





Knowledge Enhanced Zero-Shot Visual Relationship Detection

Nan Ding^{1,2}, Yong Lai^{1,2} , and Jie Liu^{1,2} 

¹ College of Computer Science and Technology, Jilin University,
Changchun 130012, China

² Key Laboratory of Symbolic Computation and Knowledge Engineering Ministry
of Education, Jilin University, Changchun 130012, China
dingnan22@mails.jlu.edu.cn, {laiy, liu_jie}@jlu.edu.cn

Abstract. In visual relationship detection (VRD), the diversity of relationships often results in many unseen (i.e. zero-shot) relationships in the test set. Predicting zero-shot relationships poses a significant challenge. Traditional methods often incorporate semantic knowledge and spatial features from images, which still rely on priors in language and annotations in the dataset. Therefore, we propose encoding spatial structure information in the knowledge graph and incorporating spatial relationships from commonsense to guide predictions. The model comprises two modules: logic tensor networks that encoded the negative domain of semantic and spatial knowledge, and a commonsense knowledge graph module updated by local spatial structure as positive domain semantic knowledge. Predictions are further constrained by region connection calculus (RCC). Experimental results demonstrate competitive performance on the Visual Relationship Datasets under the zero-shot setting and the entire subset of Visual Genome. In predicate detection, it achieves comparable results to benchmarks, while significantly outperforming benchmarks in relationship and phrase detection.

Keywords: Visual relationship detection · Zero-shot · Commonsense knowledge graph · Region connection calculus · Logic tensor networks

1 Introduction

Visual Relationship Detection (VRD) is a fundamental task in computer vision, expected to extract more detailed information from images than object detection [1]. Given an image, VRD identifies the relation types between object pairs in the form of $R(O_1, O_2)$. It is obvious that the diversity of objects and relations can blow up the number of predicted triples [2]. Accordingly, it is expensive to gather a dataset with each triple annotated in a normal scenario [2,3]. Zero-shot VRD refers to predicting the triples that were not observed during training. The approaches of zero-shot VRD have already found applications in many advanced computer vision tasks, such as image segmentation, image captioning, and human-object interactions [4].

An important approach to enhancing zero-shot VRD is to incorporate semantic knowledge, which can leverage the similarity with triples in the training set, or utilize high-level descriptions of relationships to supervise the predictions of neural networks. Lu et al. [1] utilized word embeddings to fine-tune the likelihood of a predicted relationship, while Yu et al. [5] encoded first-order logic knowledge into Markov logic networks as symbolic knowledge to correct the erroneous reasoning of neural networks. Donadello et al. [2] improved predictions by manually incorporating semantic knowledge into logic tensor networks. Wan et al. [6] and Zareian et al. [7] utilized commonsense knowledge graphs to supervise neural networks through iterative training and interaction with knowledge graphs. While manually modeling knowledge yields high accuracy and complements incomplete annotations [2], it is costly and limited. Introducing knowledge graphs can reduce costs, but existing methods typically focus only on the triples within a knowledge graph. The semantic correlation between adjacent relationships is also part of the semantic knowledge in the knowledge graph, and it is correlated with the topology between relationships [8]. For example, in predicting the relationship between *person* and *bike*, although there is no *person* – *hold* – *bike* in the knowledge graph, it can be inferred through a multi-hop path: *person* – *hand* – *hold* – *handle* – *bike*. Furthermore, it can be observed that for *hold*(*person*, *bike*), *person* – *hand* – *hold* is more important than *hand* – *hold* – *handle*. Our proposed model reduces manual workload by focusing on negative domain constraints and includes a commonsense knowledge graph module. We use multi-hop paths and different topological relationships in neighboring subgraphs to update the knowledge graph during training.



Fig. 1. Example of a bounding boxes in the picture.

