RELATIONSHIP-AWARE PRIMAL-DUAL GRAPH ATTENTION NETWORK FOR SCENE GRAPH GENERATION

Anonymous ICME submission

ABSTRACT

The relationships and interactions between objects contain rich semantic information, which plays a crucial role in scene understanding. Existing methods do not attach great importance to the expression of relational features. To tackle this problem, we propose a novel Relationship-aware Primal-Dual Graph Attention Network (RPDGAT) to extract the comprehensive semantic features of objects and explore the sparse graph inference for scene graph generation. RPDGAT mines the inherent attributes and the relationships between objects by fusing multiple features, e.g. appearance, spatial, and category features. After feature extraction, we design a trainable relationship distance measure network to construct the robust and sparse graph structure for efficient graphical message passing. Moreover, it can preserve the contextual cues and neighboring dependency for objects and relationships from the interaction between primal and dual graphs. Extensive experimental results present the improved performance of our method over several state-of-the-art methods on the visual genome datasets.

Index Terms**—** Scene graph generation, Sparse graph, Primal-dual graph attention network, Graph inference

1. INTRODUCTION

A deep understanding of the visual scene plays an important role in computer vision and machine learning. Scene graph generation (SGG) [1, 2, 3] is a frequently common and effective way to derive the rich semantic information of image data and help understand the visual scene by constructing the structured expression. Different from the independent object tasks, SGG can capture the objects and semantic relationships between object pairs, which can be widely used in many high-level vision applications, such as image retrieval [1], VQA [3], and image caption [2], and so on.

In literature, there have been many methods [4, 5] proposed to generate the scene graph of an image. Most of them centered on exploring more effective ways of feature extraction and graph inference in visual scene. For the feature extraction, Ren et al. [6, 7, 8, 9, 10] proposed the ROI features extracted from Fast RCNN for subsequent inference. Chen et al. [11, 12] pointed out that the spatial features or category features of objects also bring great value for inferring scene graph. Therefore, they imposed the spatial configurations and category distributions as complementary to the appearance features. For the graph inference, constructing dense graph [8, 13, 10] or pruning graph [9, 11, 14] are the two mainstream methods. Chen et al. [12, 7] built the fully connected graph for all objects and adopted the CRF and GRU modules to achieve feature interactions based on the graph inference. Besides, Yang et al. [9, 11, 15] proposed to construct the sparse graph by pruning some edges from dense graphs and implemented more efficient and accurate message passing based on GAT and Tree structured LSTM.

Although several SGG methods are presented with good performance, they have the following two main drawbacks. On the one hand, current methods often extract the appearance features and ignore the semantic information of spatial features and category features, which are the important auxiliary for the relationship detection between multiple objects. On the other hand, traditional methods are adopted to build dense graph structure, which is computationally expensive and not used to tackle large-scale problems. Meanwhile, dense graph inference also leads to feature saturation and redundancy and graphical message passing time-consuming.

To solve these problems, we propose a novel and efficient scene graph generation method named Relationship-aware Primal-Dual Graph Attention Network (RPDGAT) in this paper. Specifically, based on the object proposals from Faster-RCNN [6], RPDGAT extracts the comprehensive object features, e.g. the spatial and category features, to preserve the high-level semantic information, by which it maintains a trainable relationship distance matrix. Afterwards, a sparse graph with more reliable and semantic connected edges is constructed for efficient message passing. Meanwhile, we design the primal-dual graph attention network on the sparse graph to acquire the feature interactions and extract the complex relationships between multiple objects for graph inference. Finally, extensive experimental results present our proposed method outperforms the state-of-the-art models for the PredCls, SGCls, and SGGen tasks on visual genome datasets.

Our contributions are summarized as follows:

- To preserve comprehensive semantic information of scene graph, we formally deep fuse the appearance, spatial and category features of objects to mine the inherent attributes and the relationships between objects.

