSTM(a, I[b], l) \tag{3}$$

The individual visual context features of objects are encoded by another set of Bi-LSTM layers and concatenated into combined pairwise object features $v_{ij} | i \neq j; i, j = 1, ..., n$.

$$v_{ij} = concat(BiLSTM(v_i), BiLSTM(v_j)) \tag{4}$$

In the same way, the pairwise object labels (l_i, l_j) are encoded through an embedding layer to compute the language prior p_{ij}. The contextual union features u_{ij} are extracted by applying RoIAlign to the union regions of pairwise objects in F.

$$u_{ij} = conv(RoIAlign(F[b_i \cup b_j])) \tag{5}$$

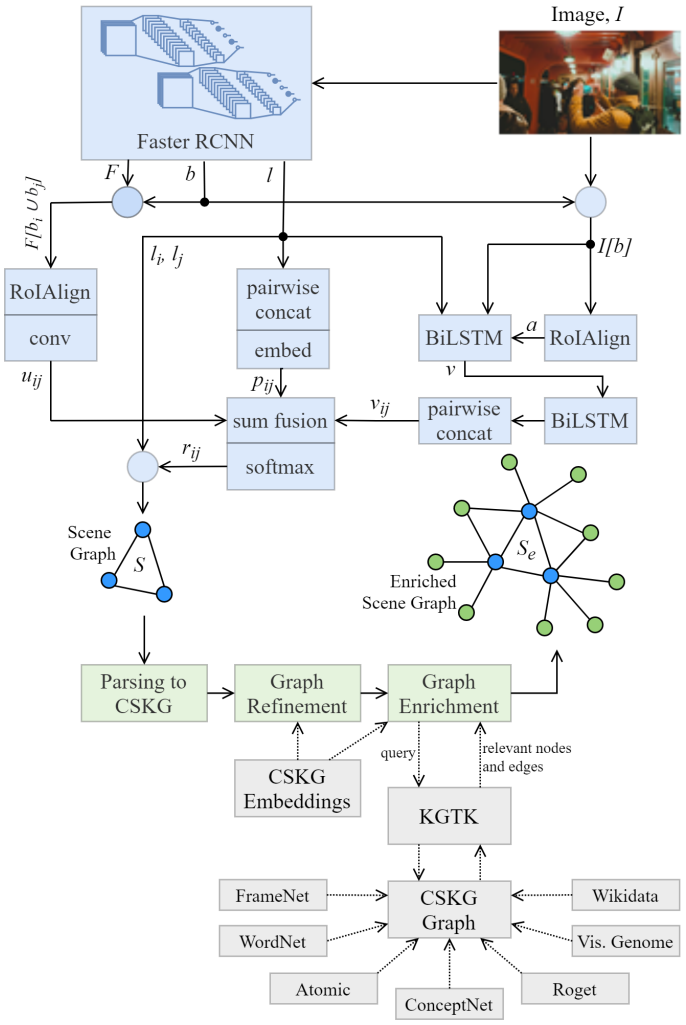

Fig. 2. The proposed commonsense knowledge-based scene graph generation method

Finally, all the three types of features representing the object pairs are fused using a summation feature fusion function [8] followed by a softmax function to predict the relationship class labels r_{ij} and the relationship class probabilities c_{ij}.

$$\{r_{ij}, c_{ij}\} = softmax(SUM(v_{ij}, u_{ij}, p_{ij})) \tag{6}$$

The scene graph S is formed by linking the pairwise objects and relationships into a graph structure.

$$S = \{l_i, r_{ij}, l_j\} \tag{7}$$

Algorithm 1: Graph refinement

Input: S, b
Output: S_r
$S_r = []$
for *each triplet* $\in S$ **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 $metric_{sim} \leqslant \tau_{sim} \wedge metric_{IoU} \leqslant \tau_{iou}$ **then**
    $S_r.append(triplet)$

Algorithm 2: Graph enrichment

Input: S, G_{cskg}
Output: S_e
$S_e = S$
for *each node* $\in S$ **do**
  $e_1 = cskg\_emb(node)$
  $triplets_{cskg} = query(G_{cskg}, node)$
  $triplets_{cskg} = preprocess(triplets_{cskg})$
  for *each triplet* $\in triplets_{cskg}$ **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 $s \geqslant \tau \wedge triplet \notin S_e$ **then**
      $S_e.append(triplet)$
  $S_e = postprocess(S_e)$

