Exploring Pairwise Spatial Relationships for Actions Recognition and Scene Graph Generation

Anfel Amirat[1(✉)], Nadia Baha[1], and Lamine Benrais[2]

[1] Computer Science Faculty, University of Science and Technology Houari Boumediene, Algiers, Algeria
{aamirat,anbahatouzene}@usthb.dz
[2] Faculty of Arts, KU Leuven, 3000 Leuven, Belgium
lamine.benrais@kuleuven.be

Abstract. Visual scene understanding is a fundamental problem and a complex task in computer vision, which not only requires identifying objects in isolation, but also the ability to understand and recognize the relationships between them. These relationships can be abstracted into a semantic representation of $< subject, predicate, object >$, resulting in a scene graph that captures much of the visual information and semantics in the scene. In recent years, scene graph generation with message-passing mechanism [1] has been an active area of research, as it has the potential to capture global dependencies between objects and their relationships. Inspired by these developments, this paper introduces a novel scene graph generation approach based on spatial relationships. Our approach performs a classification of the spatial relationship between each pair of objects to generate the initial scene graph. Then, based on the semantic features, the model detects action relationships in the scene and updates the scene graph by applying the message-passing mechanism. We conclude this paper by comparing the proposed method with the state-of-the-art approaches [1–7] and demonstrate the effectiveness of our method over the Visual Genome [1] dataset.

Keywords: Scene understanding · scene graph · visual relationships detection · spatial relationships · message-passing

1 Introduction

A scene graph is a structured representation of image content that encodes spatial and semantic information of each object and the relationship between each pair of them. Recently, inferring such a graph has gained more attention since it provides a deep understanding of the scene and improves various vision tasks such as Image Retrieval [8,9], Image Generation [10,11], Image/Video Captioning [12,13], and Visual Question Answering [14,15].

The major challenge of generating scene graphs is reasoning about relationships. Earlier works [16,17] aimed to produce a local prediction of object

Published by Springer Nature Switzerland AG 2023
I. Maglogiannis et al. (Eds.): AIAI 2023, IFIP AICT 675, pp. 381–392, 2023.
https://doi.org/10.1007/978-3-031-34111-3_32

relationships in order to simplify the process of generating visually-grounded scene graphs. The approach was to independently predict relationships between pairs of objects without considering the scene's context. In contrast, co-reasoning with contextual information could often resolve the ambiguity due to local predictions in isolation [18].

Message passing between individual objects or triplet is valuable for visual relationship detection [18]. Since objects with visual relationships are semantically related to each other, and relationships that share objects partially also have semantic relations, message passing between related elements is beneficial as it can improve the quality of visual relationship detection [2]. However, this mechanism is expensive and requires much computation time due to the numerous features to handle [19]. Moreover, visual appearance of the same relation varies significantly from one scene to another [20], making the features extraction phase more challenging. Thus, many methods focus on semantic features [21], trying to compensate for the lack of visual features.

To address these challenges and overcome the obstacle of variability in visual appearance, this work proposes a novel message-passing approach based on pairwise semantic spatial relationships. The concept is to replicate the human capacity to predict the relations between objects in a scene using their pairwise semantic spatial relationships.

In this paper, we first review past works related to message-passing scene graph generation and spatial relationships classification. Then, we introduce the proposed method in Sect. 3. In Sect. 4, the experimental results are shown and discussed. Finally, Sect. 5 concludes the paper by summarizing the obtained results.

2 Related Work

To contextualize our approach and evaluate its performance against the existing methods, we review the related work on message-passing scene graph generation and spatial relationships applications.

2.1 Message Passing

There are three levels to understanding and perceiving the context [18]: **first**, the interdependence between the different phrase components in a triplet is fundamental, the prediction of one component, such as the subject, predicate, or object, depends on the others. **Second**, triplets are not isolated, objects with relations are semantically dependent, and the relations that partly share object(s) are also semantically linked. **Third**, Visual relationships are specific to the scene, and global view features help predict relationships. Hence, message passing between objects and triplets is significant in detecting visual relationships.