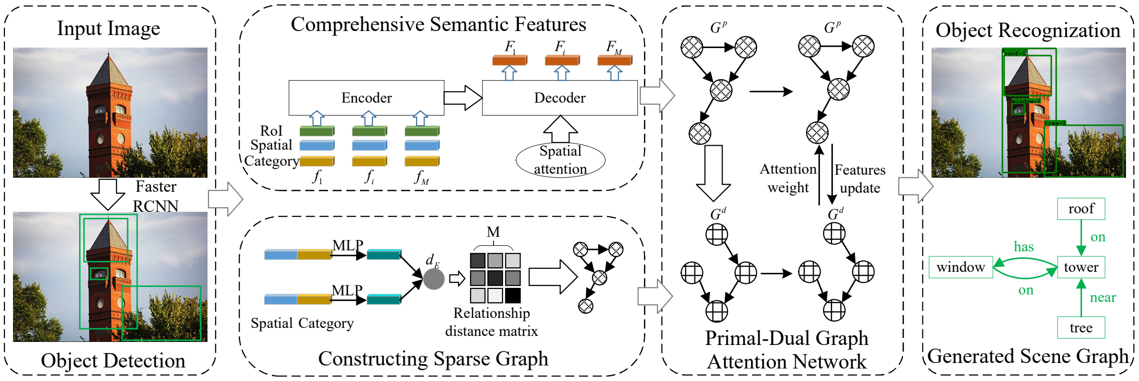

Fig. 1. A brief pipeline of our Relationship-aware Primal-Dual Graph Attention Network.

- To construct the robust graph structure, we design a trainable relationship distance measure network, which takes multiple semantic features as the inputs. And based on the distance matrix, a primal sparse graph can be screened for efficient message passing and lowering features redundancy.

- To tackle the sparse graph inference, we propose a novel primal-dual graph attention network and achieve the features and weight interactions between dual graph and primal graph, which mines the contextual cues and captures the dependencies of objects and relationships.

2. RELATED WORKS

Feature extraction in scene graph. In literature, many researchers took the RoI features as the object features to design a knowledge model or graph model for learning semantic features and context [9, 10]. Some researchers fuse the appearance features with other object features. Lu et al. [16] addressed the semantic relationship information by combining visual images and linguistic priors to improve relationships detection. Li et al. [13] incorporated the features of objects, phrases and regions for multi-task joint training. Dai et al. [12] extracted the individual region features and fused the spatial information of object and subject regions for predicate features. Tang et al. [11] concatenated the RoIAlign and spatial features to construct the dynamic tree model. Compared with these superficial fusion ways, our model can mine comprehensive semantic information with the integration of the multiple enhanced features.

Graph inference in scene graph. Dense graph has high computational complexity and could lead to saturation in the context encoding for graph inference and message passing. Xu et al. [7] adopted GRU to build the dense graph and message passing to improve the accuracy of detection. Lin et al. [15] proposed a direction-aware message passing module to extract the direction of dense edges. Sparse graph is an efficient way to enhance the efficiency of features delivery and alleviates the computation cost. Li et al. [17] proposed to prune the dense graph by integrating several nodes into a subgraph. Besides, Yang et al. [9] proposed a trainable relation proposal network (RePN) to prune the relationship edges of graph and preserve the context information by the GAT module [18]. Motivated by RePN [9], we propose a novel and efficient primal-dual graph attention network by constructing the sparse graph for graph inference.

3. METHODS

3.1. Model overview and problem formulation

To enhance the efficiency of graph inference, we propose a novel Relationship-aware Primal-Dual Graph Attention Network (RDPGAT) to construct a sparse graph for scene graph generation. The main structure of RDPGAT is shown in Figure 1 and its main procedure is summarized into the following four steps. First, we adopt the Faster-RCNN to locate the key object proposals in images. Then object features are strengthened by fusing and refining multiple features. Besides, we design a trainable relationship distance matrix for constructing the sparse graph. Finally, a primal-dual graph attention network is proposed to recognize objects and relationships.

