KnowZRel: Common Sense Knowledge-Based Zero-Shot Relationship Retrieval for Generalized Scene Graph Generation

M. Jaleed Khan, John G. Breslin, *Senior Member, IEEE*, and Edward Curry

***Abstract*—A scene graph is a key image representation in visual reasoning. The generalizability of scene graph generation (SGG) methods is crucial for reliable reasoning and real-world applicability. However, imbalanced training datasets limit this, underrepresenting meaningful visual relationships. Current SGG methods using external knowledge sources face limitations due to these imbalances or restricted relationship coverage, impacting their reasoning and generalization capabilities. We propose a novel neurosymbolic approach that integrates data-driven object detection with heterogeneous knowledge graph-based object refinement and zero-shot relationship retrieval, highlighting the loosely coupled synergy between neural and symbolic components. This combination addresses the limitations of imbalanced training datasets in SGG and enables effective prediction of unseen visual relationships. Objects are detected using a region-based deep neural network and refined based on their positional and structural similarity, followed by retrieval of pairwise visual relationships using a heterogeneous knowledge graph. The redundant and irrelevant visual relationships are discarded based on the similarity of relationship labels and node embeddings. Finally, the visual relationships are interlinked to generate the scene graph. The employed heterogeneous knowledge graph combines diverse knowledge sources, offering rich common sense knowledge about objects and their interactions in the world. Our method, evaluated using the benchmark visual genome (VG) dataset and zero-shot recall (zR@K) metric, shows a 59.96% improvement over existing state-of-the-art methods, highlighting its effectiveness in generalized SGG. The object refinement step effectively improved the object detection performance by 57.1%. Additional evaluation using the GQA dataset confirms the cross-dataset generalizability of our method. We also compared various knowledge sources and embedding models to determine an optimal combination for zero-shot SGG. The source code is available at https://github.com/jaleedkhan/zsrr-sgg.**

Received 16 October 2024; accepted 18 February 2025. Date of publication 21 February 2025; date of current version 26 November 2025. This work was supported by the Research Ireland under Grant 18/CRT/6223 and Grant 12/RC/2289_P2. This article was recommended for publication by Associate Editor Álvaro Parra upon evaluation of the reviewers' comments. *(Corresponding author: M. Jaleed Khan.)*

M. Jaleed Khan is with the Research Ireland Centre for Research Training in Artificial Intelligence, Data Science Institute, University of Galway, H91TK33 Galway, Ireland (e-mail: m.khan12@universityofgalway.ie).

John G. Breslin and Edward Curry are with the Insight Research Ireland Centre for Data Analytics, Research Ireland Centre for Research Training in Artificial Intelligence, Data Science Institute, University of Galway, H91TK33 Galway, Ireland (e-mail: john.breslin@universityofgalway.ie; edward.curry@universityofgalway.ie).

Digital Object Identifier 10.1109/TAI.2025.3544177

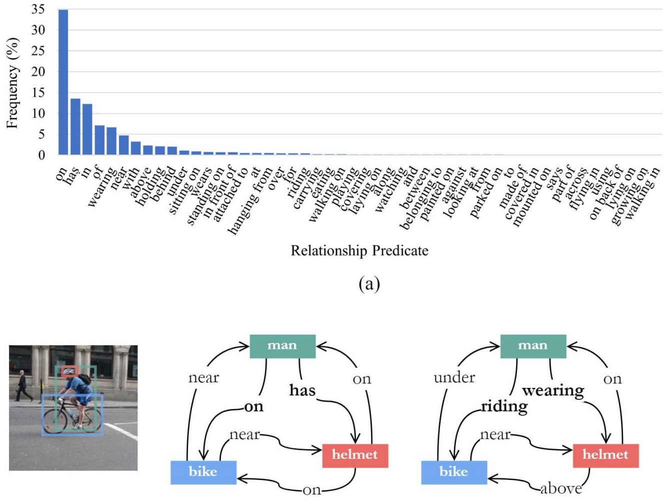
***Impact Statement*—Visual reasoning plays an essential role in understanding, interpreting, and reasoning about visual content, such as images and videos. It enables a wide range of applications of artificial intelligence, including autonomous systems, semantic image search, and assistive technologies. Scene graph generation (SGG), a key component in this process, offers semantically rich image representations that are fundamental for visual reasoning. However, its reliance on data-centric methods leads to challenges from imbalanced datasets and limited relational scope, particularly in zero-shot SGG. The proposed method, leveraging common sense knowledge, significantly enhances the generalizability of SGG. It notably boosts the zero-shot recall rate by 59.96% on the standard benchmark and demonstrates cross-dataset generalizability. This advancement facilitates more accurate and intuitive visual reasoning and encourages further research on knowledge-based approaches for generalized SGG to extend and enhance its practical applications.**

***Index Terms*—Common sense knowledge, image representation, neurosymbolic integration, scene graph, scene understanding, unseen relationships, visual reasoning, zero-shot retrieval.**

I. Introduction

HIGH-LEVEL visual reasoning tasks require semantically rich image representation to produce accurate and meaningful results. Scene graph generation (SGG) aims to extract and represent the contents of an image using a structured graph that contains objects as nodes and their relationships as edges. Generally, SGG approaches involve identifying and localizing objects in an image using deep learning-based detection and classification techniques, followed by visual relationship prediction using visual-linguitic multimodal techniques [1]. Semantic representation of images is essential for numerous downstream visual reasoning tasks that require a higher-level understanding and interpretation of the contents and context of images. Some of the downstream tasks of SGG include Visual Question Answering (VQA) [2], image captioning [3], image retrieval [4], multimedia event processing (MEP) [5], human action recognition [6], robot task planning [7], and context-aware augmented reality [8].

Visual relationships are the core of scene graphs, and their accurate prediction plays a significant role in scene understanding and visual reasoning. Most SGG methods are data-driven, relying on large-scale labeled datasets for training object detection and relationship prediction models. Despite high accuracy



(a)

(b)

Fig. 1. (a) The highly-skewed distribution of visual relationship predicates in crowdsourced datasets (VG [9]) in which a few generic predicates (e.g., "on" and "has") occur highly frequently as compared to more meaningful predicates (e.g., "riding" and "wearing"), which reflects the intrinsic bias towards generic concepts that cover a broad range of visual relationship types but are not specific and expressive enough. This leads to (b) biased visual relationship prediction and reduced generalizability of SGG methods for unseen visual relationships.

in object detection, visual relationship prediction remains challenging due to the vast number of possible relationships and their diverse appearances and interpretations across contexts and scenes. This is reflected in datasets, like visual genome (VG) [9], as a long-tailed distribution of relationships and a higher frequency of generic relationships compared to more meaningful ones. Many relationship predicates, particularly meaningful ones (e.g., "riding," "wearing," and "looking at" as shown in Fig. 1), have limited instances in datasets, complicating their feature representation learning for existing SGG methods. Additionally, visual features of relationships can vary greatly across contexts and scenes (e.g., "holding umbrella," "holding bat," and "holding food," have different visual features despite the same relationship predicate).