The literature divides message-passing technique into two types:

Local Message Passing Within Triplet. Li et al. [22] proposed a phrase-guided visual relationship detection framework that first extracts three feature branches for each triplet proposal (subject, predicate, and object). Then, it uses a phrase-guided message-passing structure to exchange information between the three branches. Dai et al. [23] proposed an efficient framework known as the Deep Relational Network (DR-Net). By using multiple units of inference that capture the statistical relationships between triplet components, the DR-Net produces the posterior probabilities of the subject, object, and relationship. Zoom-Net [2] is another interesting model. It uses a Spatiality-Context-Appearance Module consisting of 2 spatiality-aware feature alignment cells to pass messages between the different triplet components. This type of message passing ignores the global context, whereas joint reasoning using contextual information can often resolve ambiguities caused by isolated local predictions [18].

Global Message Passing Across All Elements. Li et al. [3] developed a Multi-level Scene Description Network (MSDN) in which the passage of the message is guided by a dynamic graph constructed from objects and caption region proposals. F-Net, proposed by Li et al. [24], clusters the fully-connected graph into several subgraphs. Next, it uses a Spatial-weight Message Passing structure for passing messages between subgraph and object features. MSDN and F-Net considered a subgraph as a whole when sending and receiving messages. Liao et al. [25] proposed semantics-guided graph relation neural network (SGRNN). In their approach, the target and the source must be an object or a predicate within a subgraph. When considering all other objects as carriers of global contextual information for each object, they will pass messages to each other throughout a fully-connected graph. However, propagating many types of features and inferencing on a densely connected graph is very expensive and time-consuming to train [19].

2.2 Spatial Pairwise Relationships

Apprehending the spatial relationships between objects and how they are positioned and related to one another is imperative for a deep understanding of the scene. The application of spatial relation detection is useful in visually situated dialog and Human-robot interaction. For example, when instructing a robot in a household environment to accomplish a specific task [26] or when self-driving cars are designed to provide a textual explanation for their actions [27]. Likewise, the explicit use of spatial prepositions is also helpful in automatic image captioning [28].

For the proposed approach, we decide to stimulate the human capacity to infer much information by knowing the spatial relations between the different objects in the scene to detect and infer activities and action relations between image entities.

In this work, we propose a novel approach for scene graph generation based on the global message-passing mechanism. By incorporating semantic spatial relationships, our approach aims to overcome the challenge of variability in visual appearance and make more robust predictions about the relationships between objects in a scene.

3 Proposed Method

The proposed approach for scene graph generation is divided into pairwise spatial relationships classifications and scene graph update, as Fig. 1 shows. Our model tackles the visually-grounded scene graph generation from an image by generating a graph with a spatial relationship between each object pair. Then, recognize the action relationship and update the scene graph by applying the message-passing mechanism using only semantic features (objects and spatial relations labels). To achieve this, we use two neural network architectures that focus on each task independently and stack both architectures together once they have been trained. We use ground truth objects for object detection and recognition to evaluate the approach appropriately.