Given a scene graph G, it consists of objects bounding boxes B, object categories O and relationships R. In our work, let I, E and G denote an image, the relationships between objects, and a scene graph, respectively. We decompose the scene graph generation into three parts:

$$P(G|I) = P(B|I)P(E|I,B)P(R,O|I,B,E), \quad (1)$$

where $P(B|I)$ generates a series of candidate regions of objects in images. Similar with [6, 7, 8, 9], we use the off-the-shelf Faster RCNN framework to obtain these candidate regions. $P(E|I,B)$ indicates the object features fusion, calculation of relationship distance matrix, and construction of the sparse graph. $P(R,O|I,B,E)$ classifies objects and predicts relationships based on primal-dual graph attention network to generate the entire scene graph.

3.2. Comprehensive semantic features

Given an image, we locate a series of candidate objects by Faster RCNN and then obtain the object's position by the objects bounding boxes $B = \{b_1, b_2, ..., b_n\}$. Besides, for each bounding box b_i, the appearance features f_i^a and the probability distribution p_i of the object classes are also extracted.

By the analysis, the comprehensive semantic features, such as the positions, the class probability, and the appearance features of objects are complementary to the appearance features and robust to light intensity. Therefore, we convert the position coordinates and category probabilities into the corresponding spatial and category features, and further fuse these three features in scene graph.

To extract the semantic features for individual object or multiple objects, we first expand the position coordinates of $M \times 4$ dimensional vectors to $M \times 16$ dimensional vectors by concatenation, which is a good balance between fidelity and cost, where M is the number of objects. Then they are transformed into $M \times 128$ dimensional vectors through a multilayer perceptron (MLP) to get the spatial feature vector $f_i^s = MLP(b_i||b_i)$, where $||$ is the concatenation. Similarly, the category features $f_i^c = MLP(p_i)$.

The appearance features f_i^a, spatial features f_i^s, and category features f_i^c are combined to form a linear sequence: $[(f_1^a, f_1^s, f_1^c), ..., (f_M^a, f_M^s, f_M^c)]$. Similar with [8], we sort these object regions and encode this linear sequence through a bidirectional LSTM to obtain the object fusion features:

$$C = biLSTM\left([f_i^a, f_i^s, f_i^c]\right), \tag{2}$$

where $C = [c_1, c_2, ..., c_M]$ indicates the hidden states in the last LSTM layer. The fusion features C will be sequentially decoded to obtain the refined features through LSTM based on spatial attention mechanisms. The spatial attention mechanism enables our model to extract contextual features better and improve the interpret-ability of the model. The feature vector F_i, which integrates all information about the $i-$th object in the sequence of the features is:

$$\begin{cases} e_i^t = w^T tanh(W_a h_{t-1} + W_b(c_i, f_i^s) + b_a), \\ F_i = \sum\limits_{j=1}^{M} a_j^t c_j = \sum\limits_{j=1}^{M} \dfrac{\exp(e_j^t)}{\sum_{k=1}^{M} \exp(e_k^t)} c_j, \end{cases} \tag{3}$$

where (c_i, f_i^s) is the encoded context vector of $i-$th object and its spatial features; And w^T, W_a, W_b, b_a are the parameters, which are estimated during model training; $\sum_{i=1}^{n} a_i^t = 1$ and a_i^t are attention weights. The comprehensive semantic features F_i implies the inherent attributes of objects and contextual features to identify objects and relationships.

3.3. Relationship distance matrix and the sparse graph

As mentioned earlier, a reasonable sparse graph gives more semantic connections than dense graph. Our model calculates the potential relationship distances for object pairs to form the relationship distance matrix and construct a sparse graph structure for an image.