Given the vast number of possible relationship combinations and the limited availability of training samples for rare predicates, this imbalance overfits SGG methods to frequent relationships, leading to poor generalization on unseen or infrequent relationships. Addressing this challenge is crucial for developing robust SGG models, as it directly affects their ability to generalize to unseen relationships, especially in zero-shot settings. This emphasizes the need to investigate zero-shot SGG approaches for predicting unseen visual relationships, where models must infer relationships not present or underrepresented in the training data.

The performance of existing SGG methods falls drastically in the zero-shot setting due to the challenge of predicting unseen visual relationships. Zero-shot SGG refers to the task of predicting relationships between objects in an image that were not seen during the training phase. This requires the model to generalize beyond the training data and utilize external knowledge to make accurate predictions. Recent methods have attempted to address this through relational graph neural networks [10] and graph link prediction [11]. Since visual-linguistic features alone do not generalize well to unseen relationships, some SGG methods used external knowledge sources, such as language or statistical priors [12], [13], and knowledge graphs (KGs) [14], [15], [16], to infuse common sense knowledge into the SGG pipeline. However, priors have limited applicability to unseen concepts, lack visual cues, and carry dataset bias. KGs, which capture a wider range of general concepts, have successfully complemented visual-linguistic features and moderately improved SGG performance [17]. KG embeddings encode structural and semantic information about entities and relations [18] and are used in various KG-related tasks. ConceptNet [19] has been employed for feature refinement [15] and message propagation [16] within SGG. The existing methods still require significant improvement, particularly in the zero-shot setting [20]. COACHER [14], based on ConceptNet, is the only knowledge-based SGG method primarily designed for the zero-shot setting.

Different KGs capture various types of common sense knowledge based on their scope and focus, such as ConceptNet [19], ATOMIC [21], VG [9], and WordNet [22], which offer general, procedural, visual, and lexical knowledge, respectively. A heterogeneous KG, such as a common sense knowledge graph (CSKG) [23], consolidates the unique knowledge from multiple KGs into a unified and rich representation of common sense knowledge. CSKG has proven effective in generalized common sense question answering [24] and semi-supervised expressive SGG [25]. However, its potential for unseen visual relationship prediction in zero-shot SGG remains unexplored [20]. This article proposes a novel zero-shot relationship retrieval method based on a heterogeneous knowledge graph for generalized SGG. Our method presents a simple yet novel and effective integration of Faster-RCNN for object detection, followed by a refinement process to improve the accuracy of detected objects and the zero-shot relationship retrieval mechanism using a heterogeneous knowledge graph. This approach can be categorized as loosely coupled neurosymbolic integration of deep neural networks and symbolic knowledge graphs according to the classification defined in [20]. The proposed method is illustrated in Fig. 2. It detects and refines objects in an input image, queries CSKG for relevant triples shared among object pairs, discards irrelevant or redundant triples, and constructs the scene graph. Our neurosymbolic approach leverages the strengths of both data-driven and symbolic techniques, utilizing the structural richness of a heterogeneous knowledge graph, thereby overcoming the constraints posed by training data imbalance. The experimental and comparative analysis performed using two benchmark datasets and standard evaluation metrics demonstrates how this integration improves SGG generalization, especially for unseen relationships. The key contributions of our work are as follows.

1) We proposed a novel zero-shot SGG method that comprises deep learning-based object detection and heterogeneous knowledge graph-based zero-shot relationship retrieval. The structural, semantic, and positional similarity of nodes and edges in the graph are taken into account

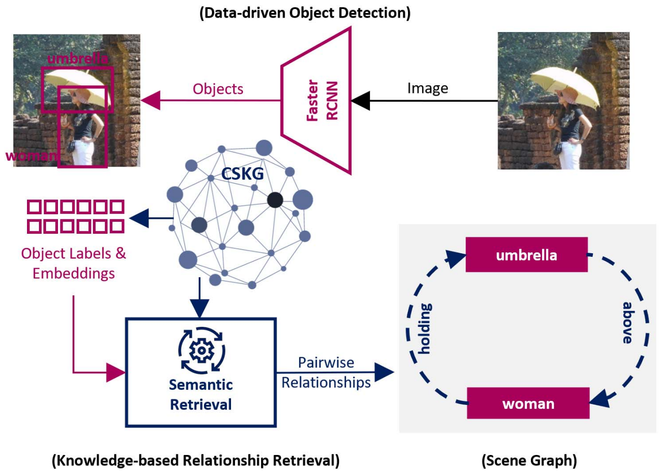


Fig. 2. Proposed zero-shot SGG approach comprising data-driven object detection and common sense knowledge-based relationship retrieval.

to refine the objects and visual relationships, followed by graph construction (Fig. 3, Algorithm 1 and Algorithm 2).

2) We performed comprehensive experimental and comparative analysis using the VG [9] benchmark dataset and GQA [26] dataset and the standard evaluation metrics for object detection and SGG. In the evaluation of object detection and refinement, the results showed significant improvements in performance after applying the refinement process, with mean Average Precision (mAP) improving by 57.1% on the VG dataset and 62.5% on the GQA dataset (Table I and Fig. 4). These improvements depict the effectiveness of the refinement process in enhancing object detection accuracy, thus validating its importance in the overall SGG pipeline.
3) The proposed method outperformed the existing state-of-the-art methods with a considerable improvement of over 59.96% in the zero-shot recall rate (Table II). This demonstrates the effectiveness of heterogeneous common sense knowledge-based relationship retrieval for generating generalized scene graphs. (Fig. 5). Encouraging results obtained on the GQA dataset depict the generalizability of the proposed method across datasets.
4) We compared different KGs used in SGG and different KG embedding models, observing that ComplEx [27] embeddings of CSKG [23] yielded the highest performance (Tables III and IV). This is due to ComplEx's capability to represent complex-valued and multidimensional relationships between entities and the richness and diversity of common sense knowledge in CSKG.