In order to infuse relevant triplets representing background knowledge and related facts from the CSKG [17], we parse the scene graph to a format compatible with the CSKG data model. Since similar entities tend to have similar vector representations in the embedding space [38], we leverage the graph embeddings to compute the similarity of nodes for various operations in the graph refinement and enrichment steps. The scene graph predictions are first refined using Algorithm 1 to discard any redundant or irrelevant predictions. The predicted objects with highly overlapping bounding boxes, similar names, or the same structural pattern in CSKG indicate the possibility of multiple redundant predictions of the same object. Such prediction errors are minimized at this stage by discarding

the object nodes that have a high intersection over union (IoU) of its bounding box or a high similarity score of CSKG embedding with another object node.

We use the Knowledge Graph Toolkit (KGTK) [15] to query CSKG and extract triplets from CSKG that include a subject or object node in the predicted scene graph. After extraction, any duplicate triplets and the triplets with both nodes similar (e.g. (person, synonym, person) and (chair, similarTo, chair)) are discarded in the preprocessing step because they do not provide any useful information. Based on the embedding similarity of the object nodes and the extracted nodes, the extracted nodes with reasonable structural similarity with the corresponding object nodes are linked via extracted edges in the scene graph. If an extracted node is already present in the scene graph, the new edge is linked to the existing node, otherwise, the new node is created and linked in the scene graph. In postprocessing, the format of the enriched scene graph is adjusted according to the original scene graph representation so that the enriched scene graphs can be evaluated for performance comparison or can be used in a downstream reasoning task. Since the predicates integrated from VG are expressed as "LocatedNear" edge type in the CSKG, we replaced the predicates in triplets extracted from the VG source in CSKG with the most frequent predicate type between the nodes in the original VG dataset. This post-processing step uses statistical prior knowledge from VG about the possible predicates between a pair of objects (nodes) in relationships to further interpret the relationship predicate. Algorithm 2 gives an overview of the steps in extracting commonsense knowledge from CSKG and integrating it into the scene graph. The thresholds in both algorithms were set to 0.5 for the experimental evaluation. These thresholds determine the trade-off between the number and the accuracy of detected and infused relationships.

4 Experiments and Results

4.1 Experimental Setup

Dataset We used the commonly used subset [59] of the Visual Genome dataset containing the most frequent 50 predicate classes and 150 object classes for training Faster RCNN, SGG model and 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% samples were used for evaluation. The longer dimension of each image was resized to 1024 pixels and the shorter dimension is adjusted accordingly. We use the pre-trained CSKG embeddings [17] for computing the similarity of nodes in the graph refinement and enrichment steps of the proposed approach.

Evaluation Protocol We used the cross-entropy loss to evaluate the training performance of the Faster RCNN and SGG models. Mean average precision (mAP) [10] was used to evaluate the object detection performance of Faster RCNN. For evaluating the performance of SGG before and after commonsense knowledge infusion, we used the most widely used metric, Recall@K ($R@K$) [31],

which is defined as the fraction of times the correct relationship is predicted in the top K confident relationship predictions. We compared the performance of the proposed method and recent SGG methods using the standard metric and benchmark dataset. We also analysed some qualitative results of the proposed method. Additionally, we employed an existing image generation method [18] as a downstream task of scene graph generation to further evaluate the proposed method by comparing the results of image generation from scene graphs before and after commonsense knowledge infusion.

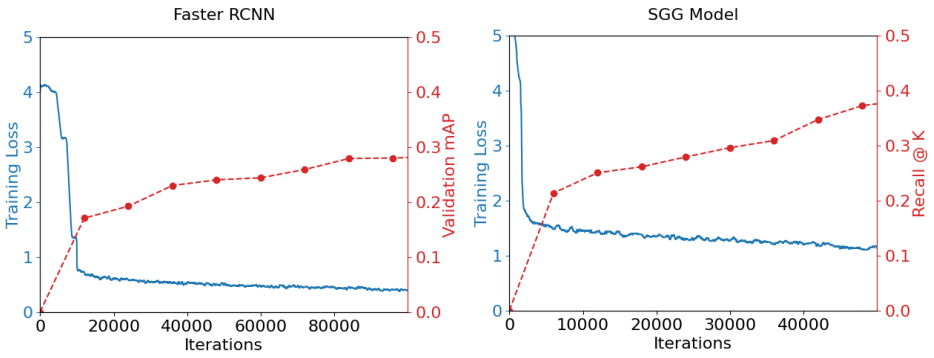

Fig. 3. Training progress plots along with periodic validation checks of the Faster RCNN and SGG models.