Relying solely on semantic and visual features may cause the model to heavily rely on priors in language and datasets [9]. For example, in Fig. 1, the relationship between the blue box and red box is likely to be predicted as *ride*(*person*, *bike*) according to common triples. However, considering the spatial positions of the boxes, we can exclude *ride*. Therefore, adding spatial features has also been an

approach for zero-shot learning. Liu et al. [10] utilized information from different views of a picture for prediction. Chiou et al. [11] and Gkanatsios et al. [12] proposed attention modules to capture spatial features. Liang et al. [3] and Jung et al. [13] fused visual, semantic, and spatial position features as inputs to neural networks. However, these methods still rely on spatial relationships between annotated boxes in the dataset. Unlike these approaches that only consider spatial relationships within pictures, we encode common spatial relationships as external knowledge as well, and utilize RCC [14] to constrain predictions. Qualitative spatial calculus plays an important role in computer vision tasks [15], and RCC is one commonly used method in this domain. RCC is a formal model which can represent spatial entities as arbitrary plane regions based on a set of jointly exhaustive and pairwise disjoint relations [16]. Research on RCC has yielded theoretical and applied results on qualitative spatial representation and reasoning [17, 18]. We can use RCC to label different predicates and use the calculus between $RCC(O_1, O_2)$ and $RCC(O_2, O_3)$ to reason about the relationship between O_1 and O_3 .

Combining the issues addressed above, this paper introduces a model¹ consisting of two main modules: LTNs incorporating negative domains of semantic and spatial knowledge as background knowledge, and a knowledge graph constructed using positive domains of semantic knowledge. The knowledge graph is updated utilizing neighboring subgraphs and weights based on relationship topology. The final prediction is constrained by RCC. Experimental results demonstrate the competitiveness of the model on datasets.

2 Related Work

VRD is an important task that recognizes complex relationships and interactions among various objects in an image. As it serves as the basis for many important tasks, research on VRD has attracted increasing attention in recent years.

2.1 Visual Relationship Detection

Fundamentally, VRD is a task of categorizing predicates. Jung et al. [19] used a softmax model to classify predicates instead of the previous binary classifier, which improved the accuracy and efficiency of the model. In addition to enhancing classification accuracy, researchers have focused more on improving visual features. Russakovsky et al. [20] proposed that input from humans can be used to intervene in object detection. Xu et al. [21] introduced a method based on multi-scale context modeling, which enhances the accuracy of scene graph generation by fusing features at different scales. Peng et al. [22] used a method that combines dataset characteristics and inputs them into a fully-connected layer to obtain fused representation.

¹ The source codes is available at <https://github.com/laigroup/K-VRD>.

2.2 Zero-Shot Learning with Knowledge

The goal of VRD is to better understand the scene in a picture by recognizing the relationships between objects. Prediction effectiveness heavily relies on training. However, due to the large workload, VRD datasets annotations are incomplete and unbalanced. Thus, evaluating predictions on zero-shot triples is crucial. The performance of the model on zero-shot learning can be improved not only by processing the visual features but also by utilizing fusion knowledge wisely.

Encoding semantic knowledge into logical constraints is a common approach to embedding semantic information into neural networks, which creates neural-symbolic systems with enhanced perception and cognition [23]. Lu et al. [1] proposed a VRD model combined with a linguistic model. Through a semantic module using pre-trained word embeddings, objects are mapped to a vector space where objects with similar semantic features are closer. Yu et al. [24] applied knowledge distillation to transfer linguistic knowledge to a visual model by pre-training a language model on a large corpus. Donadello et al. [2] extended LTNs and applied them to the classification of bounding boxes and object detection. They encoded common knowledge as logical constraints, which were trained as the background knowledge of the LTNs. The introduction of logical knowledge solves the zero-shot problem more effectively. Manigrasso et al. [25] combined LTNs with a convolution module, achieving higher accuracy in object detection compared to traditional convolutional architectures. Buffelli et al. [26] drew inspiration from LTNs to build a logic constraint module that injects semantic knowledge into various neural networks for relationship detection.

Excessive semantic knowledge can lead to predictions that heavily rely on the frequency of relationships in the knowledge. Therefore, research has been conducted on methods for fusing visual, semantic, and spatial features. Gkanatsios et al. [12] proposed ATRNet, a scene graph generation model based on attention, which combines semantic features and spatial translations between regions to build a multi-scale attention mechanism. Gkanatsios et al. [27] introduced a method combining semantic and local contextual information, improving the model’s ability to represent relationships in a visual scene. Hu et al. [28] proposed a message-passing algorithm integrating linguistic prior and spatial features to propagate contextual information. Chiou et al. [11] explored the spatial module and mask attention module to capture spatial features.

Combining these methods to improve the model, we jointly train our model using visual information, semantic knowledge, and spatial knowledge. We encode spatial structure information in a knowledge graph and common spatial relationships to minimize the effect of priors on the datasets and semantic knowledge.

