Common Sense Knowledge Infusion for Visual Understanding and Reasoning: Approaches, Challenges, and Applications

Muhammad Jaleed Khan, John G. Breslin, and Edward Curry, *Data Science Institute, National University of Ireland, Galway, H91TK33, Ireland*

Visual understanding involves detecting objects in a scene and investigating rich semantic relationships between the objects, which is required for downstream visual reasoning tasks. The scene graph is widely used for structured scene representation, however, the performance of the scene graph generation for visual reasoning is limited due to challenges posed by imbalanced datasets and insufficient attention toward common sense knowledge infusion. Most of the existing approaches use statistical or language priors for knowledge infusion. Common sense knowledge infusion using heterogeneous knowledge graphs can help in improving the accuracy, robustness, and generalizability of the scene graph generation and enable explainable higher level reasoning by providing rich and diverse background and factual knowledge about the concepts in visual scenes. In this article, we present the background and applications of the scene graph generation and the initial approaches and key challenges in common sense knowledge infusion using heterogeneous knowledge graphs for visual understanding and reasoning.

Visual understanding and reasoning is an essential part of artificial intelligence that is inspired by the ability of humans to understand, interpret, and reason about everyday visual scenes. The advancements in deep learning enabled the low-level semantic tasks in visual understanding, including image classification, object detection, and localization, and image segmentation to achieve major breakthroughs and near human-like performance. In addition to object detection and localization, the higher level reasoning tasks, such as image captioning, visual question answering (VQA), multimedia event processing (MEP), content-based image retrieval, and image generation, require the prediction of rich semantic relationships between objects in a scene. Numerous vision-language hybrid approaches have been developed for this purpose in the past decade. The scene graph[1] has emerged as a widely used structured semantic representation model of visual scenes in which objects are represented as nodes and the pairwise relationships between objects are represented as edges of a knowledge graph (KG). Many visual understanding and reasoning techniques use scene graphs to represent visual scenes and perform downstream reasoning for various applications. The performance of the downstream reasoning tasks is dependent on the efficacy of the scene graph generation (SGG) in the earlier stage, which requires accurate and robust prediction of the objects and pairwise relationships between the objects in a scene. However, SGG faces major challenges due to several factors, including the unbalanced and biased distributions of objects and relationship predicates in the training datasets and the dependence of object detection and relationship prediction models on training data.

Humans rely on implicit common sense knowledge for making sense of everyday scenes, similarly, common sense knowledge from various sources in AI has benefited language processing[2] and holds a promise to aid visual understanding and reasoning as well. In order

Digital Object Identifier 10.1109/MIC.2022.3176500
Date of publication 20 May 2022; date of current version 19 July 2022.

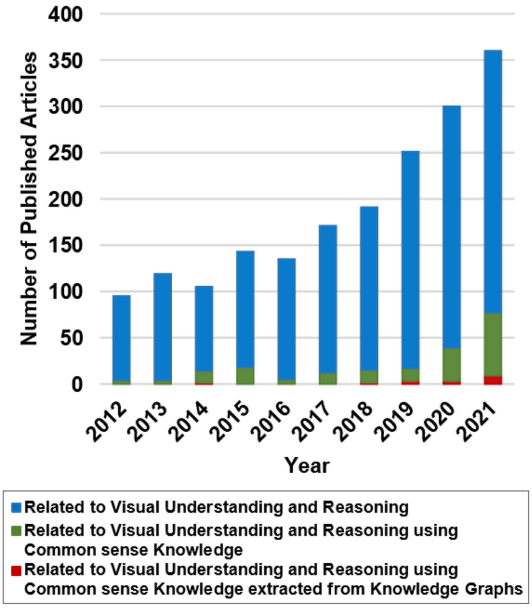

FIGURE 1. Past decade witnessing an increasing interest in visual understanding and reasoning research with increasing use of common sense knowledge from various sources and few recent works leveraging common sense knowledge extracted from KGs (data collected from Google Scholar Advanced Search).

to address the challenges posed by the long-tailed distribution problem and to improve the relationship prediction performance in SGG, numerous techniques on multimodal learning, efficient training procedures, and different ways to infuse prior knowledge have been proposed in the past decade. Most of the existing approaches use prior knowledge from statistical priors or language priors, however, the heuristics of the statistical priors do not generalize well and the limitations of semantic word embeddings affect the performance of language priors in the case of infrequent or unseen relationships. The infusion of rich and diverse common sense knowledge in the form of explicit semantics and factual knowledge from heterogeneous KG is a promising approach because it can alleviate the bias towards generic and frequently occurring relationships and give equal significance to infrequent but important relationships. However, there is a lack of attention toward common sense knowledge infusion from heterogeneous KGs in visual understanding and reasoning research. Figure 1 shows the increasing research interest in visual understanding and reasoning with an increasing number of publications focusing on common sense knowledge infusion, and only a few works leveraging KGs.