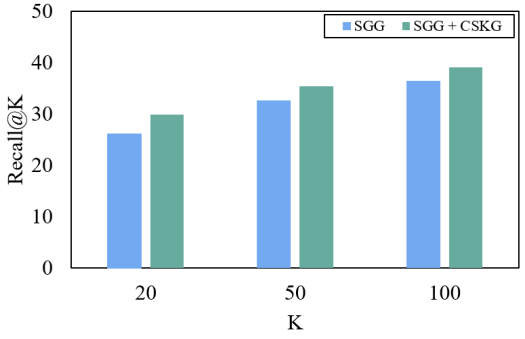

Fig. 4. Comparison of Recall@K of SGG before and after commonsense knowledge infusion.

Table 2. Comparison of the proposed method with the existing state-of-the-art SGG approaches in terms of Recall@K (R@K) on the Visual Genome dataset

SGG Method	**Approach**	**Commonsense Knowledge Source**	**R@20 (%)**	**R@50 (%)**	**R@100 (%)**
Proposed Method (SGG+CSKG)	Scene graph enrichment via commonsense knowledge infusion from different sources	**CSKG [17]**	**29.89**	**35.4**	**39.12**
GLAT [65]	Transformer-based GNN for visual commonsense reasoning	-	-	-	38.8
Unbiased SGG [48]	Causal inference and total direct effect	-	25.8	33.3	37.8
Proposed Method (SGG Only)	Scene graph generation based on fusion of visual (region and object) and text features	-	26.1	32.7	36.5
GB-Net [63]	Message passing between scene graphs and commonsense graph	ConceptNet [44], WordNet [36], Visual Genome [23]	-	29.4	35.1
VCTree [49]	Dynamic tree structures and Bi-dir TreeLSTM	-	22	27.9	31.3
IRT-MSK [12]	Instance Relation Transformer with Multiple Structured Knowledge	ConceptNet [44], Visual Genome [23]	22.2	27.2	31.2
Neural Motifs [66]	Stacked Motif Networks	-	21.7	27.3	30.5
KERN [4]	Knowledge-embedded routing network	-	-	27.1	29.8
COACHER [19]	Zero-shot relationship prediction via commonsense infusion	ConceptNet [44]	13.42	19.31	22.22
KB-GAN [11]	Commonsense and reconstruction-based object and phrase refinement	ConceptNet [44]	-	13.65	17.57
FactorizableNet [27]	Clustering-based graph factorization	-	-	13.06	16.47
MSDN [28]	Scene description at object, phrase and caption levels	-	-	10.72	14.22
Graph RCNN [60]	RPN followed by Attention GCN	-	-	11.4	13.7
IMP [59]	Object and relationship feature refinement via message passing	-	-	3.44	4.24

4.2 Results

Training and Evaluation of Models We trained the Faster RCNN model on the images and groundtruth annotations of objects in the Visual Genome dataset with Stochastic Gradient Descent (SGD) as an optimizer, batch size of 2 and initial learning rate of 0.002 which was decayed by a factor of 10 after 60k and 80k iterations. We froze the trained Faster RCNN and trained the whole SGG model on the images and groundtruth annotations of objects and relationships in the Visual Genome dataset using SGD as an optimizer, batch size of 4 and initial learning rate of 0.04 which was decayed by a factor of 10 twice during training when the validation performance stops improving noticeably. The plots of training loss and validation mAP for object detection and training loss and R@K for scene graph detection are shown in Figure 3, which show a smooth convergence of the models during the training process. The Faster RCNN model achieved $29.19mAP$ (using 0.5 IoU threshold), while the SGG model achieved $R@K = 26.1, 32.7, 36.5$ for $K = 20, 50, 100$ on the test set. The training and evaluation of the SGG model was performed in the Scene Graph Detection (SGDet) setting.