II. Related Work

A. Scene Graph Generation (SGG)

SGG is a structured and semantically-grounded representation of visual scenes. SGG methods typically follow a bottom-up mechanism in which the objects in the scene are identified and localized, and their attributes are recognized using deep learning-based detection and classification techniques, followed by multimodal approaches for pairwise relationship prediction based on visual features, linguistic features, symbolic rules, or hybrid techniques. Since most of the SGG methods are supervised or semi-supervised, these methods struggle with the performance of visual relationship prediction due to the challenges posed by the biased distribution of relationships in crowdsourced training datasets and the complex nature and intra-class variability of visual relationships. To address these challenges, researchers have adopted new approaches, such as transfer learning [28], causal inference [29], and self-supervised learning [30], and examined new aspects of visual relationships, such as heterophily [31] and saliency [32]. BGTNet [33] is a transformer-based SGG method that uses bi-directional gated recurrent unit (GRU) layers for object-object communication to enhance object information and transformer encoders for object classification and the creation of object-specific edge information. Xu et al. [34] proposed an iterative message-passing approach based on standard recurrent neural networks (RNN) and contextual cues for a joint inference for visual relationship prediction is SGG. Tang et al. [35] proposed VCTree, a visual context tree model comprising dynamic tree structures to encode the inherent parallel and hierarchical relationships between objects in an image, followed by end-task supervised learning integrated with tree structure reinforcement learning for SGG. In their recent work [29], Tang et al. presented an SGG framework based on causal inference; a causal graph is constructed and used to perform traditional biased training first, followed by counterfactual causality to infer and remove the effects of bias quantified using total direct effect (TDE) predicate score. The fully convolutional SGG model (FCSGG) [36] employed a bottom-up approach that encodes objects as bounding box centre points and relationships as 2D vector fields called relation affinity fields (RAFs) to allow for encoding both semantic and spatial features for simultaneous object detection and visual relationship prediction. The Union Message-based SGG (USGG) architecture [37] leverages relational semantics from union regions around object pairs to enhance feature extraction and relational representations through a union embedding and fusion network for SGG. RU-Net [38] introduced unrolled message passing with graph regularization and a group diversity enhancement module to address ambiguous object representations and low diversity in relationship predictions in SGG. Some existing SGG methods have been adapted and evaluated in the zero-shot setting, showing a drastic decrease in performance. Only a few SGG methods have been proposed primarily for zero-shot SGG, which are discussed in the next section.

B. Zero-Shot Relationship Prediction in SGG

It is essential for SGG models to perform well in the zero-shot setting because it enables them to generalize to new situations and handle previously unseen visual concepts. This is particularly useful in downstream reasoning tasks and real-world applications in which collecting large and diverse labeled datasets for training is difficult or expensive. Therefore, zero-shot SGG reduces the need for large amounts of labeled data and increases the generalizability and applicability of SGG. Relational graph neural network (RGNN) [10] is an encoder-decoder framework for zero-shot SGG, which uses semantic correlations between relationships to make predictions for unseen relationships; the encoder aggregates the training samples of visual relationships into a graph and uses a GNN to learn entity embeddings that are used to guide the vision-based decoder to predict visual relationships. The text-image-joint SGG (TISGG) model [39] enhances the generalization of SGG by aligning visual and text features and refining predictions using factual knowledge from the dataset by employing strategies like character-guided sampling and informative re-weighting to mitigate dataset biases. Goel et al. [40] proposed a two-stage model-agnostic SGG training pipeline with label refinement and latent space augmentation, leveraging relation label informativeness and imputing missing informative relations for enhanced training. The quaternion relation embedding (QuatRE) approach [41] for SGG leverages hypercomplex representations to better capture entity interactions by employing the Hamilton product in quaternion space to represent relations, offering enhanced expressiveness and generalization. Lexical knowledge-aware memory network (LKMN) [42] distills lexical knowledge of different objects and constructs multimodal representations of pairwise objects to reduce the intra-class variation of the relationship predicate, followed by the creation of a compact semantic space in which the representations of unseen relationships are reconstructed based on the seen relationships, allowing the model to make predictions for unseen relationships using the learned relationships as a guide. XKGC [11] employed a graph completion strategy for zero-shot SGG in which the information about unseen relationships is integrated with the visual features extracted from the image, transforming the zero-shot relationship prediction into a graph link prediction task. Peng et al. [43] introduced a causal graph for training and a degree-of-difficulty loss (DDloss) to mitigate bias toward dominant classes in imbalanced datasets. The triple calibration and reduction (T-CAR) framework [44] enhanced compositional generalization and inference on unseen triples in SGG by calibrating and reducing unrealistic unseen triples through a triple calibration loss and unseen space reduction loss. Kim et al. [45] leveraged large language models to address semantic over-simplification and low-density scene graph issues in weakly-supervised SGG by utilizing chain-of-thought reasoning and few-shot learning strategies to enhance triple extraction and alignment. Since visual information is insufficient to generalize effectively to unseen visual relationships, the zero-shot performance of SGG methods is quite limited. The zero-shot approach, coupled with diverse common sense knowledge related to visual concepts from external sources, has the potential to improve SGG performance significantly.

C. *Common Sense Knowledge-Based SGG*

The initial SGG methods employed statistical priors and language priors as external sources of common sense knowledge. Zellers et al. [13] proposed stacked motif networks that capture higher-order motifs that are regularly occurring substructures in scene graphs and incorporate information across motifs to improve SGG performance; pre-computed frequency priors are used to incorporate common sense knowledge from dataset statistics. Liang et al. [12] proposed a deep variation-structured reinforcement learning framework (VRL) in which a deep Q-network is used to learn the reward function for a given visual relationship prediction task while infusing additional information about semantic correlations between the visual concepts from the language prior. However, priors lack visual cues, carry forward the bias in training datasets, and offer limited applicability to unseen visual relationships, as 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 in the zero-shot setting.

Since KGs capture a wider range of general concepts in the world in a structured way, they have been observed to be better at complementing visual-linguistic features in SGG [17]. Gu et al. [15] introduced a knowledge-based feature refinement with auxiliary image generation (KBGAN) method that extracts commonsense knowledge from ConceptNet to refine object and phrase features and uses an auxiliary image reconstruction path to regularize the SGG network and minimize the effects of dataset bias. Similarly, Zareian et al. [16] used ConceptNet for knowledge-based message propagation in graph neural networks for visual relationship prediction. A heterogeneous KG combines the diversity and coverage of commonsense knowledge from multiple KGs into a rich, diverse, and unified knowledge source that provides a more comprehensive and complete representation of common sense knowledge. Khan et al. [25] proposed knowledge enrichment of scene graphs using a heterogeneous KG, CSKG [23], that contains rich and diverse common sense knowledge integrated from seven different large-scale KGs, and reported encouraging improvement in visual relationship prediction and downstream reasoning; however, it is a semi-supervised approach and has not yet been evaluated in the zero-shot setting. On the other hand, COACHER [14] is an explicit zero-shot SGG approach leveraging ConceptNet for visual relationship retrieval; however, it does not leverage heterogeneous KGs. Leveraging heterogeneous common sense knowledge for zero-shot relationship prediction has not yet been explored; however, it has the potential to enable high-level structural and semantic understanding of unseen visual relationships, which is essential for generalized SGG.

III. Proposed Method

The proposed method is a simple yet effective combination of data-driven and knowledge-based approaches for generalized SGG. Fig. 3 provides an overview of the proposed method. Objects are detected using a region-based deep neural network and refined based on their positional and structural similarity (Algorithm 1). The pairwise visual relationships between the objects are retrieved using a heterogeneous knowledge graph (Algorithm 2). The redundant and irrelevant visual relationships are discarded based on the similarity of relationship labels and node embeddings. Finally, the visual relationships are interlinked to construct the scene graph.