In this article, we have discussed the prominent role of SGG in visual understanding by reviewing the latest approaches and applications of SGG. We have also reviewed the SGG approaches involving common sense knowledge infusion based on statistical and language priors. We argue about the potential and need of attention toward common sense knowledge infusion in SGG based on heterogeneous KGs, which will help in extending the accuracy and robustness of relationship prediction and improving the performance and interpretability of the downstream visual reasoning tasks. Moreover, we have identified and presented the key challenges in relationship prediction and knowledge infusion in SGG based on the limitations of the existing approaches and sources.

SCENE GRAPH GENERATION

The scene graph is a structured representation that captures the semantics of a visual scene, such as objects and pairwise relationships, between the objects, and represents them in a graphical form. SGG techniques generally follow a bottom-up approach (see Figure 2) in which objects are detected and localized using object detectors and the pairwise relationships between the objects are predicted by leveraging vision-language hybrid features of the objects; triplets are formed by linking these semantic elements, which are then connected to generate the scene graph. The most challenging task in SGG is the prediction of pairwise visual relationships between objects, which has attracted a lot of interest in this research area.[1] Generally, a region proposal network is employed to generate triplet proposals (ROIs of an object, subject, and relationship pair) from input images. Subsequently, the multimodal features of each proposal, including object features, region features, and language features are encoded and fused together. Attention-based or message-passing approaches are used to refine the feature representations, followed by classification into object and relationship categories and construction of the scene graph.

Approaches

SGG is an active research topic and a variety of related approaches have been proposed. The earlier feature representation methods in SGG are focused on multimodal vision-language feature extraction approaches, while some of the current approaches also leverage common sense knowledge from statistical or language priors to extract complementary features, as shown in Table 1. In addition, numerous approaches have been proposed for feature refinement in SGG, including message passing, attention-based, and visual translation embedding approaches, etc. A variety of state-of-the-art

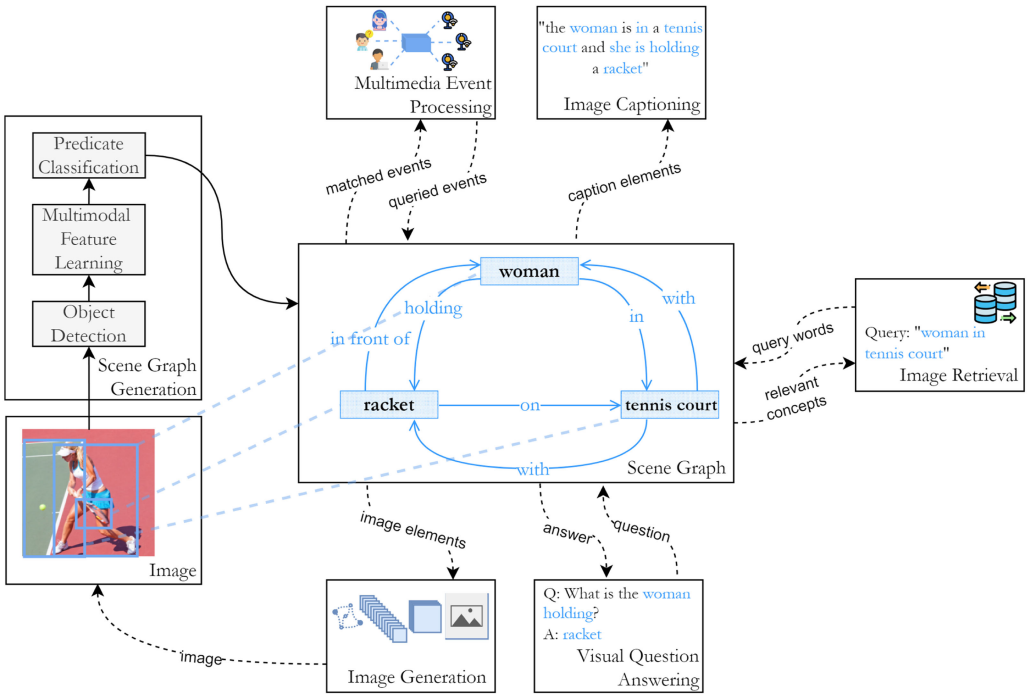

FIGURE 2. General workflow of SGG and its applications in visual understanding and reasoning, including image captioning, VQA, MEP, image generation, and image retrieval.