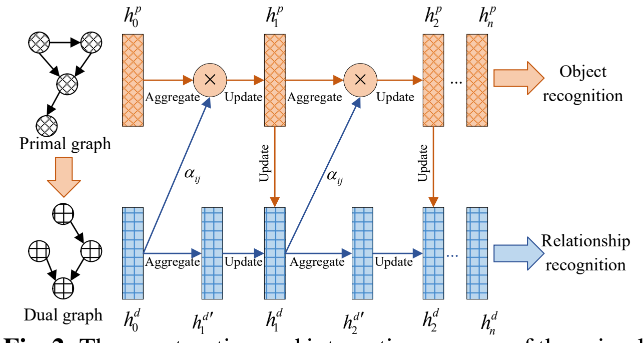

Fig. 2. The construction and interaction process of the primal-dual graph attention network.

Empirically, a reliable sparse graph mainly considers two aspects: the spatial distance between objects and object category. For the spatial distance, closer objects are more likely to be connected. For the object category, some object categories have meaningless relationships, such as hats and glasses, even though the spatial distance between them is usually small. Conversely, the relationships between some categories are more likely to exist with longer spatial distance, such as 'people-play-football'. In summary, the categories of objects and their corresponding locations are able to determine whether a meaningful relationship exists between two objects.

We calculate the relationship distance matrix between objects through category features and spatial features. First, the spatial features and category features are normalized and concatenated for $M \times 256$ features. To conveniently calculate the distance between features and make the parameters learnable, we use the MLP to compress the features into $M \times 32$ features. Finally, we calculate the Euclidean distance for any two object features to obtain $M \times M$ distance matrix, the item m_{ij} is calculated as:

$$m_{ij} = d_E\left(MLP(f_i^s||f_i^c), MLP(f_j^s||f_j^c)\right), \tag{4}$$

where d_E donates the euclidean distance function.

After obtaining the distance matrix for all object pairs, we sort the scores and select the top K object pairs to construct sparse graph structure, which is with n nodes and K edges, to improve the robustness of graph and the efficiency of message passing, and to lower the feature redundancy.

3.4. Primal-Dual graph attention network

In this subsection, we design a primal-dual graph attention network to solve the sparse graph inference problem. First, given the primal graph (i.e., the spares graph in Section 3.3), its dual graph is constructed automatically, in which nodes represent the relationships in the primal graph. Then, the features and attention weights based on GAT take multiple rounds of interactions between the dual graph and the primal graph. Finally, the model will recognize the objects and their relationships according to the node features of the primal graph and dual graph, respectively. The construction and interaction process of the primal-dual graph attention network is presented in Figure 2.

Constructing the dual graph. Without loss of generality, we regard the sparse graph generated by relationship distance matrix as the primal graph $G^p = (n^p, e^p)$, where n^p and e^p represent the nodes of the object proposals and the edges of the relationships between object pairs, respectively. Given the primal graph G^p, its corresponding dual graph $G^d = (n^d, e^d)$ is constructed as follows: (1) Each edge e^p in the primal graph G^p will be a node n^d in the dual graph G^d. (2) For two edges r_i and r_j, $(r_i, r_j \subseteq e^p)$, if the object node of edge r_i is the subject node of edge r_j, then we create an edge e^d_{ij} in G^d connecting n^d_i and n^d_j. Therefore, the dual graph G^d has K nodes and is a more sparse structure than its primal graph.

Nodes features assignment of the primal and dual graphs. After extracting the object fusion features F_i and constructing the graph structures G^p and G^d, we assign the node-wised features to the corresponding nodes in the primal-dual graph. For the primal graph G^p , the node features h^p_i are initialized by using the fused object features F_i, i.e. $h^p_i = F_i$. For the dual graph G^d, to generate the node features h^d_u, the features of the subject node F_i and the features of the object node F_j in G^p will be concatenated sequentially. And then we adopt to use the fully connected layer to compress the dimensionality of F_i and F_j, i.e. $h^d_u = \eta(F_i || F_j)$, where η is the Leaky ReLU and node u in G^d is related with e_{ij} in G^p.