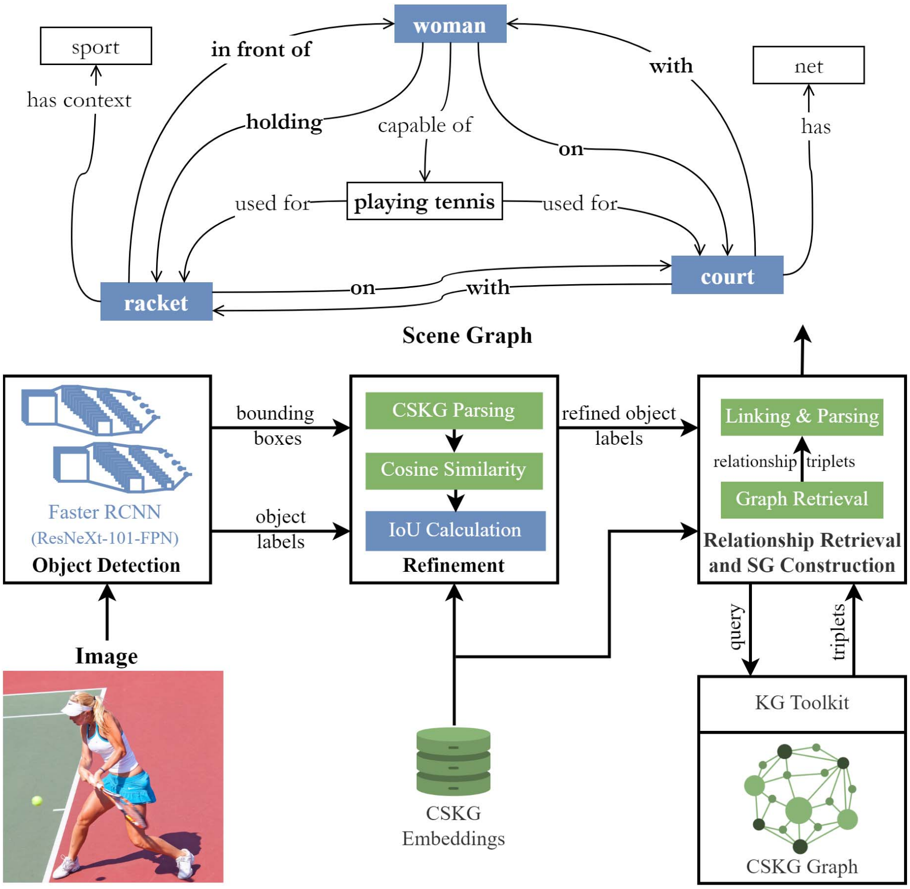


Fig. 3. The proposed zero-shot SGG method, comprising deep learning-based object detection and heterogeneous knowledge graph-based object refinement (Algorithm 1) and zero-shot relationship retrieval (Algorithm 2) for each object pair, followed by graph construction. The example scene graph represents objects and pairwise visual relationships (zero-shot) in the scene, e.g., *(woman, holding, racket)* and *(woman, on, court)*, and background information and related facts, e.g., *(woman, capable of, playing tennis)* and *(racket, used for, playing tennis)*, which enables high-level reasoning to deduce that *"the woman is playing tennis."*

A. Object Detection and Refinement

Faster RCNN [46] is employed for object detection with a ResNeXt-101-FPN [47] backbone feature extractor. Given an input image I, Faster RCNN outputs bounding boxes and class labels for each detected object, as presented in Algorithm 1. The bounding boxes, denoted as b, represent the position as coordinates (i.e., x and y coordinates for the top-left and bottom-right corners of the box), and the class labels, denoted as l, represent the object category (e.g., person and car). The positional similarity of each pair of objects is computed using the Intersection over Union (IoU) of their bounding boxes, denoted as $metric_{IoU}$. IoU measures the extent to which the bounding boxes overlap in the image. The object labels are parsed to the CSKG representation model and CSKG [23] embeddings of the object labels, denoted as e_1 and e_2, are used to compute the cosine similarity $metric_{sim}$ of the object nodes. The cosine similarity metric shows the structural and semantic similarity between two objects as similar entities in a KG have similar vector representations in the embedding space. A pair of objects with a high intersection over the union (IoU) of their bounding boxes or a high cosine similarity score of their CSKG embeddings indicates a redundant object, so one of the objects in the pair is eliminated. This is repeated for all pairs of detected objects to refine the list of detected objects based on the similarity scores $metric_{IoU}$ and $metric_{sim}$, and a refined list of objects, l_r is generated.

B. Relationship Retrieval and Scene Graph Construction

CSKG [23] is employed to extract structured background knowledge about the objects and their interactions in the form of relationship triples. The knowledge graph toolkit (KGTK) [48] is used to query CSKG and retrieve triples $triples$, having a detected object $l_r[i]$ as a node. The triples that have the same nodes (e.g., *person, similar to, person)* or nodes with similar graph embeddings *(man, is a, person*) at both ends indicate redundant triples that do not provide any new information and are discarded. The extracted triples with both nodes representing the detected objects are taken as direct relationships between the objects. For each CSKG triple that has one node representing a detected object (object node), we compare the graph embedding of the other (non-object) node with the graph embeddings of the detected objects, excluding the object node. If the non-object node is found to be structurally and semantically similar enough to a detected object, it is replaced by that detected object, and

Algorithm 1: Object Detection and Refinement.

```
Input: $I$ (image)
Output: $l_r$ (refined object labels)

$[b, l] = FasterRCNN(I)$
$l_r = []$.
$l = parse_{cskg}(l)$.
$l_r.append(l[0])$.
$i = 1$.
while $i < length(b)$ do
    $e_1 = cskg\_embedding(l[i])$.
    $b_1 = b[i]$.
    $is\_redundant = False$.
    $j = 0$.
    while $j < i$ do
        $e_2 = cskg\_embedding(l[j])$.
        $b_2 = b[j]$.
        $metric_{sim} = cosine\_sim(e_1, e_2)$.
        $metric_{IoU} = compute\_IoU(b_1, b_2)$.
        if $metric_{sim} \geq \tau_{sim}$ and $metric_{IoU} \geq \tau_{iou}$ then
            $is\_redundant = True$.
        end if
        $j = j + 1$.
    end while
    if $is\_redundant == False$ then
        $l_r.append(l[i])$.
    end if
    $i = i + 1$.
end while
```