deep learning networks are used in SGG. The graph-based representation in SGG suits the architecture of graph neural networks (GNNs), which are used in attention-based approaches to integrate attention modules in the graph structure for the identification of salient regions for object and relationship prediction.[3,4]

Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) are actively employed in SGG because these networks capture the dependencies in data and contextual information of the objects, which are crucial for relationship prediction in SGG.[5] Convolutional neural networks are most commonly used in SGG

TABLE 1. Summary of the existing SGG approaches using common sense knowledge infusion.*

SGG approach	Technique	Common Sense knowledge Source	Recall@20/ 50/100 (%)	Challenges	Application
KERN[3]	GNN with gated mechanism for counting co-occurrence probability of relationships	Statistical prior	- / 27.1 / 29.8	1	SGG only
Neural Motifs[5]	RNN-LSTM-based stacked motif networks	Statistical prior	21.7 / 27.3 / 30.5	1	SGG only
DR-Net[13]	Deep relational network exploiting object-relationship dependencies	Statistical prior	- / - / -	1	SGG only
Lu *et al.*[6]	RCNN followed by relationship prediction using semantic word embeddings	Language prior	- / - / -	2	Image retrieval
Liang *et al.*[14]	Variation-structured reinforcement learning (VRL)	Language prior	- / 13.34 / 12.57	2	SGG only
KB-GAN[15]	Common sense and reconstruction-based object and phrase refinement	KG: ConceptNet	- / 13.65 / 17.57	3,7	Image generation
IRT-MSK[7]	Instance relation transformer with multiple structured knowledge	KG: ConceptNet, VG	22.2 / 27.2 / 31.2	3,4,7	SGG only
GB-Net[4]	Message passing between scene graphs and common sense graph	KG: ConceptNet, WordNet, VG	- / 29.4 / 35.1	3,4,5,6,7	SGG only

*Results are reported on the standard VG dataset. Current challenges are linked from the list of challenges on the "Challenges in Knowledge Infusion in SGG" section.

for extracting the global and local images features required for classification of relationships between object pairs.[6] Moreover, the state-of-the-art transformer models are also employed in SGG.[7] The infusion of common sense knowledge from various sources helps in the prediction of relationships and ensures efficient and accurate SGG. The existing approaches of knowledge infusion for improved relationship prediction in SGG and the sources used for this purpose are further discussed in the next section.

Applications

The common applications of SGG in visual understanding and reasoning, including image captioning, VQA, MEP, image retrieval, and image generation are illustrated in Figure 2 and briefly discussed in this section.

- *Image captioning* methods utilize the semantic relationships between objects in scene graphs to generate accurate language descriptions of the scene, which is difficult to achieve with only visual features of the scenes. Based on the idea of abstraction of scenes into symbols for providing a clear path for the generation of text descriptions, Chen *et al.*[8] proposed abstract scene graph that identifies and makes use of user's intentions in addition to semantics in scene graphs to generate desired as well as diverse image captions.
- *Visual question answering (VQA)* involves multimodal feature learning which leverages the essential semantic information in scene graphs. For example, Ziaeefard *et al.*[9] proposed a graph attention networks-based approach to encode scene graphs and related knowledge from ConceptNet for VQA.
- *Multimedia event processing (MEP)* uses graph-based approaches for representing multimedia streams for real-time event processing in the middleware for the Internet of Multimedia Things.[10] MEP approaches use graph-based semantic models for representing video streams; deep learning models are used to detect objects and symbolic rules are employed to identify relationships between objects, which are required for matching high-level video events queried by users.
- *Image retrieval* tasks use scene graphs to precisely describe the semantics of images to ensure interpretable and open content-based retrieval of images. For example, Schroeder *et al.*[11] proposed structured query-based image retrieval that uses structured queries (instead of text) and models visual relationships in scene graphs as directed subgraphs for the graph matching task in image retrieval based on scene graph embeddings.
- *Image generation* from the scene graph representations of visual scenes is a promising application because it is more robust and flexible as compared to image generation from textual scene descriptions, which struggles with maintaining its performance with the increasing number of objects and their interactions in the text.[12]