3 Preliminaries

3.1 Logic Tensor Network

LTNs combine tensor networks [29] with reasoning through first-order logic \mathcal{L} . We use grounding (denoted by \mathcal{G}) to interpret \mathcal{L} in the real world and the

Łukasiewicz t-parameter for the logical operator as in Ref. [2]. Grounded theory (GT) is a pair $\langle \mathcal{K}, \hat{\mathcal{G}} \rangle$, where \mathcal{K} is the knowledge base consisting of a set of clauses and $\hat{\mathcal{G}}$ is partial grounding. GT can be achieved by finding $\hat{\mathcal{G}}$ such that the truth values of the formulas in \mathcal{K} all close to 1. VRD is transformed into optimizing the parameters of LTNs to find grounding functions.

3.2 Knowledge Graph

A commonsense knowledge graph (CKG) is a knowledge graph constructed with a set of nodes \mathcal{N} and edges \mathcal{E} , where the nodes represent different semantic labels (entities or predicate classes) and the edges represent relationships between nodes. Each edge is labeled with a confidence value indicating the probability of the relationship between the connected nodes. If two concepts are not related, there is no edge between them.

3.3 RCC8

RCC is a model used for spatial qualitative representation and inference [14]. RCC8 [17] is the most prevalent subset of RCC, specifying eight basic relationships between two spatial regions. Figure 2 shows examples of these relationships. These relations are disjoint and cover all possible relationships between regions. Since the relationship between any two regions is for only one of RCC8 relationships, RCC8 can be used to represent definite knowledge between two spatial regions, while indefinite knowledge can be represented by the disjunction of possible relationships. The composition table for RCC8 relations can be seen in Ref. [17].

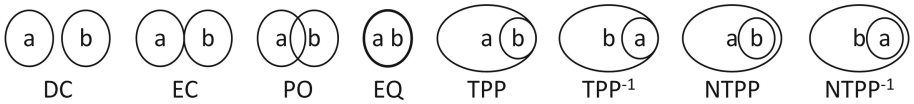


Fig. 2. Different relationships in RCC8.

4 Method

This section details our method, and the architecture is shown in Fig. 3.

4.1 Neural Network Prediction

As in Ref. [2], we use LTNs to encode VRD tasks. Let $\mathcal{T}_{RD} = \langle \mathcal{K}_{RD}, \mathcal{G}_{RD} \rangle$ be the grounded theory, where \mathcal{K}_{RD} is the knowledge base of LTNs. The knowledge

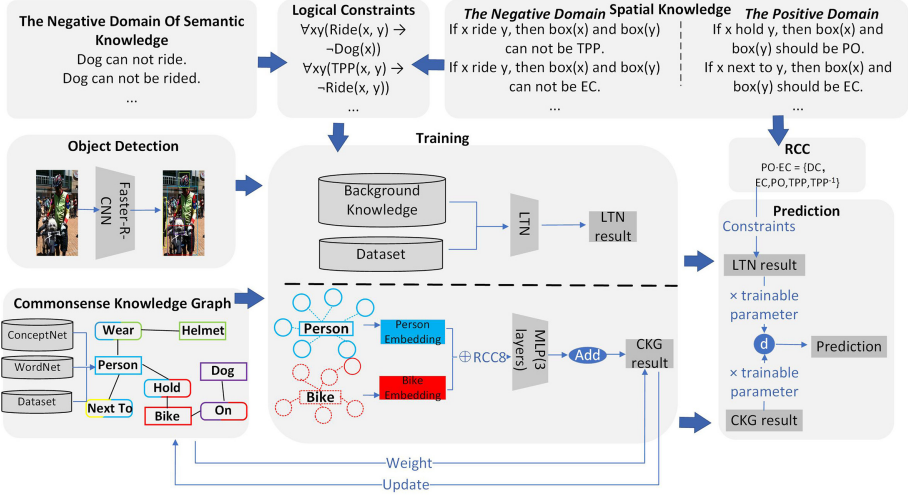


Fig. 3. An overview of the proposed framework.

base \mathcal{K}_{RD} includes two parts: positive and negative examples obtained by analyzing annotations of the dataset, and background knowledge. For background knowledge, we not only use semantic knowledge but also take spatial knowledge into account, which is one of our innovations. For ease of representation, the background knowledge is encoded in the form of negation of logical constraints. For semantic knowledge, we consider manually writing common sense as semantic knowledge. For spatial knowledge, we refer to RCC8 in Fig. 2 for encoding. By analyzing all the images in the training set, we can obtain the impossible RCC8 relations of predicates. For instance, for *talk*, the set of negative RCC8 relations is $\{EQ, EC, TPP, NTPP\}$. One corresponding representation is $\forall xy(EQ(x, y)) \rightarrow \neg \text{talk}(x, y)$.