Algorithm 2: Relationship Retrieval and Scene Graph Construction.

```
Input: $l_r$ (refined object labels), $G_{cskg}$ (CSKG graph)
Output: $S$ (scene graph)

$S = []$.
$i = 0$.
while $i < length(l_r)$ do
    $e_1 = cskg\_embedding(l_r[i])$.
    $triples = query(G_{cskg}, l_r[i])$.
    $j = 0$.
    while $j < length(triples)$ do
        $triple = triples[j]$.
        if $triple[node2] \notin l_r$ then
            $l_r.append(triple[node2])$.
        end if
        $e_2 = cskg\_embedding(triple[node2])$.
        if $cosine\_sim(e_1, e_2) \leq \tau$ and $triple \notin S$ then
            $S.append(triple)$.
        end if
        $j = j + 1$.
    end while
    $i = i + 1$.
end while
$S = parse_{sgg}(S)$.
```

a new direct relationship is formed between that object and the object node already present in this CSKG triple. However, if the non-object node is not similar to any detected object, it is added as a new object (or concept), and a new relationship is established between this new object and the object node already present in this CSKG triple. Once this process is completed for all the triples, a rich set of visual relationships S is obtained that encompasses relationships between the detected objects and relationships of the detected objects with related concepts in CSKG. These relationships provide background information about how the detected objects interact with each other and with other concepts in the world, along with associated facts about the detected objects. Finally, all the relationships are interlinked to construct a graph representation of the image. The format of this graph is parsed to the scene graph representation model by leveraging the statistical prior knowledge of the VG knowledge base without revealing the specific visual relationships present in VG. The predicates of CSKG triples with the VG knowledge base as the source are expressed as a generic "LocatedNear" edge type in CSKG. To correctly interpret the relationships, these predicates are replaced by the most common predicate type between the objects found in the original knowledge base. This complete process is presented in Algorithm 2.

The thresholds for IoU (τ_{iou} in Algorithm 1) and cosine similarity (τ_{sim} in Algorithm 1 and τ in Algorithm 2) metrics, used in both object refinement and relationship retrieval, strike a balance between the quantity and precision of the visual relationships in the generated scene graphs.

IV. Experiments and Results

A. Experimental Setup

1) Platform and Tools: We used PyTorch[1] and KGTK[2] for implementation and experiments on 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.

2) Datasets: VG [9] is the most widely used and standard dataset for SGG performance evaluation. It contains 108 K labelled images and annotations for objects and visual relationships, with the most frequent 150 object classes and 50 relationship classes included in the standard split [34] that were used to train the object detector and compare the retrieved relationships with the groundtruth for evaluation respectively. We also used the GQA dataset [26], which contains 113 018 images, 1702 object classes and 310 relationship types, with an 80-10-10 split for training, validation and testing. Given the long-tailed distribution of objects and relationships in the dataset that impacts the performance of SGG, we used the common subset of the dataset with the most frequent 800 object classes and 170 relationship classes.

[1] https://pytorch.org/

[2] https://kgtk.readthedocs.io/

3) Evaluation Metrics: We used the standard object detection metrics, including intersection over union (IoU), precision, recall, F-score, and mean average precision (mAP), to evaluate the Faster RCNN model. We used zero-shot Recall at K ($zR@K$) metric [49], [29] for quantitative evaluation and benchmark comparison of zero-shot SGG. It is the standard metric for zero-shot SGG and a variant of Recall at K ($R@K$). $R@K$ is defined as the fraction of times that the correct relationship is predicted among the top K confident relationship predictions. The confidence score of each prediction is considered when computing $R@K$, which means that the relationship labels must be correctly predicted and have a higher score to be counted as correct. $zR@K$ calculates $R@K$ for unseen relationships, i.e., those not in the training set. The recall rates reported in the results are based on thresholds for IoU (τ_{iou} in Algorithm 1) set to 0.5 and cosine similarity (τ_{sim} in Algorithm 1 and τ in Algorithm 2) set to 0.8.

4) Comparative Methods: We compared the performance of the proposed method with the existing state-of-the-art SGG methods evaluated in the zero-shot setting. The comparison includes

1) methods leveraging common sense knowledge, such as COACHER [14], MOTIF [13], KBGAN [15], and DeepVRL [12], as well as,
2) data-centric methods, such as TISGG [39], SHAGCL-DD [43], T-CAR [44], LLM4SGG [45], USGG [37], QuatRE [41], RU-Net [38], XKGC [11], MIL-SGG [40], RGNN [10], BGTNet [33], FCSGG [36], TDE [29], VCTree [35], and IMP [34].

The results of the comparative methods are reported in their best settings on the standard split of the dataset.

5) KG Embedding Models: We compare the performance of the following four KG embedding models in the zero-shot retrieval task.

1) TransE [50] is a KG embedding model that uses a translation operation to model relationships between entities.
2) RESCAL [51] represents entities as vectors and relationships as matrices, allowing it to capture higher-order relationships.
3) DistMult [52] uses a distance-based multiplication operation to model relationships.
4) ComplEx [27] extends DistMult by using complex-valued vectors.

The specific approach used by each model differs; however, they all aim to map entities and relationships in the KG to points in a high-dimensional vector space, such that the geometric relationships between the points reflect the relationships between the entities and relationships in the KG. We used the cosine similarity of KG embeddings of concepts in Algorithm 1 and Algorithm 2.

B. Results and Discussion

TABLE I
EVALUATION OF OBJECT REFINEMENT

Dataset	Method	Precision	Recall	F-score	mAP
VG	w/o refinement	76.4%	72.5%	74.4%	0.28
	w/ refinement	**82.7%**	**79.1%**	**80.9%**	**0.44**
GQA	w/o refinement	75.2%	71.3%	73.2%	0.24
	w/ refinement	**81.5%**	**77.6%**	**79.5%**	**0.39**

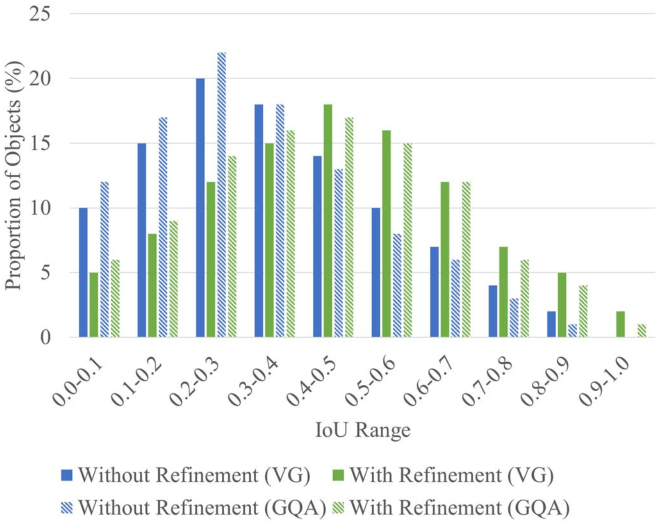


Fig. 4. Distribution of IoU scores for object detections with and without refinement.