Challenges in SGG

SGG has remained an active research topic in visual understanding research and numerous approaches[3–7,13–15] have been proposed by researchers in this field to address the limitations of SGG during the past decade, however, there are still significant efforts required to mitigate the existing challenges for effective use of SGG in the downstream reasoning tasks. The major challenges in SGG due to limitations of the existing approaches are summarized as follows.

- *Imbalanced training datasets* is one of the key challenges in SGG. The relationship predicates are highly imbalanced in the training datasets; a large number of predicates have only a few instances in the datasets. As a result, it is very challenging to effectively learn representations of the rarely occurring relationship predicates.
- *Accurate and robust relationship prediction* remains a challenging part of SGG because the relationships have a wider semantic space than the objects as they comprise different object pairs and the training datasets do not provide enough samples for all relationships. Due to the long-tailed distribution problem in training datasets, most of the relationships only include the common relationship predicates (e.g., "in," "has," "on," etc.), which limits the accuracy and expressiveness of SGG by ignoring the more descriptive predicates (e.g., "person holding racket" and "person lying on beach" are more expressive than "person has racket" and "person on beach") that are more useful in visual understanding and reasoning tasks.
- *A huge number of possible relationships* is possible if there are a large number of objects and predicates because a relationship is a combination of two objects and a predicate. The machine learning (ML) models for classification and detection require limited categories due to which the traditional approach of object detection followed by relationship prediction would be inefficient for SGG.
- *Relationship prediction between distant objects* is unexplored in SGG. The current SGG techniques predict relationships between closely

located objects in scenes. This is mainly because the currently available datasets only include small-scale images that mostly cover closely located objects only.

- *Prediction of time-varying relationships in videos* is an emerging problem in SGG for videos. The existing techniques only focus on instantaneous relationships between objects, however, visual relationships in videos can have time-varying patterns in addition to the spatial patterns.

COMMON SENSE KNOWLEDGE INFUSION IN SGG

Common sense knowledge is essential for visual understanding and reasoning because it stimulates the common sense reasoning process. Some of the latest SGG approaches have employed common sense knowledge in the form of prior knowledge based on statistical priors[3] and language priors[14] in an effort to address the challenges in SGG. A few recent approaches have utilized background knowledge and related facts from KGs as common sense knowledge for relationship reasoning in SGG.[4,7,15] The existing approaches are summarized in Table 1.

Approaches Based on Statistical and Language Priors

Statistical priors, commonly used as prior knowledge, aim to model the statistical correlations between object pairs and relationships. Chen *et al.*[3] proposed knowledge-embedded routing network (KERN), which uses a structured graph to represent the statistical knowledge and integrates it into the deep propagation network as supplementary information, which minimizes the uncertainty in prediction by regularizing the distribution of potential relationship triplets. In addition, the statistical correlations between relationship triplets are leveraged for SGG using deep relational network[13] and LSTM-based approaches.[5]

Language priors are also used to guide relationship prediction in SGG by leveraging the semantic relationships of words. These approaches use semantic word embeddings,[6] priori predicate distribution and compact semantic associations in language priors[14] for relationship prediction using different multimodal learning approaches.

KGs as Common Sense Knowledge Source

ML models leverage the explicit semantics and factual knowledge in KGs as common sense knowledge, which improves the performance and robustness of the models.[16] The infusion of common sense knowledge using KGs enhances the reasoning capabilities of the models by improving their interpretability.[17] In addition, this also enables the models to alleviate the bias toward generic and frequently occurring concepts and give equal significance to infrequent but important concepts, which improves the recall of the models while maintaining precision.[18] The scale of infusion of common sense knowledge varies from shallow to deep infusion in ML models. The use of KGs as a common sense knowledge source within the state-of-the-art neuro-symbolic approaches[19] is a promising research direction in visual understanding and reasoning. SGG techniques can benefit from the related facts and background knowledge of visual concepts in effectively capturing and interpreting detailed semantics in images. This can improve the performance of relationship prediction in SGG as well as the downstream reasoning tasks for different applications. Several knowledge bases have been developed to store common sense knowledge, such as related facts and background knowledge, in various forms as concepts or entities, attributes, and relationships between the concepts.