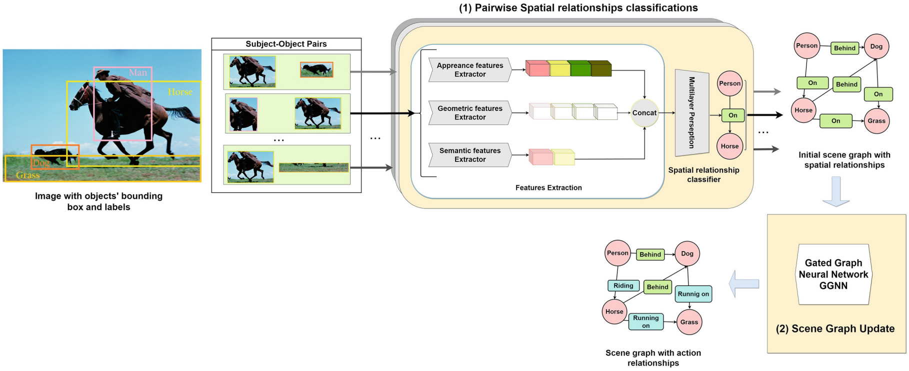

(1) Pairwise Spatial relationships classifications
Man
Horse
Dog
Grass
Image with objects' bounding box and labels
Subject-Object Pairs
...
Appreance features Extractor
Geometric features Extractor
Semantic features Extractor
Concat
Features Extraction
Multilayer Perseption
Person
On
Horse
Spatial relationship classifier
Person
Behind
Dog
On
Behind
On
Horse
On
Grass
Initial scene graph with spatial relationships
Gated Graph Neural Network GGNN
(2) Scene Graph Update
Person
Behind
Dog
Riding
Behind
Runnig on
Horse
Running on
Grass
Scene graph with action relationships

Fig. 1. An overview pipeline of our image scene graph generation model.

Before delving into the proposed model, we describe the scene graph structure. Formally, a scene graph is a structured representation of a scene's content. It comprises the objects' labels with bounding box coordinates and the relationship between each object pair.

A scene graph is defined as a 3-tuple set $G = \{B, O, R\}$:

$B = \{b_1, b_2, ..., b_n\}$ is the bounding box set, $b_i \in R^4$ corresponds to the bounding box of the i^{th} region.

$O = \{o_1, o_2, ..., o_n\}$ object's label set, o_i corresponds to the label class of the region b_i.

$R = \{r_{1\to 2}, r_{1\to 3}, ..., r_{n\to n-1}\}$ relationship triplet set, where $r_{i\to j}$ is a triplet of the object (o_j, b_j), the subject (o_i, b_i), and the relationship class $a_{i\to j}$.

3.1 Pairwise Spatial Relationships Classifications

Features Extraction. This module aims to get the objects' appearance, semantic cues, and relative spatial locations between pairwise objects. This approach is inspired by [29] to extract three types of features to classify the semantic spatial relationship between each object's pair in the scene.

Geometric Features: we exploit the spatial contextual information from the subject, object, union, and intersection boxes. For each box (x_1, y_1, x_2, y_2), a 9-dimensional vector is calculated as (1) shows:

$$V = (\frac{c_x}{W}, \frac{c_y}{H}, \frac{w}{W}, \frac{h}{H}, \frac{x_1}{W}, \frac{h_1}{H}, \frac{x_2}{W}, \frac{h_2}{H}, \frac{w*h}{W*H}) \tag{1}$$

where $(c_x, c_y) = (\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$ is the box's centroid, $(w, h) = (x_2 - x_1, y_2 - y_1)$ denotes the width and the height of the box, and (W, H) the width and the height of the image. For an empty intersection box, a zero vector represents the intersection box's geometric features. Then, all four vectors are concatenated to compose the geometric features.

Appearance Features: for the subject bounding box region, object bounding box region, union box, and intersection box, we use the FC7 layer from VGG16 [30] pre-trained on ImageNet [31] to extract the appearance feature vector (4096-d). For an empty intersection box, a zero vector represents the intersection box's appreence features. Then we concatenate all four vectors to compose the appearance features of the spatial relationship.

Semantic Features: glove [32] is used as a word embedding engine to encode objects' label names for the subject and the object. For phrase names, the mean vector is calculated. By concatenating the two encoded name features, the semantic relation features are composed.

Finally, the relation features are obtained by concatenating geometric, appearance, and semantic features.

Spatial Relationship Classification. After concatenating the extracted features described in 3.1 for each object's pair, we feed them to a multilayer perceptron neural network architecture (MLP) to classify the spatial relationships. Then, the initial scene graph with only pairwise spatial relationships is generated.

3.2 Scene Graph Update

We aim to update the scene graph relationships generated in 3.1 by applying the message-passing mechanism to have more meaningful semantic information with activities and action relationships.

Action Relationship Recognition. This step aimes to update edge representation while keeping node representations constant by using a variant of GGNN [33] to propagate information among edges. For each edge $a_{s\to o}$, three steps, as Fig. 2 shows, are needed: pass preparation, information aggregation, and edge update.