Evaluation after commonsense knowledge infusion We repeated testing of the scene graph generation method after adding the proposed commonsense knowledge infusion steps and achieved $R@K = 29.89, 35.4, 39.12$ for $K = 20, 50, 100$ on the test set, which is considerably higher than the R@K values achieved for the scene graph generation without commonsense knowledge infusion steps, as shown in Figure 4. The diverse commonsense knowledge integrated into the scene graphs from CSKG includes visual cues about the spatial proximity of objects in the scene relative to each other and physical interactions between the objects from the knowledge base of Visual Genome. This helps in mitigating some missed or wrong predictions made during scene graph generation and improves the recall rate for relationship prediction.

Comparative Analysis A detailed comparative analysis of the proposed approach with the existing scene graph generation methods is presented in Table 2. The proposed method incorporates the latest, largest and most diverse commonsense knowledge source from a consolidation of 7 distinct sources, and thus achieves higher recall score ($R@K = 29.89, 35.4, 39.12$ for $K = 20, 50, 100$) for SGG on the benchmark Visual Genome dataset as compared to the state-of-the-art technique ($R@K = 25.8, 33.3, 37.8$ for $K = 20, 50, 100$).

Qualitative Results Some qualitative results of the proposed method on Visual Genome images are shown in Figure 5. In addition to the objects and their pairwise visual relationships, the commonsense knowledge-based scene graphs contain the background facts about the underlying concepts, additional knowledge about the spatial proximity of objects in the scene relative to each other, and possible physical interactions between the objects. The useful background

facts include *(person, requires, eating)* and *(food, usedFor, eating)* in Figure 5(a). The commonsense relationships about spatial proximity such as *(tree, on, street)* in Figure 5(b) and the commonsense relationships about object interactions such as *(person, holding, surfboard)* in Figure 5(c) complement the scene graph representations.

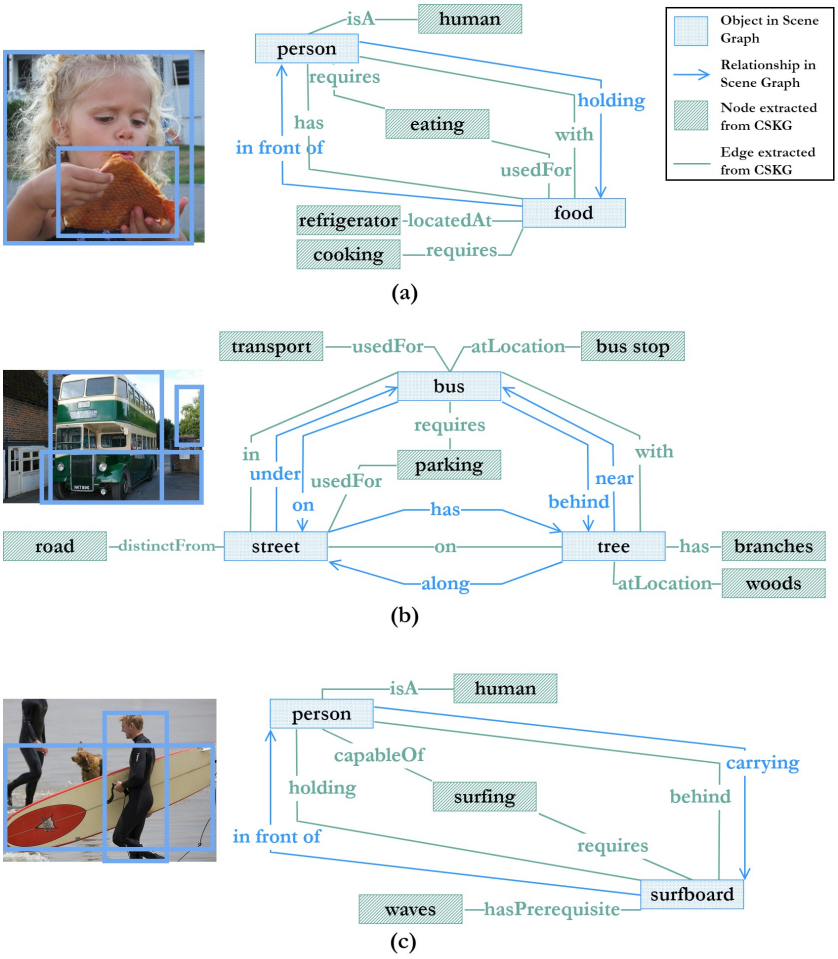

Fig. 5. Some qualitative results of the proposed commonsense knowledge-based scene graph generation method.