Approaches Based on KGs

Most of the techniques in visual understanding and reasoning extract relevant facts from a knowledge source and embed them within the ML model at a certain stage.[15] The recent graph-based approaches use message passing to embed the structural information from the source in the representations of the model.[4] The knowledge bases covering different domains and contexts of common sense knowledge can be leveraged in a consolidated form [such as the CommonSense Knowledge Graph (CSKG)[20]] as a unified, rich, and heterogeneous source of common sense knowledge. For example, GB-Net[4] links the entities and edges in a scene graph to the corresponding entities and edges in a common sense graph extracted from Visual Genome (VG), WordNet, and ConceptNet, and iteratively refines the scene graph using GNN-based message passing. Similarly, Guo *et al.*[7] employed an instance relation transformer to extract relational and common sense knowledge from VG and ConceptNet for SGG. However, the potential of consolidated KGs in visual understanding and reasoning needs to be explored in more depth, which will help in mitigating the existing challenges and trigger more practical applications of visual understanding and reasoning.

Challenges in Knowledge Infusion in SGG

While the investigation of common sense knowledge infusion from language priors, statistical priors, and

KGs is highly invaluable for mitigating the existing challenges, the limitations of priors, effective acquisition, efficient extraction and integration, and full utilization of common sense knowledge emerge as challenging research problems in this direction. The key challenges due to limitations of the existing approaches and sources are listed in the following.

1) *Limitations of statistical priors:* A variety of the existing approaches use prior knowledge from statistical priors for relationship prediction in SGG, however, the statistical priors mostly use heuristic approaches (such as co-occurrence probability of relationships predicates), which are hard-coded and not generalized.[3,5,13]
2) *Limitations of language priors:* The effectiveness of language priors for knowledge infusion in SGG can be affected by the limitations of semantic word embeddings, especially in generalizing to the infrequent objects in the datasets. Moreover, the visual appearance of relationship predicates can vary across different scenes, and semantically different relationship predicates can have a similar visual appearance.[6,14]
3) *Multihop relationship reasoning using the common sense KGs* needs to be explored in SGG because the existing approaches mostly integrate only triplets from the knowledge sources and ignore the rich structural information beyond individual triplets. For instance, the relationships between the pairs of objects that are uncommon in training datasets can be inferred by the use of semantically related facts and background knowledge from the common sense KGs.[4,7,15]
4) *The knowledge representation methods of different KGs are different*, for example, the same concept is represented in different KGs in different ways. The infusion of common sense knowledge from multiple sources is important for the sake of diversity and completeness of common sense knowledge, however, it introduces the challenge of flexibility and robustness to different knowledge representation methods.[4,7]
5) *The consolidated KGs can be quite noisy* apart from being rich and heterogeneous sources of common sense knowledge. Common sense knowledge infusion using such sources can infuse noise in the form of redundant, irrelevant, and incorrect triplets, which can affect the SGG performance.[4]
6) *The KG consolidation efforts can compromise the rich semantic knowledge provided by individual knowledge bases* in an attempt to create huge and heterogeneous sources of common sense knowledge. For example, CSKG retains only the structure of relationships between objects in VG during the consolidation and expresses all the relationship predicates taken from the VG knowledge base as a single "LocatedNear" predicate. This makes the consolidation simple but results in the loss of the important visual cues about spatial proximity or interactions between objects provided by the visual relationship predicates in VG, thus limiting the applicability of CSKG in visual understanding and reasoning.[4]
7) *The interpretability offered by KGs can be affected by the application of nonlinear ML methods* for visual understanding. Specialized strategies for knowledge infusion and ML need to be designed and adopted in order to preserve the interpretability of KGs to ensure explainable visual reasoning.[4,7,15]

The existing challenges indicate the need for the development of SGG techniques that can effectively learn representations for a large number of relationships from small amounts of training samples by leveraging the state-of-the-art efficient model training approaches, as well as, infusion of external common sense knowledge from new sources, such as KGs.