Interactions between the primal and dual graphs. To preserve the more discriminative and robust representations of the nodes in the primal and dual graphs, we propose to apply the graph attention network to update the features and attention weights iteratively by using the interactions between two graphs. The features of i-th node $\overline{h}^p_i$ in primal graph are aggregated and updated as:

$$\overline{h}^p_i = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W^p_{ij} \overline{h}^p_j\right), \tag{5}$$

where $N(i)$ represents the neighbors of the i-th node, σ is the activation function, W denotes the learned parameters and α_{ij} is the attention weight and computed as:

$$\alpha_{ij} = \frac{exp\left(\eta(W^d_u a^d \overline{h}^d_u)\right)}{\sum\limits_{v \in E(i)} exp\left(\eta(W^d_v a^d \overline{h}^d_v)\right)}, \tag{6}$$

where a^d is a fully connected layer; $\overline{h}^d_u$ denotes the $u-$th node features in dual graph; and $E(i)$ represents the nodes in dual graph whose corresponding edges include the $i-$th node in primal graph. Then, the node features $\overline{h}^d_u$ in dual graph are aggregated and updated as:

$$\begin{cases} \overline{h}^{d'}_u = \sigma\left(\sum\limits_{v \in N(u)} \frac{1}{c_{uv}} W^d_{uv} \overline{h}^p_v\right), \\ \overline{h}^d_u = \eta\left(\overline{h}^{d'}_u || \eta(\overline{h}^p_i || \overline{h}^p_j)\right), \end{cases} \tag{7}$$

Table 1. Comparison our model with other state-of-the-art methods on R@50 and R@100.

Method	SGGen		SGCls		PredCls		Mean
	R50	R100	R50	R100	R50	R100	
VRD	0.3	0.5	11.8	14.1	27.9	35.0	14.93
IMP	3.4	4.2	21.7	24.4	44.8	53.0	25.25
GRNN	11.4	13.7	29.6	31.6	54.2	59.1	33.27
IMP+	20.7	24.5	34.6	35.4	59.3	61.3	39.30
Freq	26.2	30.1	32.3	32.9	60.6	62.2	40.72
SMN	27.2	30.3	35.8	36.5	65.2	67.1	43.68
KERN	27.1	29.8	36.7	37.4	65.8	67.6	44.07
VCTree	27.9	31.1	38.1	38.8	66.4	68.1	45.07
GPS	**28.4**	**31.7**	39.2	40.1	66.9	**68.8**	45.85
Ours	28.2	**31.7**	**39.7**	**40.3**	**67.1**	68.6	**45.93**

where c_{uv} denotes the normalization factor.

The objects and their related relationships are recognized by the above model based on these output features. Specifically, the categories of the detected objects are classified by the node features of the primal graph, and the relationships of multiple objects are detected by the node features of dual graph. Noting that we have added an edge for each node pointed to itself. Our primal-dual graph attention network captures the contextual cues and the dependencies of objects and relationships in graph inference process, which dramatically improves the performance of scene graph generation.

4. EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed method on the visual genome datasets from the aspects of scene graph generation tasks and ablation study compared with several state-of-the-art methods.

4.1. Experimental settings and evaluation indicators

Dataset. We evaluate our model on the Visual Genome (VG) dataset [19], which is the largest and most commonly used benchmark in scene graph generation. VG dataset includes 33877 objects categories and 108077 images in total. Similar with [7, 8, 9, 10], we select the most frequent 151 object categories, including background, and 50 predicate relationships as the evaluation criteria. Within these categories and predicate relationships, an image contains 11.5 objects and 6.2 relations on average. Besides, same as [7], we also adopt the division method as the pre-processing and data partitioning methods in our experiments. We randomly select 70% of the images as the training set and the remaining 30% of the images as the test set.