Downstream Task The rich and heterogeneous scene representations generated by the proposed method can significantly improve the downstream visual

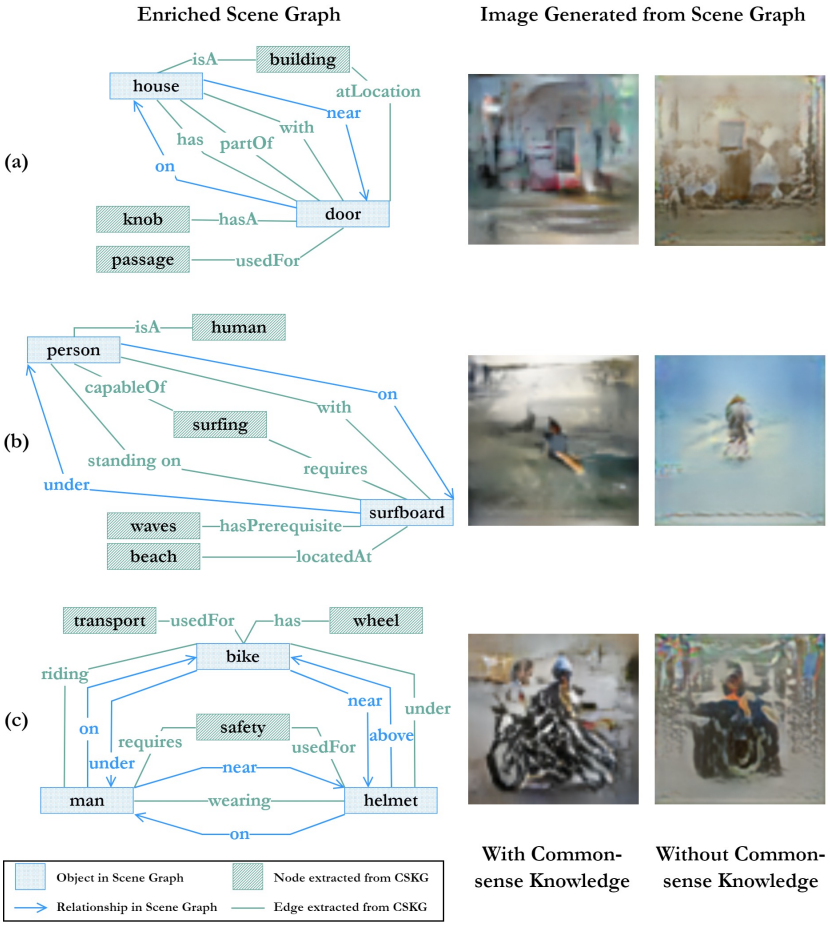

Fig. 6. Results of image generation using scene graphs generated by the proposed method.

reasoning tasks including image captioning, image generation, VQA, image retrieval, visual storytelling and multimedia event processing.

We employed an existing image generation method [18] as a downstream task of scene graph generation to further evaluate the proposed method. We trained the image generation network on the Visual Genome subset that was used to train the scene graph generation model. The trained network was used to generate images from scene graphs before and after commonsense knowledge infusion. The results of image generation from scene graphs are presented in Figure 6, which shows that the commonsense knowledge-based scene graphs generate more realistic images in which the semantic concepts in the input scene graph can be more clearly observed.