After calculating the grounding value of the atomic formula, the logical constraints in the \mathcal{K}_{RD} are combined based on logical operators. However, the calculation of quantifiers is susceptible to special relations with a small number of samples, which can seriously affect the subsequent steps. Therefore, Ref [2] proposed using the harmonic mean to improve the calculation of quantifiers. Building upon this, we introduce the frequency of predicates to further minimize the impact.

4.2 Knowledge Graph Prediction

Our CKG is primarily derived from ConceptNet, WordNet, and datasets. We query the CKG for word embedding \mathbf{n} , and according to the positional relationship of bounding boxes, we use a one-hot vector to represent RCC8 relationships. Then, we concatenate these vectors to get the relationship embedding.

The relationship embedding is updated iteratively during training. The relationship between two related classes may involve a multi-hop path in CKG. The nodes in the path can influence the prediction of the relationship. Such multi-hop paths contain more information than one-hop paths. Therefore, during training, we obtain multi-hop paths between subjects and objects and segment them into closed subgraphs for iterative updating. In the process of constructing the subgraph, for each node in the subgraph, we label it with $\langle d_{sub}, d_{obj} \rangle$, where d_{sub} denotes the shortest path distance of the node from the subject (likewise for d_{obj}). The updating formula for node embedding is as shown in Eq. 1.

$$n = n_{init} + \sum_{n \in subgraph} \frac{2 \times h + 1 - d_{sub} - d_{obj}}{2 \times h} \prod_{end=sub, obj} \prod_{e \in path_{n, end}} w_e, \quad (1)$$

where h denotes the hop number of the subgraph, $path_{n, end}$ denotes the path from n to end , and w_e is the edge weight of e in the knowledge graph.

The topology in CKG may also affect the correlation between the relationships [8]. Therefore, we assign different weights to edges in the subgraph based on topological relationships. We mainly use four topologies: Head-Head, Tail-Tail, Not connected, and Parallel. Thus, we can derive the update formula for relationship embedding as shown in Eq. 2:

$$re_{update} = \sum_{e \in subgraph} \langle \mathbf{n}_{subject}^e : \mathbf{n}_{object}^e \rangle \times w_e \times w_{topo} \quad (2)$$

where e is the set of edges in the subgraph, and w_{topo} is the weight corresponding to the topological relationship.

4.3 Spatial Constraints

Based on common sense, we manually assign a spatial relationship in RCC8 to each predicate. When predicting the relationship, the neural network’s prediction is constrained based on the spatial relationships between the pairs of boxes related to subject and object. For example, in Fig. 1, when predicting the relationship between *person* and *bike*, we first obtain the relationships R_1 between *person* and other entities, and the relationships R_2 between corresponding entities and *bike*. We select a probability threshold θ and filter the sets R_1 and R_2 to obtain smaller sets r_1 and r_2 containing relationships with probabilities greater than θ . We pair the relations in the two sets to infer impossible relationships between *dog* and *bike*. We can obtain $next_to(O_{yellow}, O_{blue})$, $hold(O_{blue}, O_{red})$ from Fig. 1, and $RCC(next_to) = EC$, $RCC(hold) = PO$ from knowledge. Following composition, $RCC(O_{yellow}, O_{red})$ can’t be EC, which means $R(O_{yellow}, O_{red})$ can’t be ride. We constrain the neural network prediction results according to Eq. 3.

$$Loss = p_{r_1} \times p_{r_2} \times \log(\prod p_i), \text{ if } re_i \text{ in } re_{neg} \quad (3)$$

where p_{r_1} represents the probability of r_1 , p_i denotes the probability of the neural network predicting the relationship between subject and object, and re_{neg} is the

set of impossible relationships, as described in Sec. 1. At the end of each round of iterative training, the LTNs and CKG prediction results are combined to obtain the final prediction results with the trainable parameters as weights.