Task settings. Scene graph aims to predict a series of subject-relationship-object triples and generate the semantic graph. We evaluate the model through the following three task setups: (1) Predicate classification (PredCls): given the object category and bounding box, the model predicts the relationship labels; (2) Scene graph classification (SGCls): given the object bounding box, the model predicts the object category and the relationship labels; (3) Scene graph generation (SGGen): model needs to detect and identify the objects, and predict the relationship labels.

Evaluation metrics. Similar with traditional models, R@K is the recall value at top K and usually used as the main performance metric, which measures the truly recalled instance among the top K most confident triplet predictions. Chen et al. [10] proposed to use the mean recall@K (mR@K) to evaluate the performance of all relationships more comprehensively. Therefore, we adopt the Recall@K and mean Recall@K measures to evaluate our model in experiments, including R@50, R@100 and mR@50, mR@100.

4.2. Comparison with other state-of-the-art methods

We compare the performance of our model on VG dataset with the following start-of-the-art methods, such as Visual Relationship Detection (VRD) model without context [16], Graph RCNN model (GRNN) [9], Iterative Message Passing method (IMP) [7] and its improved version (IMP+) [7], associative embedding model (AE) [14], the best frequency baseline (Freq) in [8], Stacked Motif Networks (SMN) [8], Knowledge-Embedded Routing Network (KERN) [10], visual context tree model (VCTree) [11] and GPS-Net (GPS) [15]. VRD, IMP, and IMP+ methods use language models, message passing, and global context to extract relationship features. Freq method predicts the most frequent relationship between the object pairs who have given the category label. KERN model fuses the statistical correlation of object pairs and their relationships with deep neural networks. The VCTree model fuses the appearance features with the spatial features and constructs a dynamic tree structure. Compared with them, PRDGAT extracts comprehensive semantic features, composes the robust graph structure, and captures the dependencies of objects and relationships based on the Primal-Dual graph attention network.

The comparison results are presented in Table 1, and the best performance is highlighted in boldface. As shown in Table 1, our proposed model outperforms other start-of-the-art models in most of cases. Although the pipline of our model is similar to Graph RCNN, our PRDGAT still obtains 12.66% improvement than Graph RCNN. Because it composes the robust graph structure and captures the dependencies of objects and relationships, our model can construct the sparse graph with more reliable and semantic connected edges than Graph RCNN models, which is the foundation for graph inference and message passing. Besides, SMN implicitly captures the statistical correlation between objects by encoding the global context and obtains 43.68% average Recall performance, which is higher 2.96% than Freq. Noting that the statistical correlation plays a vital role in scene graph tasks. Thus, our model gets 45.93% mean recall value, which is higher about 1.86%, 0.86% and 0.1% than KERN, VCTree and GPS-Net, respectively.

Table 2. Comparison our model with state-of-the-art methods on mR@50 and mR@100.

	SGGen		SGCls		PredCls		
Method	mR50	mR100	mR50	mR100	mR50	mR100	Mean
IMP	0.6	0.9	3.1	3.8	6.1	6.8	3.75
IMP+	3.8	4.8	5.8	6.0	9.8	10.5	6.78
SMN	5.3	6.1	7.1	7.6	13.3	14.4	8.97
KERN	6.4	7.3	9.4	10.0	17.7	19.2	11.67
VCTree	6.9	8.0	10.1	10.8	17.9	19.4	12.18
Ours	**7.3**	**8.5**	**10.8**	**11.5**	**19.2**	**21.0**	**13.05**

Table 3. Comparison our model with other methods on mR@50 and mR@100 without graph constraint.