5 Conclusion

The use of commonsense knowledge for expressive and accurate visual understanding is inevitable due to its potential in complementing scene representations by providing necessary information for higher-level reasoning. In this paper, we propose a commonsense knowledge-based scene graph generation approach, which enriches the scene graph of an image with background knowledge and relevant facts extracted from CSKG, which is the latest, largest, and most diverse commonsense knowledge source. In the experimental and comparative analysis on the benchmark Visual Genome dataset, the proposed method achieved a higher recall rate ($R@K = 29.89, 35.4, 39.12$ for $K = 20, 50, 100$) as compared to the existing state-of-the-art technique ($R@K = 25.8, 33.3, 37.8$ for $K = 20, 50, 100$). We further evaluated the proposed method by employing image generation as a downstream task and showed improved qualitative results of image generation from scene graphs after commonsense knowledge infusion. The promising results depict the effectiveness of the rich and heterogeneous commonsense knowledge-based scene graph representations in improving the expressiveness and performance of visual reasoning tasks. In future work, we will investigate zero-shot and few-shot SGG using consolidated commonsense knowledge to reduce computational costs and requirement of training data and to allow the SGG model to predict unseen or rare object and predicate categories. We will also evaluate the efficacy of the proposed method in downstream reasoning tasks including multimedia event processing, image captioning, visual question answering and image retrieval.

References

1. Baier, S., Ma, Y., Tresp, V.: Improving visual relationship detection using semantic modeling of scene descriptions. In: International Semantic Web Conference. pp. 53–68. Springer (2017)
2. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley framenet project. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. pp. 86–90 (1998)
3. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: Scene graphs: A survey of generations and applications. arXiv preprint arXiv:2104.01111 (2021)
4. Chen, T., Yu, W., Chen, R., Lin, L.: Knowledge-embedded routing network for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6171 (2019)
5. Curry, E., Salwala, D., Dhingra, P., Pontes, F.A., Yadav, P.: Multimodal event processing: A neural-symbolic paradigm for the internet of multimedia things. IEEE Internet of Things Journal (2022)
6. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE conference on computer vision and Pattern recognition. pp. 3076–3086 (2017)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)

8. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1933–1941 (2016)
9. Gangemi, A., Alam, M., Asprino, L., Presutti, V., Recupero, D.R.: Framester: A wide coverage linguistic linked data hub. In: European Knowledge Acquisition Workshop. pp. 239–254. Springer (2016)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587 (2014)
11. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1969–1978 (2019)
12. Guo, Y., Song, J., Gao, L., Shen, H.T.: One-shot scene graph generation. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 3090–3098 (2020)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969 (2017)
14. Hung, Z.S., Mallya, A., Lazebnik, S.: Contextual translation embedding for visual relationship detection and scene graph generation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
15. Ilievski, F., Garijo, D., Chalupsky, H., Divvala, N.T., Yao, Y., Rogers, C., Li, R., Liu, J., Singh, A., Schwabe, D., et al.: Kgtk: a toolkit for large knowledge graph manipulation and analysis. In: International Semantic Web Conference. pp. 278–293. Springer (2020)
16. Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D.L., Szekely, P.: Dimensions of commonsense knowledge. arXiv preprint arXiv:2101.04640 (2021)
17. Ilievski, F., Szekely, P., Zhang, B.: Cskg: The commonsense knowledge graph. In: European Semantic Web Conference. pp. 680–696. Springer (2021)
18. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1219–1228 (2018)
19. Kan, X., Cui, H., Yang, C.: Zero-shot scene graph relation prediction through commonsense knowledge integration. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 466–482. Springer (2021)
20. Khan, M.J., Curry, E.: 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)
21. Kipfer, B.: Roget's 21st century thesaurus in dictionary form (éd. 3). new york: The philip lief group (2005)
22. Koner, R., Li, H., Hildebrandt, M., Das, D., Tresp, V., Günnemann, S.: Graphhopper: Multi-hop scene graph reasoning for visual question answering. In: International Semantic Web Conference. pp. 111–127. Springer (2021)
23. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
24. Lee, C.W., Fang, W., Yeh, C.K., Wang, Y.C.F.: Multi-label zero-shot learning with structured knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1576–1585 (2018)