5 Evaluation

5.1 Evaluation Protocol

The experimental dataset used are Visual Relationship Datasets (VRDs) [1] and a subset of Visual Genome proposed in Ref. [21], called VG150. VRDs includes 5000 images containing 70 predicates and 100 entities, which comprise 37,993 visual relation instances and 6,672 triples types. There are 1,877 relationships that appear only in the test set, which are used to evaluate zero-shot learning. VG150 includes 108,703 images containing 50 predicates and 150 entities. It has 1,174,692 visual relation instances and 19,237 triples types in total.

Following Ref. [1] we use Recall as our performance metric, denoted as $Recall@x$ ($R@x$), which refers to the proportion of relationships that actually exist among the first x relationships predicted. We evaluate our method on the three classical visual relationship detection tasks proposed in Ref. [1]: predicate detection, phrase detection, and relationship detection.

5.2 Results

To validate the effectiveness of our method, we selected six methods for zero-shot prediction using either semantic or spatial information for comparison: LP [1] integrates visual and language features before detecting visual relationships. JVSE [30] proposes a model using the similarity between visual vectors and semantic vectors. Multi-view [10] proposes a multi-view image generation approach that utilizes spatial information to transfer the 2D visual space to the 3D multi-view space. IVDRC [6] utilizes a knowledge graph constructed from the dataset and iteratively updated by a neural network for prediction. DSR [3] introduces a framework, which facilitates the co-occurrence of relationships. LTN [2] is the first to propose the LTNs module for VRD. In predicate detection, the entity pairs are fixed. Therefore, spatial constraints do not work. Instead, we use word frequency to constrain the neural network predictions.

Table 1 shows the results of comparison with other state-of-the-art (SOTA) techniques that utilize semantic or spatial features for visual relationship detection. Phrase and relationship detection are more complex because they require object detection of the image, and the results of the object detector also affect the results of the two tasks. We use the object detection results provided in Ref. [1], detected by RCNN [31], the same object detection data used for JVSE and LTN. Faster-RCNN for object detection is used by others. Compare with the original method using LTN for VRD, our method performs better on all three tasks, especially predicate detection. Our method exploits knowledge from different domains, which can express more information and allow for more accurate

Table 1. Performances of different methods on VRDs under the zero-shot setting and entire VG150. We use “-” to indicate that the performance has not been reported in the original paper.

Datasets	VRDs-zero shot						VG150	
Task	Predicate Det.		Phrase Det.		Relation Det.		Predicate Det.	
Evaluation	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50
LP	—	—	3.57	3.36	3.52	3.13	—	—
JVSE	—	—	6.16	5.05	5.73	4.79	—	—
Multi-view	42.6	42.6	11.5	11.5	6.6	6.6	—	—
IVDRC	78.52	61.76	8.73	6.92	4.53	4.53	88.26	85.4
DSR	79.81	60.9	—	—	9.2	5.25	74.37	69.06
LTN	75.31	55.68	15.55	10.99	14.27	10.03	95.38	88.82
Our method	78.46	57.6	16.98	12.31	15.59	11.04	96.59	90.51

reasoning on the visual relationships. From Table 1 we can see, in VRDs under the zero-shot setting, for predicate detection, our framework achieves similar results to SOTA. In relationship detection and phrase detection, our method has achieved significant improvements. Our model demonstrates better fault tolerance than others in cases of inaccurate object detection. Such observations support the effectiveness of structure information in CKG and spatial knowledge in zero-shot relation prediction, which can compensate for the inaccuracy of object detection and indicate its superior generalizability. In the entire VG150, for predicate detection, our method outperforms LTN and SOTA. This demonstrates the effectiveness of our approach not only in zero-shot learning but also performs well in complex visual relationship detection accuracy after training.

5.3 Ablation Study

Both multi-hop information and topology positively impact predicting zero-shot triples, as shown in Table 2. We provide variants of our model in 2 dimensions: multi-hop paths in CKG and topology between relationships in CKG. Multi-hop paths can fully leverage the information of nodes and edges between entities pairs, incorporating path information into the prediction process. By assigning different weights to various relationships based on their topology, the latent spatial relationships in CKG can be effectively utilized. Table 2 demonstrates the effectiveness of our method in utilizing latent information in CKG.

To determine the contribution of each component in our method, we also compared different variants of the method, including: LTN: only the logical tensor network module; LTN-s: adding spatial knowledge to the knowledge base; LTN-s-w: adding weights of the knowledge based on word frequency; LTN-s-w-kg: adding the knowledge graph module; LTN-all: the complete method.