1) Object Refinement Evaluation: We evaluated the accuracy of object detection with and without the refinement process in object detection (Algorithm 1) on the VG and GQA datasets to demonstrate the effectiveness of the refinement process. The results of the object detection accuracy with and without the refinement process are summarized in Table I, which shows a significant improvement in performance after applying the refinement process. Specifically, on the VG dataset, mAP improved from 0.28 to 0.44, and on the GQA dataset, mAP improved from 0.24 to 0.39. In addition, Fig. 4 illustrates the distribution of IoU scores for object detections before and after refinement. The refined detections show a higher concentration of IoU scores closer to 1, indicating more accurate bounding boxes. These improvements depict the effectiveness of the refinement process in enhancing object detection accuracy, thus validating its importance in the overall SGG pipeline.

2) Zero-Shot SGG Evaluation: The proposed method achieved an overall recall rate of $zR@K = 14.22, 25.43$ and 35.65 for $K = 20, 50$, and 100, respectively, on the VG test set in the zero-shot setting of SGG. The detailed recall rate, $zR@100$, for each relationship predicate, along with its proportion in the dataset, is shown in Fig. 5. Contrary to the conventional data-centric approaches, it is interesting to note that the proposed method achieves a consistent recall rate for most of the relationship predicates, irrespective of their distribution in the dataset. However, comparatively lower recall rates are noted for a few relationship predicates, such as "with," "and," and "to," due to their complexity and ambiguity in the KG domain. Fig. 6 demonstrates how the recall rate, $zR@100$, varies with the cosine similarity threshold used in Algorithm 2. As the threshold increases initially, recall

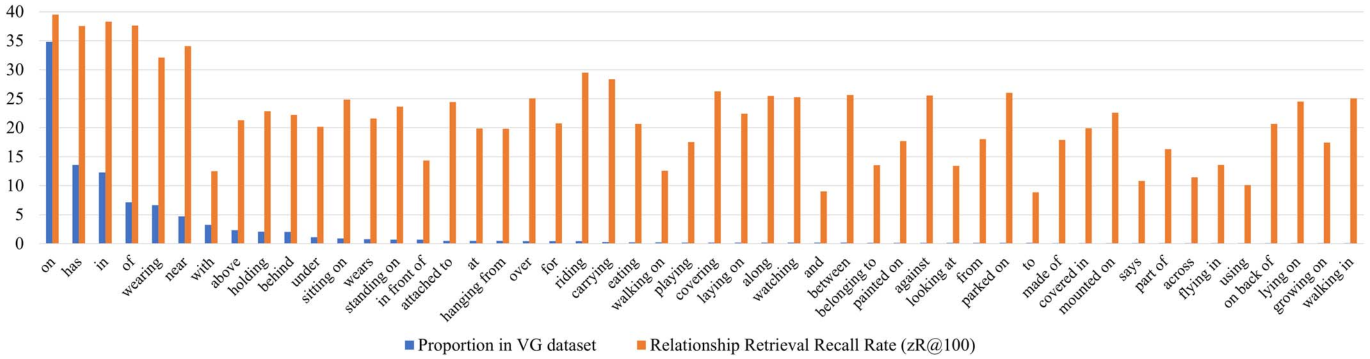


Fig. 5. Detailed result of recall rate for each relationship predicate in VG dataset.

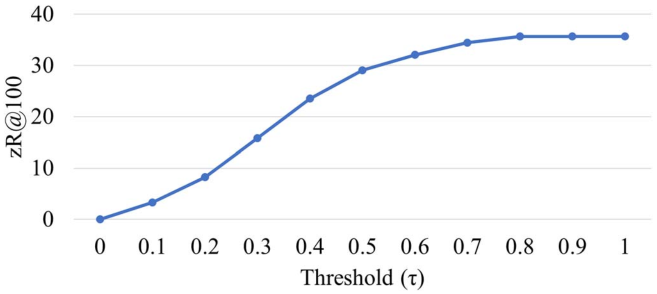


Fig. 6. Effect of varying cosine similarity threshold on the recall rate.

rises significantly, capturing more relationships between the detected objects. Between the threshold values of 0.5 and 0.7, recall growth diminishes, and plateaus afterwards, indicating the addition of more irrelevant than meaningful relationships. The best performance, i.e. $zR@100 = 35.65$, is achieved at 0.8 threshold.

Additionally, we evaluated the proposed method on the GQA dataset and observed an overall zero-shot recall rate of $zR@K = 12.47, 22.51$ and 29.56 for $K = 20, 50$, and 100, respectively. This test justifies that the proposed method is generalizable across datasets.

3) Benchmark Comparison: The performance of the proposed method is compared with the state-of-the-art methods in Table II, which shows that the proposed method achieved a considerably higher recall rate than all the comparative methods in the zero-shot setting. The proposed method surpassed the performance of the previous state-of-the-art method [14] by 59.9% in terms of $zR@100$, which is a significant advancement. Leveraging rich, diverse common sense knowledge in heterogeneous KGs enables the proposed method to bring the visual concepts closer to their high-level semantics for predicting unseen visual relationships.

4) Comparison of KGs and Embedding Models: The recall rates obtained by the proposed method with ComplEx embeddings of different KGs are shown in Table III. Due to its heterogeneous nature and broader coverage of common sense knowledge, CSKG achieved a significantly higher recall rate compared to ConceptNet and WordNet. The recall rates obtained by the proposed method with different embedding models for CSKG are shown in Table IV. Due to their capability to represent complex-valued and multidimensional relationships between entities in CSKG, ComplEx and DistMult achieved higher recall rates than TransE and RESCAL in the same setting, meaning that ComplEx and DistMult are more expressive and better suited for zero-shot relationship retrieval in SGG. ComplEx achieved the highest performance compared to the rest of the embedding methods.

C. Limitations and Future Directions

Heterogeneous KGs are currently the richest and most diverse sources of common sense knowledge, as they capture detailed structural and semantic features of general concepts in the world. However, KGs have limited contextual knowledge [18], which can be limiting if the heterogeneous KGs lack contextually valid information about visual concepts in a specific scene. The proposed method performed consistently for most of the relationship predicates. Still, it could not disambiguate a few complex relationship predicates, which could be attributed to the limited effectiveness of the background knowledge in the KG for handling such cases. To address this challenge, a more sophisticated relationship reasoning approach [53] incorporating linguistic and visual features in addition to KG embeddings may be necessary. This work focuses on relationship retrieval, which is quite a challenging task within SGG, even in supervised settings, due to the high complexity of the problem. The proposed method has a considerably higher recall rate as compared to the previous state-of-the-art method, but the recall rate is still quite limited for its practical use. Developing a full-fledged generalized SGG approach requires more accurate and robust zero-shot relationship prediction as well as unseen object detection; therefore, it remains an open research problem. Additionally, knowledge transfer and distillation techniques [54], [55] present another promising direction in which the knowledge of previously seen visual relationships can be used to guide unseen relationship prediction, enhancing the generalization capabilities and practicality of SGG in real-world scenarios.