CONCLUSION

The visual understanding and reasoning tasks involve multimodal techniques for the prediction of visual components, followed by reasoning to predict higher level semantic events. As shown by numerous approaches based on prior knowledge in statistical and language priors, common sense knowledge plays an important role in fine-tuning relationship prediction for SGG. Despite its significant potential, only a few techniques have used KGs as a common sense knowledge source. In this article, we have discussed SGG as the mainstream image representation model in visual understanding and reasoning approaches, the applications of SGG, and the substantial research on common sense knowledge infusion in SGG. We argued about the potential and need for attention toward common sense knowledge infusion using heterogeneous KGs, which can extend the accuracy and robustness of SGG and improve the performance and interpretability of the downstream reasoning tasks by providing related, rich, and diverse factual and background information about the semantic elements in the scenes. We have identified the key challenges in relationship prediction and knowledge infusion in SGG. This is a promising and challenging research direction in visual understanding and reasoning.

ACKNOWLEDGMENTS

This work was supported by the Science Foundation Ireland under Grant 18/CRT/6223 and Grant 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CCBY public copyright licence to any author accepted manuscript version arising from this submission.

REFERENCES

1. G. Zhu *et al.*, "Scene graph generation: A comprehensive survey," 2022, *arXiv:2201.00443*.
2. K. Faldu, A. Sheth, P. Kikani, and H. Akbari, "Ki-Bert: Infusing knowledge context for better language and domain understanding," 2021, *arXiv:2104.08145*.
3. T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6163–6171.
4. A. Zareian, S. Karaman, and S.-F. Chang, "Bridging knowledge graphs to generate scene graphs," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 606–623.
5. 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.
6. C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 852–869.
7. Y. Guo, J. Song, L. Gao, and H. T. Shen, "One-shot scene graph generation," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3090–3098.
8. S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9962–9971.
9. M. Ziaeefard and F. Lécué, "Towards knowledge-augmented visual question answering," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1863–1873.
10. 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.*, early access, Jan. 14, 2022, doi: 10.1109/JIOT.2022.3143171.
11. 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.
12. J. Johnson, A. Gupta, and L. Fei-Fei, "Image generation from scene graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1219–1228.
13. B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3076–3086.
14. 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.
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. R. Wickramarachchi, C. Henson, and A. Sheth, "Knowledge-infused learning for entity prediction in driving scenes," *Front. Big Data*, vol. 4, p. 759110, 2021, doi: 10.3389/fdata.2021.759110.
17. F. Lecue, "On the role of knowledge graphs in explainable AI," *Semantic Web*, vol. 11, no. 1, pp. 41–51, 2020.
18. U. Kursuncu, M. Gaur, and A. Sheth, "Knowledge infused learning (K-IL): Towards deep incorporation of knowledge in deep learning," 2019, *arXiv:1912.00512*.
19. M. J. Khan and E. Curry, "Neuro-symbolic visual reasoning for multimedia event processing: Overview, prospects and challenges," in *Proc. ACM Int. Conf. Inf. Knowl. Manage. Workshops*, 2020.
20. F. Ilievski, P. Szekely, and B. Zhang, "CSKG: The common sense knowledge graph," in *Proc. Eur. Semantic Web Conf.*, 2021, pp. 680–696.

MUHAMMAD JALEED KHAN is a Ph.D. researcher in artificial intelligence with the Data Science Institute, National University of Ireland, Galway, H91TK33, Ireland. His research interests include multimedia analysis, image representation and reasoning, deep learning, computer vision, and hyperspectral image analysis. He is also a member of the Artificial Intelligence and Computer Vision (iVision) Lab at Institute of Space Technology, Islamabad, and the Pakistan Pattern Recognition Society (PPRS), an IAPR society. Contact him at m.khan12@nuigalway.ie.

JOHN G. BRESLIN is a personal professor in electronic engineering with the College of Science and Engineering, the National University of Ireland, Galway, H91TK33, Ireland. His research interests include electronic engineering, sensors, social semantics, social media, and semantic web. Contact him at john.breslin@nuigalway.ie.

EDWARD CURRY is the established professor of data science and a director of the Insight SFI Research Centre for Data Analytics and the Data Science Institute, National University of Ireland, Galway, H91TK33, Ireland. His research interests include distributed systems, middleware, event processing, data management, and smart cities. Contact him at edward.curry@nuigalway.ie.