Table 2. Performances of different usages of CKG on VRDs under the zero-shot setting.

Task	Phrase Det.		Relationship Det.	
	R@100	R@50	R@100	R@50
LTN-ckg	16.53	11.64	14.99	10.52
LTN-no multi-hop	16.08	11.38	14.71	10.39
LTN-no topology	15.95	11.34	14.63	10.38

Table 3. Performances of different variants on VRDs under the zero-shot setting.

Task	Predicate Det.		Phrase Det.		Relationship Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
LTN	75.31	55.68	15.55	10.99	14.27	10.03
LTN-s	75.95	55.89	15.6	11.15	14.27	10.08
LTN-s-w	78.43	57.23	16.4	11.65	15	10.66
LTN-s-w-kg	78.46	57.57	16.9	12.27	15.48	11.02
LTN-all	78.46	57.57	16.98	12.31	15.59	11.04

From Table 3, each module shows a certain improvement in zero-shot learning. Adding spatial knowledge to the background knowledge of LTNs can enable neural networks to better distinguish the correspondence between visual relationships and spatial relationships, and to classify relationships from a spatial perspective. After weighting the knowledge in the knowledge base by word frequency, the more frequently occurring knowledge is more fully utilized, which correspondingly reduces the influence of uncommon knowledge, resulting in a more significant improvement. The second notable improvement occurs after adding a common knowledge graph. It incorporates more commonsense knowledge, enabling the network to learn more triples. Spatial constraints can modify the neural network’s incorrect predictions from the perspective of spatial commonsense. By analyzing the prediction results, we found that our method performs well in predicting spatial relationships. After incorporating the knowledge graph, some uncommon relationships are successfully predicted, such as, *on(hat, horse)*. The addition of spatial knowledge also allows for a better expression of spatial relationships. For example, not only can it predict *in_front_of(bear, tree)*, but it can also predict *behind(tree, bear)*. However, the effect on predicate detection is not satisfactory enough. This may be because introducing too much knowledge can affect the judgment of visual information on the results, which can be explored in future research to better integrate visual information with knowledge.

6 Conclusion

Due to the large scale of the relation detection dataset and the complexity of the images, annotating all images in the training set with complete detail is unfeasible. Consequently, zero-shot learning emerges as a critical metric in VRD. In this study, we address this challenge from both semantic and spatial perspectives. Semantic knowledge primarily stems from multi-hop information and the topology of paths in CKG, while spatial knowledge is derived from RCC8. We encode common spatial relations as spatial knowledge and propose a model that integrates external semantic and spatial knowledge for VRD. The final prediction fuses results from the neural network and knowledge graph modules, further constrained by RCC8. Through experimental comparisons, we observe that employing the semantic and spatial knowledge fusion method improves the model’s recall. Our approach achieves comparable or even superior performance to the state-of-the-art under the zero-shot setting. Future research will focus on integrating visual features with external knowledge to further enhance predictions.

Acknowledgments. This work was supported by Jilin Province Natural Science Foundation under Grant No. 20240101378JC, the National Natural Science Foundation of China under Grant Nos. U22A2098 and 62172185, and the Science Research Foundation of Jilin Provincial Department of Education under Grant Nos. JJKH20241286KJ and JJKH20230336KJ.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*, pp. 852–869. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51
2. Donadello, I., Serafini, L.: Compensating supervision incompleteness with prior knowledge in semantic image interpretation. In: *IJCNN*, pp. 1–8. IEEE (2019)
3. Liang, K., Guo, Y., Chang, H., Chen, X.: Visual relationship detection with deep structural ranking. In: *AAAI*, vol. 32 (2018)
4. Zhan, Y., Yu, J., Yu, T., Tao, D.: On exploring undetermined relationships for visual relationship detection. In: *CVPR*, pp. 5128–5137 (2019)
5. Yu, D., Yang, B., Wei, Q., Li, A., Pan, S.: A probabilistic graphical model based on neural-symbolic reasoning for visual relationship detection. In: *CVPR*, pp. 10609–10618 (2022)
6. Wan, H., et al.: Iterative visual relationship detection via commonsense knowledge graph. *Big Data Res.* **23**, 100175 (2021)