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART SGG METHODS IN ZERO-SHOT SETTING ON VG TEST SET

Method	**Approach**	**Knowledge Source**	**zR@K**		
			K=20	**K=50**	**K=100**
Proposed Method	Heterogeneous knowledge-based relationship retrieval for SGG	CSKG	**14.22**	**25.43**	**35.65**
TISGG [39]	Triple calibration and reduction for zero-shot SGG	-	13.60	20.20	22.30
COACHER [14]	Visual relationship prediction via knowledge graph mining	ConceptNet	13.42	19.31	22.22
MOTIF [13]	Stacked neural motif networks for scene graph generation	Statistical Prior	13.05	19.03	21.98
XKGC [11]	Visual relationship classification with knowledge graph completion	-	12.61	18.82	21.94
BGTNet [33]	Bidirectional GRU transformer for scene graph generation	-	12.23	18.31	21.51
RGNN [10]	Relational graph neural network for scene graph generation	-	12.59	18.38	21.21
KBGAN [15]	Knowledge-based feature refinement for SGG	ConceptNet	12.35	18.10	21.13
IMP [34]	Scene graph generation via iterative message passing	-	12.17	17.66	20.25
MIL-SGG [40]	Visual relationship reasoning based on informativeness	-	9.33	14.43	-
TDE [29]	Total direct effect for unbiased scene graph generation	-	-	14.40	18.20
QuatRE [41]	Quaternion relation embedding for expressive SGG	-	-	11.91	15.16
VCTree [35]	Dynamic visual context tree model for SGG	-	-	10.80	14.30
FCSGG [36]	Relation affinity fields-based fully convolutional SGG	-	-	8.60	10.90
SHAGCL-DD [43]	Causality-guided loss for unbiased SGG	-	5.88	7.52	8.50
DeepVRL [12]	Variation-structured Q-learning using deep Q-network	Language Prior	-	7.14	6.27
T-CAR [44]	Unseen triple calibration and reduction	-	3.2	4.7	6.0
LLM4SGG [45]	Weakly supervised SGG via LLMs	-	-	2.2	3.02
USGG [37]	Union message based SGG architecture	-	-	1.0	1.8
RU-Net [38]	Regularized unrolling network for SGG	-	0.36	0.73	1.15

TABLE III
COMPARISON BETWEEN DIFFERENT KGS USING COMPLEX EMBEDDINGS ON VG AND GQA TEST SETS

KG	**zR@100 (VG)**	**zR@100 (GQA)**
CSKG [23]	**35.65**	**29.56**
ConceptNet [19]	26.33	21.02
WordNet [22]	15.39	13.57

TABLE IV
COMPARISON BETWEEN DIFFERENT KG EMBEDDINGS OF CSKG ON VG AND GQA TEST SETS

KG Embedding Model	**zR@100 (VG)**	**zR@100 (GQA)**
ComplEx [27]	**35.65**	**29.56**
DistMult [52]	34.12	28.71
TransE [50]	30.33	25.63
RESCAL [51]	29.78	24.89

V. CONCLUSION

We presented a novel zero-shot relationship retrieval method that leverages a heterogeneous knowledge graph for generalized SGG, addressing critical limitations arising from imbalanced datasets. Our neurosymbolic method combines data-driven object detection with knowledge-based relationship retrieval, refining predictions based on positional and structural similarity. Evaluation on the VG benchmark dataset shows a 59.96% improvement in zero-shot recall rate, demonstrating that the proposed method is simple yet effective in handling data imbalances, predicting unseen relationships, and generating generalized scene graphs. The object refinement step effectively improved the object detection performance by 57.1%. ComplEx embeddings proved to be well-suited for generalized SGG. Additional experimental evaluation using the GQA dataset depicts the generalizability of the proposed method across datasets. Future work includes exploring use cases in downstream reasoning tasks and investigating more sophisticated approaches based on graph mining and multihop KG reasoning for zero-shot relationship retrieval.

REFERENCES

[1] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, “Scene graphs: A survey of generations and applications,” 2021, *arXiv:2104.01111*.

[2] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, and S. Günnemann, “Graphhopper: Multi-hop scene graph reasoning for visual question answering,” in *Proc. Int. Semantic Web Conf.*, Springer, 2021, pp. 111–127.

[3] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, “Comprehensive image captioning via scene graph decomposition,” in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 211–229.

[4] B. Schroeder and S. Tripathi, “Structured query-based image retrieval using scene graphs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 178–179.

[5] 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 Things J.*, vol. 9, no. 15, pp. 13705–13724, Aug. 2022.

[6] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, “Action genome: Actions as compositions of spatio-temporal scene graphs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10236–10247.

[7] C. Agia et al., “Taskography: Evaluating robot task planning over large 3d scene graphs,” in *Proc. Conf. Robot Learn.*, PMLR, 2022, pp. 46–58.

[8] T. Tahara, T. Seno, G. Narita, and T. Ishikawa, “Retargetable AR: Context-aware augmented reality in indoor scenes based on 3D scene graph,” in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 249–255.

[9] R. Krishna et al., “Visual genome: Connecting language and vision using crowdsourced dense image annotations,” *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[10] X. Yu, J. Li, S. Yuan, C. Wang, and C. Wu, “Zero-shot scene graph generation with relational graph neural networks,” in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 1894–1900.

[11] X. Yu et al., “Zero-shot scene graph generation with knowledge graph completion,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 1–6.

[12] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 848–857.

[13] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5831–5840.

[14] X. Kan, H. Cui, and C. Yang, "Zero-shot scene graph relation prediction through commonsense knowledge integration," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Berlin, Germany: Springer, 2021, pp. 466–482.

[15] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1969–1978.

[16] A. Zareian, S. Karaman, and S.-F. Chang, "Bridging knowledge graphs to generate scene graphs," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany: Springer, 2020, pp. 606–623.

[17] M. J. Khan, J. G. Breslin, and E. Curry, "Common sense knowledge infusion for visual understanding and reasoning: Approaches, challenges, and applications," *IEEE Internet Comput.*, vol. 26, no. 4, pp. 21–27, Jul./Aug. 2022.

[18] A. Ettorre, A. Bobasheva, C. Faron, and F. Michel, "A systematic approach to identify the information captured by knowledge graph embeddings," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol.*, 2021, pp. 617–622.

[19] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4444–4451.

[20] M. J. Khan, F. Ilievski, J. G. Breslin, and E. Curry, "A survey of neurosymbolic visual reasoning with scene graphs and common sense knowledge," *Neurosymbolic Artif. Intell.*, no. Pre-press, pp. 1–24, 2024.

[21] M. Sap et al., "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3027–3035.

[22] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[23] F. Ilievski, P. Szekely, and B. Zhang, "CSKG: The commonsense knowledge graph," in *Proc. Eur. Semantic Web Conf.*, Berlin, Germany: Springer, 2021, pp. 680–696.

[24] 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 *Proc. 35th AAAI Conf. Artif. Intell.*, 2021.

[25] M. J. Khan, J. G. Breslin, and E. Curry, "Expressive scene graph generation using commonsense knowledge infusion for visual understanding and reasoning," in *Proc. Eur. Semantic Web Conf.*, Berlin, Germany: Springer, 2022, pp. 93–112.