Pass Preparation: for each node from the subject node (o_s, b_s), and the object (o_o, b_o), its set of neighbors (o_i, b_j) is selected .

Information Aggregation: for each node from subject node (o_s, b_s) and object node (o_o, b_o), information is summarized by computing incoming information from its neighbors as shown in (2) :

$$m_k = o_k + \sum a_{i\to k} \cdot o_i - \sum a_{k\to j} \cdot o_j \tag{2}$$

Edge Update: after information aggregation, we concatenate m_s and m_o. Then it is passed with the current state $a_{S\to O}$ to Gated Recurrent Unit (GRU) to update the edge label. Finally, a scene graph with, in addition, pairwise action relationships is obtained.

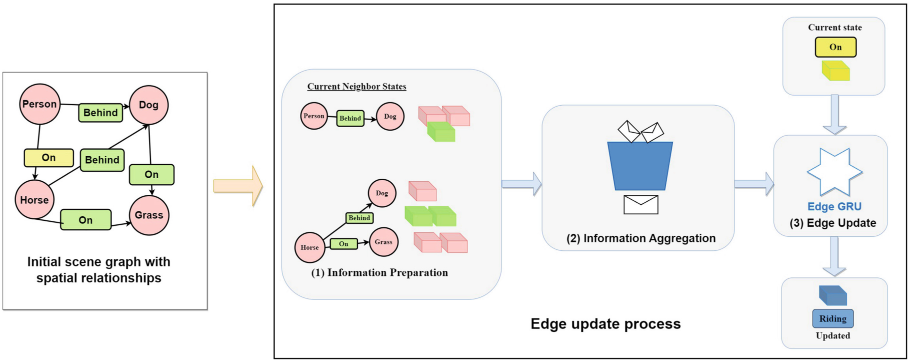


Fig. 2. Relation update process. After computing the information *(Informations Aggregation)* from the selected neighbors *(Information Preparation)*, the state of the edge **(on)** is updated to **(riding)** by passing both the information computed and the current state **(on)** to the GRU *(Edge Update)*.

4 Test and Results

This section presents a details evaluation of the proposed model. First, an evaluation of the spatial relationship classifier is processed. Then, we pass to the model of scene graph generation. Tests are conducted on a personal computer with an i7 processor, 16 GB memory, and a 2 GB Nvidia GPU.

4.1 Pairwise Spatial Relationships Classifications

Dataset. We conduct the experiments and evaluate the Spatial relationships classifier on the SpatialSense dataset [34], a collected benchmark for spatial relation recognition that contains 17498 spatial relations on 11596 images. All images are collected from Flickr and NYU [37]. The annotated spatial relation in the dataset covers 3679 unique object classes and 9 unique predicates (i.e., above, behind, in, in front of, next to, on, to the left of, to the right of, under). The SpatialSense dataset provides positive and negative examples of spatial relationships. To train the spatial relationship classifier, only positive triplets are considered. Following the official split in [34], we take 65% of relations for training, 15% for validation, and 20% for testing.

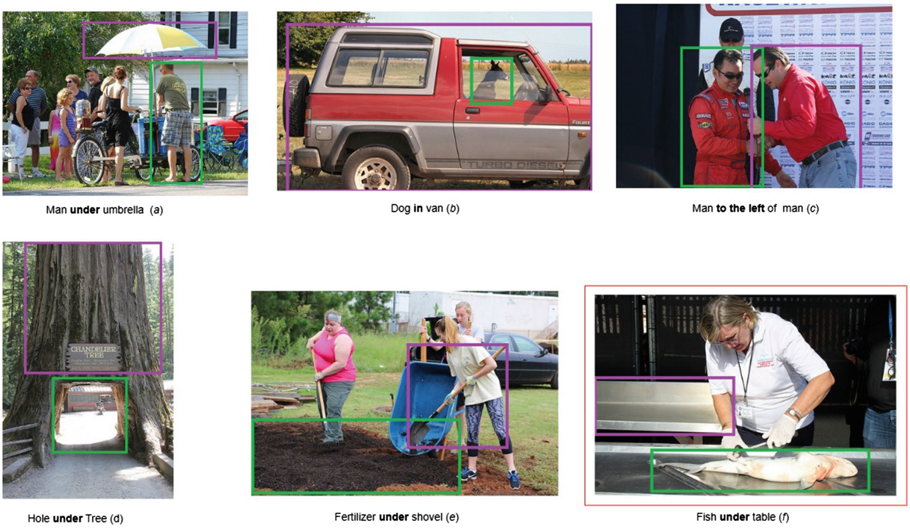
Evaluation Metric. The proposed classifier's ability to classify pairwise spatial relationships can be evaluated using classification accuracy [35] as a reliable and fair measure.