	SGGen		SGCls		PredCls		
Method	mR50	mR100	mR50	mR100	mR50	mR100	Mean
AE	1.6	2.5	6.0	7.8	15.1	19.5	8.75
IMP+	5.4	8.0	12.1	16.9	20.3	28.9	15.27
Freq	5.9	8.9	13.5	19.6	24.8	37.3	18.33
SMN	9.3	12.9	15.4	20.6	27.5	37.9	20.60
KERN	11.7	16.0	19.8	26.2	36.3	49.0	26.50
Ours	**12.9**	**17.1**	**21.8**	**28.5**	**39.2**	**51.0**	**28.42**

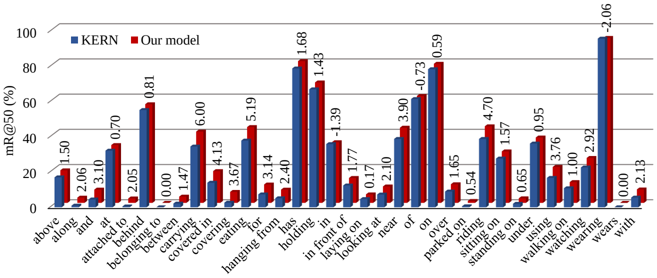

Fig. 3. The mR@50 improvement in PredCls of our model and KERN on the Top-35 frequency relationships.

The distribution of different relationships in VG dataset is uneven and imbalanced. We also use mean Recall@K (mR@K) to evaluate the performance of relationship identification. The results of mR@50 and mR@100 on the VG dataset for the three tasks are shown in Table 2. Our model also gets the best results under the mR@K measure. Specifically, the average mRecall is 13.05%, a 1.38% improvement over KERN and a 0.87% improvement over VCTree. For a more comprehensive comparison, we conduct an extra experiment on mR@50 and mR@100 without graph constraint in Table 3, which implies that for each pair of objects, all possible predicates are valid candidates. And our RPDGAT still achieves outstanding performance in this metric. Hence, our model also has an obvious effect on long-tailed distribution. Figure 3 shows the mR@50 improvement in PredCls of our model and KERN on the Top-35 frequency relationships.

4.3. Ablation study

In our model, we mainly improve the scene graph generation through three aspects: the fusion and refining of the multiple comprehensive semantic features (CSF), construct-

Table 4. The results of ablation studies with three different tasks on R@50 and R@100.

		PD-	SGGen		SGCls		PredCls		
CSF	RSG	GAT	R50	R100	R50	R100	R50	R100	Mean
✓	✓	×	23.7	28.2	33.0	35.0	59.4	62.8	40.35
✓	×	✓	27.0	30.5	34.8	35.5	63.2	65.1	42.68
×	✓	✓	28.0	31.4	37.9	38.6	66.1	67.9	44.98
✓	✓	✓	**28.2**	**31.7**	**39.7**	**40.3**	**67.1**	**68.6**	**45.93**

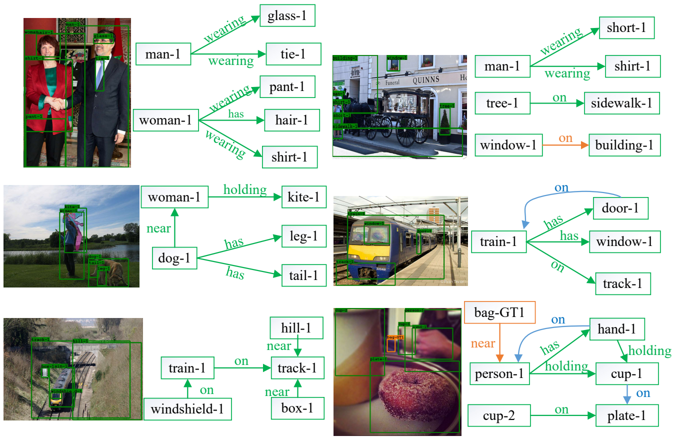

Fig. 4. The qualitative results of our model. Green boxes and arrows indicate the correct results of objects and relationships predicted by our model, while yellow boxes and arrows indicate that the false and mismatching predicted results. The blue edges are false positives.

ing the robust sparse graph structures (RSG), and primal-dual graph attention network (PD-GAT). To study the effectiveness of these modules, we analyze the impact of each module on the model performance in detail through ablation research. We design three different experiments: replacing the CSF with the ROI features, replacing the SSG with the RePN in Graph RCNN, and replacing the PD-GAT with the vanilla GCN [20, 21]. The experimental results of their ablation study are shown in Table 4 and the three leftmost columns indicate whether the model uses CSF, SSG, and PD-GAT.