[26] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6700–6709.

[27] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, 2016, pp. 2071–2080.

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

[29] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3716–3725.

[30] A. Prakash et al., "Self-supervised real-to-sim scene generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 16044–16054.

[31] X. Lin, C. Ding, Y. Zhan, Z. Li, and D. Tao, "Hl-Net: Heterophily learning network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19476–19485.

[32] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Boosting scene graph generation with visual relation saliency," *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, 2022.

[33] N. Dhingra, F. Ritter, and A. Kunz, "bgt-net: Bidirectional GRU transformer network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2150–2159.

[34] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5410–5419.

[35] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6619–6628.

[36] H. Liu, N. Yan, M. Mortazavi, and B. Bhanu, "Fully convolutional scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11546–11556.

[37] S. Sun, D. Huang, Z. Qin, X. Tao, C. Pan, and G. Liu, "USGG: Union message based scene graph generation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Piscataway, NJ, USA: IEEE Press, 2023, pp. 2575–2579.

[38] X. Lin, C. Ding, J. Zhang, Y. Zhan, and D. Tao, "RU-NET: Regularized unrolling network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19457–19466.

[39] Q. Di, W. Ma, Z. Qi, T. Hou, Y. Shan, and H. Wang, "Towards unseen triples: Effective text-image-joint learning for scene graph generation," 2023, *arXiv:2306.13420*.

[40] A. Goel, B. Fernando, F. Keller, and H. Bilen, "Not all relations are equal: Mining informative labels for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15596–15606.

[41] Z. Wang, X. Xu, G. Wang, Y. Yang, and H. T. Shen, "Quaternion relation embedding for scene graph generation," *IEEE Trans. Multimedia*, vol. 25, pp. 8646–8656, 2023.

[42] Y. Li, X. Yang, X. Huang, Z. Ma, and C. Xu, "Zero-shot predicate prediction for scene graph parsing," *IEEE Trans. Multimedia*, vol. 25, pp. 3140–3153, 2023.

[43] R. Peng et al., "A causality guided loss for imbalanced learning in scene graph generation," *Neurocomputing*, 2024, Art. no. 128042.

[44] J. Li, Y. Wang, and W. Li, "Zero-shot scene graph generation via triplet calibration and reduction," *ACM Trans. Multimedia Comput., Commun. Appl.*, 2023.

[45] K. Kim et al., "Weakly supervised fine-grained scene graph generation via large language model," 2023, *arXiv:2310.10404*.

[46] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2016.

[47] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2117–2125.

[48] F. Ilievski et al., "KGTK: A toolkit for large knowledge graph manipulation and analysis," in *Proc. Int. Semantic Web Conf.*, Berlin, Germany: Springer, 2020, pp. 278–293.

[49] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Visi.*, Berlin, Germany: Springer, 2016, pp. 852–869.

[50] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[51] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011.

[52] B. Yang, W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," 2014, *arXiv:1412.6575*.

[53] X. Li et al., "HAPZSL: A hybrid attention prototype network for knowledge graph zero-shot relational learning," *Neurocomputing*, vol. 508, pp. 324–336, 2022.

[54] X. Yang, H. Zhang, and J. Cai, "Auto-encoding and distilling scene graphs for image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2313–2327, May 2022.

[55] J. Peyre, I. Laptev, C. Schmid, and J. Sivic, "Detecting unseen visual relations using analogies," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1981–1990.

M. Jaleed Khan received the Ph.D. degree in artificial intelligence from the University of Galway, Galway, Ireland, in 2024.

He is a Senior Researcher with AI and Data Analytics, Fujitsu Research, London, and an Honourary Fellow with the University of Oxford, Oxford, U.K. His research has resulted in over 40 peer-reviewed publications and several book chapters and open source projects. He is actively involved in the AI Research Community as a PC Member of top-tier conferences (ECAI and ECML), a Reviewer for journals (IJCV and TNNLS), a Professional Member of ACM and IAPR and a Grant Panellist for funding bodies (FFG and NCN). His research interests include artificial intelligence, multimodal learning, neurosymbolic reasoning, machine learning, and computer vision.

John G. Breslin (Senior Member, IEEE) is a Professor of electronic engineering, the National University of Ireland Galway, Galway, Ireland. Associated with three Taighde Éireann - Research Ireland Centres, he is a Principal Investigator with Insight (Data Analytics), University of Galway, Galway, and EDIH Data2Sustain, University of Galway, and a Funded Investigator with VistaMilk (AgTech) University of Galway. With a h-index of 50, over 12,000 citations, and various Best Paper Awards, he has jointly written over 300 peer-reviewed academic publications, including books on *The Social Semantic Web and Social Semantic Web Mining*. He co-created the SIOC framework, implemented in hundreds of applications (by Yahoo, Boeing, Vodafone, etc.) on at least 65,000 websites with 35 million data instances. He has featured in 275 mainstream media items including articles from CNN, Forbes, New Scientist, Washington Post, and France 24, and has been on 35 panels and given over 80 talks at venues including Stanford, Tsinghua University, and the Library of Congress. He co-authored the Irish bestsellers Old Ireland in Colour, Old Ireland in Colour 2 and Old Ireland in Colour 3.

Prof. Breslin is a Co-Founder of boards.ie (Ireland's largest discussion forum website; Wikipedia article), adverts.ie (classified ads website), and StreamGlider (real-time streaming newsreader app). He has won two IIA Net Visionary Awards, an ITAG Outstanding Contribution to the ICT Sector Award, a Galway Chamber President's Award, and a Best Irish-Published Book Award. He is a Co-Founder of the PorterShed (Galway City Innovation District), and serves on the Steering Group of Scale Ireland. He is Cathaoirleach (Chair) of Gaillimh le Gaeilge, a Former Director and Past Chair of WestBIC, and the Director Emeritus of the American Council of Exercise. He undertook the Entrepreneurship Development Program with Bill Aulet at MIT, in 2017. He also runs the Breslin Archive, since 2019 and the Tomita Fansite, since 1995.

Edward Curry is a Professor of computer science, the National University of Ireland Galway, Galway, Ireland. He has made substantial contributions to semantic technologies, incremental data management, event processing middleware, software engineering, as well as distributed systems, and information systems.

Prof. Curry combines strong theoretical results with high-impact practical applications. The excellence and impact of his research have been acknowledged by numerous awards, including best paper awards and the University of Galway President's Award for Societal Impact in 2017. His team's technology enables intelligent systems for smart environments in collaboration with several industrial partners. He is an Organiser and the Programme Co-Chair of major international conferences, including CIKM 2020, ECML 2018, IEEE Big Data Congress, and European Big Data Value Forum. He is a Co-Founder and elected Vice President of the Big Data Value Association, an Industry-led European Big Data Community, has built consensus on a joint European big data research and innovation agenda, and influenced European data innovation policy to deliver on the agenda.