Compared with State-of-the-art Methods. We compare our classifier with various recent methods.

Table 1. Classification accuracy comparison on the test split of the SpatialSense dataset (All Values Expressed as Percentages). IFO = in front of, TTFO = to the left of, TTRO = to the right of. **Bold** font represents the highest accuracy; underline means the second highest.

Model	overall	above	behind	in	IFO	next to	on	TTFO	TTRO	under
Vip-CNN [22]	67.2	55.6	68.1	66.0	62.7	62.3	72.5	69.7	73.3	66.6
Peyre et al. [36]	67.5	59.0	67.1	69.8	57.8	65.7	75.6	56.7	69.2	66.2
PPR-FCN [38]	66.3	61.5	65.2	70.4	64.2	53.4	72.0	69.1	71.9	59.3
DRNet [23]	71.3	**62.8**	**72.2**	69.8	66.9	59.9	79.4	63.5	66.4	75.9
VTranE [30]	69.4	61.5	69.7	67.8	64.9	57.7	76.2	64.6	68.5	76.9
Language-only [34]	60.1	60.4	62.0	54.4	55.1	56.8	63.2	51.7	54.1	70.3
2D-only [34]	68.8	58.0	66.9	70.7	63.1	62.0	76.0	66.3	74.7	67.9
Language+2D [34]	71.1	61.1	67.5	69.2	66.2	64.8	77.9	**69.7**	74.7	77.2
DSRR [40]	**72.7**	61.5	71.3	**71.3**	**67.8**	**65.1**	79.8	69.4	**75.3**	**78.6**
The proposed approach	71.6	62.1	67.0	70.2	66.6	64.5	**79.9**	65.9	73.2	72.6

Table 1 shows the performance of different approaches on the SpatialSense dataset. Vip-CNN [22], Peyre et al. [36], PPR-FCN [38], DRNet [23], and VtransE [39], initially designed for visual relationship detection, are based only on visual appearance. Language-only,2D-only, and Language+2D [34], designed for spatial relation recognition, are based on 2D/Language features. Our classifier takes into consideration the three main types of features: appearance features, semantic features, and geometric features. Overall, the results of the accuracy score indicate that our proposed classifier outperforms almost all existing



Man **under** umbrella (*a*)

Dog **in** van (*b*)

Man **to the left** of man (*c*)

Hole **under** Tree (d)

Fertilizer **under** shovel (*e*)

Fish **under** table (*f*)

Fig. 3. Classification examples of spatial relationships by the proposed classifier on the SpatialSense dataset: *a, b, c, d,* and *e* are correct classifications, and in contract *f* is a misclassification. We believe that with depth information, our classifier could predict the proper label *in front of* instead of *under* for the misclassification *f.*

approaches in terms of overall accuracy, except DSRR (by only 1.1%) [40], which exploits depth information with an additional depth estimation model. With the additional depth, we expect our classifier to gain another performance boost and correctly classify complex cases that were previously misclassified, as Fig. 3 shows.

4.2 Scene Graph Generation

After training and testing our classifier for spatial relationships between pairs of objects, this sub-section evaluates the whole scene graph generation process.

Dataset. To evaluate the proposed approach, we use VG150 [1]. It is a widely adopted subset of Visual Genome for evaluating scene graph generation tasks. It contains 108073 images and covers 150 object categories and 50 predicate categories. We follow the same split in [1] for evaluating our approach.