25. Lee, S., Kim, J.W., Oh, Y., Jeon, J.H.: Visual question answering over scene graph. In: 2019 First International Conference on Graph Computing (GC). pp. 45–50. IEEE (2019)
26. Li, Y., Ouyang, W., Wang, X., Tang, X.: Vip-cnn: Visual phrase guided convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1347–1356 (2017)
27. Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., Wang, X.: Factorizable net: an efficient subgraph-based framework for scene graph generation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 335–351 (2018)
28. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE international conference on computer vision. pp. 1261–1270 (2017)
29. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). pp. 2117–2125 (2017)
30. Liu, L., Wang, M., He, X., Qing, L., Chen, H.: Fact-based visual question answering via dual-process system. Knowledge-Based Systems p. 107650 (2021)
31. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision. pp. 852–869. Springer (2016)
32. Ma, C., Sun, L., Zhong, Z., Huo, Q.: Relatext: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks. Pattern Recognition **111**, 107684 (2021)
33. Ma, K., Ilievski, F., Francis, J., Bisk, Y., Nyberg, E., Oltramari, A.: Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In: 35th AAAI Conference on Artificial Intelligence (2021)
34. McCarthy, J., et al.: Programs with common sense. RLE and MIT computation center (1960)
35. Mi, L., Chen, Z.: Hierarchical graph attention network for visual relationship detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13886–13895 (2020)
36. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
37. Narasimhan, M., Schwing, A.G.: Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In: Proceedings of the European conference on computer vision (ECCV). pp. 451–468 (2018)
38. Palmonari, M., Minervini, P.: Knowledge graph embeddings and explainable ai. Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges, IOS Press,, Amsterdam pp. 49–72 (2020)
39. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1981–1990 (2019)
40. Prakash, A., Debnath, S., Lafleche, J.F., Cameracci, E., Birchfield, S., Law, M.T., et al.: Self-supervised real-to-sim scene generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16044–16054 (2021)
41. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(6), 1137–1149 (2016)
42. Sadeghi, M.A., Farhadi, A.: Recognition using visual phrases. In: CVPR 2011. pp. 1745–1752. IEEE (2011)

43. Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., Choi, Y.: Atomic: An atlas of machine commonsense for if-then reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3027–3035 (2019)
44. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Thirty-first AAAI conference on artificial intelligence. pp. 4444–4451 (2017)
45. Su, Z., Zhu, C., Dong, Y., Cai, D., Chen, Y., Li, J.: Learning visual knowledge memory networks for visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7736–7745 (2018)
46. Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.: Energy-based learning for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13936–13945 (2021)
47. Tang, K.: A scene graph generation codebase in pytorch (2020), `https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch`
48. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3716–3725 (2020)
49. Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.: Learning to compose dynamic tree structures for visual contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6619–6628 (2019)
50. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
51. Wan, H., Ou, J., Wang, B., Du, J., Pan, J.Z., Zeng, J.: Iterative visual relationship detection via commonsense knowledge graph. In: Joint International Semantic Technology Conference. pp. 210–225. Springer (2019)
52. Wang, H., Zhang, F., Xie, X., Guo, M.: Dkn: Deep knowledge-aware network for news recommendation. In: Proceedings of the 2018 world wide web conference. pp. 1835–1844 (2018)
53. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(10), 2413–2427 (2017)
54. Wang, R., Wei, Z., Li, P., Zhang, Q., Huang, X.: Storytelling from an image stream using scene graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 9185–9192 (2020)
55. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph matching for relationship-aware image-text retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1508–1517 (2020)
56. Wang, X., Ye, Y., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6857–6866 (2018)
57. Wu, X., Sahoo, D., Hoi, S.C.: Recent advances in deep learning for object detection. Neurocomputing (2020)
58. Xie, Y., Pu, P.: How commonsense knowledge helps with natural language tasks: A survey of recent resources and methodologies. arXiv preprint arXiv:2108.04674 (2021)
59. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5419 (2017)

60. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph r-cnn for scene graph generation. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–685 (2018)
61. Yang, X., Zhang, H., Cai, J.: Auto-encoding and distilling scene graphs for image captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
62. Ye, K., Kovashka, A.: Linguistic structures as weak supervision for visual scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8289–8299 (June 2021)
63. Zareian, A., Karaman, S., Chang, S.F.: Bridging knowledge graphs to generate scene graphs. In: European Conference on Computer Vision. pp. 606–623. Springer (2020)
64. Zareian, A., Karaman, S., Chang, S.F.: Weakly supervised visual semantic parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3736–3745 (2020)
65. Zareian, A., Wang, Z., You, H., Chang, S.F.: Learning visual commonsense for robust scene graph generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16. pp. 642–657. Springer (2020)
66. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5831–5840 (2018)
67. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9185–9194 (2019)