From Table 4, we know that the primal-dual graph attention network has made the greatest contribution to our model. Under its influence, the performance of the model has increased from 40.35% to 45.93%, which demonstrates that the contextual cues and dependencies of objects and relationships play a vital role in scene graph generation. Besides, the robust sparse graph generates more reasonable edges for efficient message passing, and it gets 3.25% increase than other prune structure. Meanwhile, the model extracting comprehensive semantic features increases the mean score by 0.95% compared with ROI features. Figure 4 shows the qualitative results of our model. The results in Figure 4 also verify that our proposed model can distinguish the accurate and meaningful relationships between multiple objects and identify their corresponding relationship labels.

5. CONCLUSION

In this paper, we propose a novel Relationship-aware Primal-Dual Graph Attention Network (RPDGAT) to address some shortcomings of traditional scene graph generation. Specifically, the comprehensive semantic features are extracted to consolidate the inherent attributes of objects. Besides, the robust sparse graph is automatically built to achieve efficient graphical message passing. Furthermore, the primal-dual graph attention network is designed for sparse graph inference, which captures contextual cues and the dependencies of objects and relationships. Finally, the experimental results validate the effectiveness of our RPDGAT on real-world visual genome datasets. In the future, we plan to extend our method to apply into the practical areas, such as image retrieval [1] and VQA [3].

6. REFERENCES

[1] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei, "Image retrieval using scene graphs," in *CVPR*, 2015, pp. 3668–3678.

[2] Senmao Ye, Junwei Han, and Nian Liu, "Attentive linear transformation for image captioning," *IEEE TIP*, vol. 27, no. 11, pp. 5514–5524, 2018.

[3] Drew A Hudson and Christopher D Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019, pp. 6700–6709.

[4] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon, "Linknet: Relational embedding for scene graph," in *NeurIPS*, 2018, pp. 560–570.

[5] Zih-Siou Hung, Arun Mallya, and Svetlana Lazebnik, "Contextual translation embedding for visual relationship detection and scene graph generation," *IEEE TPAMI*, 2020, doi: 10.1109/TPAMI.2020.2992222.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.

[7] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei, "Scene graph generation by iterative message passing," in *CVPR*, 2017, pp. 5410–5419.

[8] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi, "Neural motifs: Scene graph parsing with global context," in *CVPR*, 2018, pp. 5831–5840.

[9] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh, "Graph r-cnn for scene graph generation," in *ECCV*, 2018, pp. 670–685.

[10] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin, "Knowledge-embedded routing network for scene graph generation," in *CVPR*, 2019, pp. 6163–6171.

[11] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu, "Learning to compose dynamic tree structures for visual contexts," in *CVPR*, 2019, pp. 6619–6628.

[12] Bo Dai, Yuqi Zhang, and Dahua Lin, "Detecting visual relationships with deep relational networks," in *CVPR*, 2017, pp. 3076–3086.

[13] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang, "Scene graph generation from objects, phrases and region captions," in *CVPR*, 2017, pp. 1261–1270.

[14] Alejandro Newell and Jia Deng, "Pixels to graphs by associative embedding," in *NeurIPS*, 2017, pp. 2171–2180.

[15] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao, "Gps-net: Graph property sensing network for scene graph generation," in *CVPR*, 2020, pp. 3746–3753.

[16] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei, "Visual relationship detection with language priors," in *ECCV*, 2016, pp. 852–869.

[17] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang, "Factorizable net: an efficient subgraph-based framework for scene graph generation," in *ECCV*, 2018, pp. 335–351.

[18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," in *ICLR*, 2018.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[20] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun, "Spectral networks and locally connected networks on graphs," in *ICLR*, 2014.

[21] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.