Evaluation Metric. We aim to generate the scene graph for images. The key points are relationship classification and graph generation, while we no longer evaluate the accuracy of object detection or recognition. We evaluate the model performance from the aspect of predicate classification (PredCls) as we use both ground truth boxes and object labels directly. We use R@50 and R@100 to evaluate the performance. R@K computes the fraction of times a true relationship is predicted in an image's top k confident relation predictions.

Compared with State-of-the-art Methods. We report predicate classification on Visual Genome [1] in Table 2. This experiment is meant to serve as a benchmark against existing message-passing scene graph approaches.

Table 2. Evaluation results of the predicate classification task on the visual Genome dataset [1].

Model	R@50	R@100
MP [1]	41.8	55.5
Zoom-Net [2]	67.25	77.51
MSDN [3]	67.03	71.01
AGGNN [4]	65.1	67.2
ReRN* [5]	62.1	63.7
Dornadula et al. [6]	56.65	57.21
SGRN [7]	64.2	66.4
Proposed Approach	73.09	78.1

The experiments prove the effectiveness of our proposed method. We outperform existing models that use Visual Genome supervision for PredCls by 6,06 recall@50 and 0.51 recall@100. Message Passing [1], and Zoom-Net [2] are local message-passing-based methods. In contrast, the rest are all global message-passing-based methods.

Visual features for the same relation vary greatly from scene to scene, making relation predicting more challenging, especially for rare and unseen configurations and relations. For example, the visual features that represent the "riding" relation between a person and a horse can be very different from one image to another, depending on the pose, the background, the lighting condition, etc. In contrast, considering the semantic pairwise spatial relationships between the objects in the scene, we can infer from "the man on the horse and horse on the grass" that the action relation between man and horse is "riding". That is why focusing on semantic features like semantic pairwise spatial relationships can improve predicate classification tasks.

5 Conclusion

This paper investigates a novel message-passing scene graph generation approach based on semantic spatial relationships. First, we classify the spatial relationship between each pair of objects in the scene by extracting geometric, appearance, and semantic features and then passing them to an MLP architecture. After, we apply the message-passing mechanism as a second step to detect action relationships and update the scene graph.

Experimental results demonstrate its efficiency and competitiveness compared to the state-of-the-art approaches with 73.09 for R@50 and 78.1 for R@100. However, there are several prospective paths for improving this approach further. Firstly, incorporating additional depth information into the spatial relationship classifier can improve the accuracy and robustness of the model. Moreover, training the spatial relationships classifier on datasets with other spatial relationship classes, such as between, near, and far can be useful in scenes with more diverse spatial configurations. Furthermore, extending the proposed method to work with multi-spatial relations instead of single-spatial relations can boost our model, as it can capture more nuanced relationships between objects.

By incorporating these improvements, the proposed method can be enhanced and upgraded to achieve even better performance.

These prospective paths can be explored in future research and can contribute to advancing the field of scene understanding.

References

1. Xu, D., Zhu, Y., Choy, C. B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5410–5419 (2017)
2. Yin, G., et al.: Zoom-net: Mining deep feature interactions for visual relationship recognition. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 322–338 (2018)
3. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1261–1270 (2017)
4. Li, S., Tang, M., Zhang, J., Jiang, L.: Attentive gated graph neural network for image scene graph generation. Symmetry **12**(4), 511 (2020)
5. Tian, P., Mo, H., Jiang, L.: Exploring correlation of relationship reasoning for scene graph generation. Int. J. Mach. Learn. Cybern. **13**(9), 2479–2493 (2022)
6. Dornadula, A., Narcomey, A., Krishna, R., Bernstein, M., Li, F.F.: Visual relationships as functions: Enabling few-shot scene graph prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)
7. Liao, W., Lan, C., Zeng, W., Yang, M.Y., Rosenhahn, B.: Exploring the semantics for visual relationship detection. arXiv preprint arXiv:1904.02104 (2019)
8. Johnson, J., et al.: Image retrieval using scene graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3668–3678 (2015)
9. Ramnath, S., Saha, A., Chakrabarti, S., Khapra, M.M.: Scene Graph based Image Retrieval-A case study on the CLEVR Dataset. arXiv preprint arXiv:1911.00850 (2019)
10. Fang, F., Yi, M., Feng, H., Hu, S., Xiao, C.: Narrative collage of image collections by scene graph recombination. IEEE Trans. Visual Comput. Graph. **24**(9), 2559–2572 (2017)
11. Herzig, R., Bar, A., Xu, H., Chechik, G., Darrell, T., Globerson, A.: Learning canonical representations for scene graph to image generation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12371, pp. 210–227. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58574-7_13

12. Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G.: Unpaired image captioning via scene graph alignments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10323–10332 (2019)
13. Xu, N., Liu, A.A., Liu, J., Nie, W., Su, Y.: Scene graph captioner: image captioning based on structural visual representation. J. Vis. Commun. Image Represent. **58**, 477–485 (2019)
14. Yang, Z., Qin, Z., Yu, J., Hu, Y.: Scene graph reasoning with prior visual relationship for visual question answering. arXiv preprint arXiv:1812.09681 (2018)
15. Qian, T., Chen, J., Chen, S., Wu, B., Jiang, Y.G.: Scene graph refinement network for visual question answering. In: IEEE Transactions on Multimedia (2022)
16. 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.) ECCV 2016. LNCS, vol. 9905, pp. 852–869. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51
17. Yu, R., Li, A., Morariu, V.I., Davis, L.S.: Visual relationship detection with internal and external linguistic knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1974–1982 (2017)
18. Zhu, G., et al.: Scene graph generation: A comprehensive survey. arXiv preprint arXiv:2201.00443 (2022)
19. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5831–5840 (2018)
20. Cong, W., Wang, W., Lee, W.C.: Scene graph generation via conditional random fields. arXiv preprint arXiv:1811.08075 (2018)
21. 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.) ECCV 2016. LNCS, vol. 9905, pp. 852–869. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_51
22. Li, Y., Ouyang, W., Wang, X., Tang, X.O.: Vip-cnn: Visual phrase guided convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1347–1356 (2017)
23. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3076–3086 (2017)
24. Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., Wang, X.: Factorizable net: an efficient subgraph-based framework for scene graph generation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 335–351 (2018)
25. Liao, W., Lan, C., Zeng, W., Yang, M.Y., Rosenhahn, B.: Exploring the semantics for visual relationship detection. arXiv preprint arXiv:1904.02104 (2019)
26. Fasola, J., Mataric, M.: Using spatial language to guide and instruct robots in household environments. In: 2012 AAAI Fall Symposium Series (2012)
27. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 563–578 (2018)
28. Ghanimifard, M., Dobnik, S.: What goes into a word: generating image descriptions with top-down spatial knowledge. In: Proceedings of the 12th International Conference on Natural Language Generation, pp. 540–551 (2019)
29. Zhang, Y., Pan, Y., Yao, T., Huang, R., Mei, T., Chen, C.W.: Boosting scene graph generation with visual relation saliency. ACM Trans. Multimed. Comput. Commun. Appl. **19**(1), 1–17 (2023)

30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
31. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. IEEE (2009)
32. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
33. Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.: Gated graph sequence neural networks. arXiv preprint arXiv:1511.05493 (2015)
34. Yang, K., Russakovsky, O., Deng, J.: Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2051–2060 (2019)
35. Japkowicz, N., Shah, M.: Evaluating learning algorithms: a classification perspective. Cambridge University Press (2011)
36. Peyre, J., Sivic, J., Laptev, I., Schmid, C.: Weakly-supervised learning of visual relations. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5179–5188 (2017)
37. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (5) 7576, 746–760 (2012)
38. Zhuang, B., Liu, L., Shen, C., Reid, I.: Towards context-aware interaction recognition for visual relationship detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 589–598 (2017)
39. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5532–5540 (2017)
40. Ding, X., Li, Y., Pan, Y., Zeng, D., Yao, T.: Exploring depth information for spatial relation recognition. In: 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 279–284. IEEE (2020)