7. Zareian, A., Karaman, S., Chang, S.-F.: Bridging knowledge graphs to generate scene graphs. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pp. 606–623. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58592-1_36
8. Chen, J., He, H., Wu, F., Wang, J.: Topology-aware correlations between relations for inductive link prediction in knowledge graphs. In: *AAAI*, vol. 35, pp. 6271–6278 (2021)
9. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: *CVPR*, pp. 3716–3725 (2020)
10. Liu, X., Gan, M.G., He, Y.: Multi-view visual relationship detection with estimated depth map. *Appl. Sci.* **12**(9), 4674 (2022)
11. Chiou, M.J., Zimmermann, R., Feng, J.: Visual relationship detection with visual-linguistic knowledge from multimodal representations. *IEEE Access* **9**, 50441–50451 (2021)
12. Gkanatsios, N., Pitsikalis, V., Koutras, P., Maragos, P.: Attention-translation-relation network for scalable scene graph generation. In: *ICCV Workshops*, pp. 0–0 (2019)
13. Jung, J., Park, J.: Improving visual relationship detection using linguistic and spatial cues. *ETRI J.* **42**(3), 399–410 (2020)
14. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. *KR* **92**, 165–176 (1992)
15. Wang, S., Wen, C., Lai, Y., Liu, W., Liu, D.: Interactive activity learning from trajectories with qualitative spatio-temporal relation. *Chin. J. Electron.* **24**(3), 508–512 (2015)
16. Li, S., Cohn, A.G.: Reasoning with topological and directional spatial information. *Comput. Intell.* **28**(4), 579–616 (2012)
17. Li, S., Ying, M.: Extensionality of the rcc8 composition table. *Fund. Inform.* **55**(3–4), 363–385 (2003)
18. Binong, J., Hazarika, S.M.: Extracting qualitative spatiotemporal relations for objects in a video. In: Mandal, J.K., Saha, G., Kandar, D., Maji, A.K. (eds.) *Proceedings of the International Conference on Computing and Communication Systems*, pp. 327–335. Springer Singapore, Singapore (2018). https://doi.org/10.1007/978-981-10-6890-4_31
19. Jung, J., Park, J.: Visual relationship detection with language prior and softmax. In: *IPAS*, pp. 143–148. IEEE (2018)
20. Russakovsky, O., Li, L.J., Fei-Fei, L.: Best of both worlds: human-machine collaboration for object annotation. In: *CVPR*, pp. 2121–2131 (2015)
21. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: *CVPR*, pp. 5410–5419 (2017)
22. Peng, J., Zhang, Y., Huang, W.: Visual relationship detection with image position and feature information embedding and fusion. *IEEE Access* **10**, 117170–117176 (2022)
23. Yu, D., Yang, B., Liu, D., Wang, H., Pan, S.: A survey on neural-symbolic learning systems. *Neural Networks* (2023)
24. Yu, R., Li, A., Morariu, V.I., Davis, L.S.: Visual relationship detection with internal and external linguistic knowledge distillation. In: *ICCV*, pp. 1974–1982 (2017)

25. Manigrasso, F., Miro, F.D., Morra, L., Lamberti, F.: Faster-LTN: a neuro-symbolic, end-to-end object detection architecture. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) *Artificial Neural Networks and Machine Learning – ICANN 2021: 30th International Conference on Artificial Neural Networks*, Bratislava, Slovakia, September 14–17, 2021, *Proceedings, Part II*, pp. 40–52. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86340-1_4
26. Buffelli, D., Tsamoura, E.: Scalable theory-driven regularization of scene graph generation models. In: *AAAI*, vol. 37, pp. 6850–6859 (2023)
27. Gkanatsios, N., Pitsikalis, V., Maragos, P.: From saturation to zero-shot visual relationship detection using local context. In: *BMVC* (2020)
28. Hu, Y., Chen, S., Chen, X., Zhang, Y., Gu, X.: Neural message passing for visual relationship detection. *arXiv preprint [arXiv:2208.04165](https://arxiv.org/abs/2208.04165)* (2022)
29. Socher, R., Chen, D., Manning, C.D., Ng, A.: Reasoning with neural tensor networks for knowledge base completion. In: *Advances in Neural Information Processing Systems*, vol. 26 (2013)
30. Li, B., Wang, Y.: Visual relationship detection using joint visual-semantic embedding. In: *ICPR*, pp. 3291–3296. IEEE (2018)
31. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *CVPR*, pp. 580–587 (2014)