# PMP-NET: RETHINKING VISUAL CONTEXT FOR SCENE GRAPH GENERATION

*Xuezhi Tong[1], Rui Wang[2,3,*], Chuan Wang[3], Sanyi Zhang[3], Xiaochun Cao[3]*

[1]College of Intelligence and Computing, Tianjin University
[2]Zhejiang Lab
[3]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences

## ABSTRACT

Scene graph generation aims to describe the contents in scenes by identifying the objects and their relationships. In previous works, visual context is widely utilized in message passing networks to generate the representations for classification. However, the noisy estimation of visual context limits model performance. In this paper, we revisit the concept of incorporating visual context via a randomly ordered bidirectional Long Short Temporal Memory (biLSTM) based baseline, and show that noisy estimation is worse than random. To alleviate the problem, we propose a new method, dubbed Progressive Message Passing Network (PMP-Net) that better estimates the visual context in a coarse to fine manner. Specifically, we first estimate the visual context with a random initiated scene graph, then refine it with multi-head attention. The experimental results on the benchmark dataset Visual Genome show that PMP-Net achieves better or comparable performance on all three tasks: scene graph generation (SGGen), scene graph classification (SGCls), and predicate classification (PredCls).

***Index Terms*—** Scene graph generation, visual context, multi-head attention, message passing

## 1. INTRODUCTION

Scene graph generation is a fundamental visual understanding task, which aims to describe scenes by detecting objects and predicting the relationships between objects in images. Because of the rich semantic information provided by scene graph generation, it can be applied to tasks ranging from basic ones like multi-label classification [1] and image retrieval [2] to high-level ones like image captioning [3], visual question answering [4] and visual reasoning[5].

Based on the studies on relationship detection [6, 7], early attempts [8, 9] propose to model visual context via message passing. In some works [10, 11], a scoring network is adopted to generate "relation proposals" for better context structure estimations. More recent works [12, 13] focus on the problem of long-tailed distribution of predicate categories in the Visual Genome dataset (VG) [14]. However, the performance is still limited even for the most frequent categories (Recall@100 below 50% for SGCls and SGGen on the VG dataset [14]). This fact shows that there still remains large room for improvement on predicting frequent categories [15].

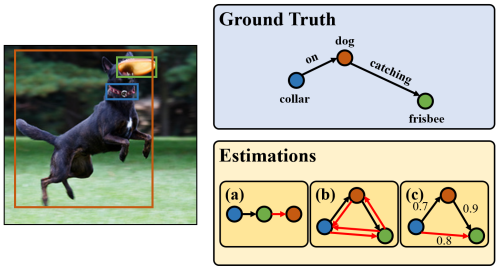


**Fig. 1**. Illustration of the existing methods for approximating the visual context in images. The three estimations are: (a) sequential graph ordered by a certain rule (bounding box size in this case); (b) fully connected graph; (c) scene graph estimated from a scoring network. The black arrows represent the estimated connections that are consistent with ground truth, while the red ones are wrong predictions.

To alleviate the aforementioned problem, we study and rethink the concept of "modeling visual context", which has been widely used in the community rooted in early attempts [8]. The modeling of visual context is always achieved by applying pre-defined or estimated graphs to message passing networks. This means these graphs are regarded as the structure of the visual context. However, a critical problem is that the actual structure is not available beforehand, because determining whether two objects are related is also a part of the task. As illustrated in Fig. 1, existing works consists of three types of approximations of visual context: fully connected graph for GNN-based methods [1, 8], sequential graph for RNN-based methods [9, 12], estimated graph predicted by an edge scoring network [10, 11]. Apparently, neither pre-

This work is supported by the National Key R&D Program of China Grant No. 2020AAA0109304, the National Natural Science Foundation of China Under Grants No. 6217024002 and No.U20B2066, the Open Research Projects of Zhejiang Lab (No. 2021KB0AB01).

*Corresponding author

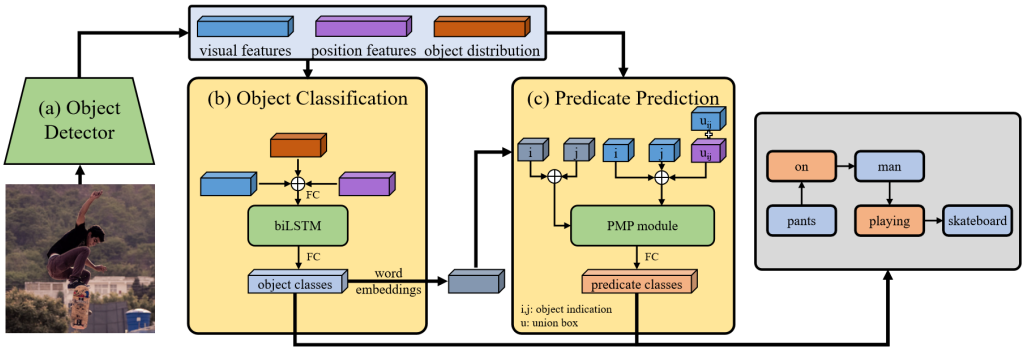


**Fig. 2**. Illustration of the framework of the proposed method. For the biLSTM-based baseline, the PMP Module is replaced with two LSTMs followed by a biLSTM.

defined graph structure nor estimated graph is optimal for defining the visual context, due to the intra-image variance of scenes and the label incompleteness of current datasets. We introduce a strong baseline by modeling the visual context using LSTMs with random-ordered nodes. The performance of the baseline is comparable to state-of-the-art, which shows that there still remains large room for improvement for estimating the visual context.

Based on our analysis, we propose a novel scene graph generation architecture, Progressive Message Passing Network (PMP-Net), that explores the visual context in a coarse to fine manner. The PMP module is constructed as follows: For the first step, visual and semantic information are encoded via two separate LSTMs with random ordered graphs to coarsely model the visual context in a manner of random walk. For the second step, a multi-head attention layer is adopted to further capture visual context by learning the correlations between all the nodes in a graph.

Unlike previous methods [10, 11] that supervise independent visual context estimation networks with ground-truths, we estimate the visual context implicitly based on the refined features from the LSTMs. Since object pairs in the images commonly share subject/object with others, the nodes in the message passing network for predicate prediction are highly correlated globally. As a result, we use a multi-head attention layer, which learns global self-attention, to explore the visual context for predicate prediction.

Experiments on the VG dataset [14] demonstrate the efficacy of the proposed PMP-Net. The results show that the biLSTM-based framework alone obtains superior performance. With the adoption of the PMP module for the predicate prediction, the performance achieves a further improvement.

## 2. METHOD

In this section, we formulate the task of scene graph generation in Sec. 2.1, and introduce a strong baseline based on biLSTM in Sec. 2.2. Finally, we describe the proposed PMP-Net in Sec. 2.3. Fig. 2 provides an overview of the PMP-Net: An object detector is used to generate region proposals and extract the corresponding features. A biLSTM is used for object classification and a PMP module is used for predicate prediction.

### 2.1. Problem Formulation

A *scene graph* $G = (U, E)$, is used to describe the contents of a scene. The node set $U = (B, O)$ of the scene graph represents the objects in the scene, where $B$ and $O$ are the bounding box set and class label set of the objects, respectively. The edge set $E$ represents the relationships between objects. The relationship between object $i$ and $j$ is denoted as a triplet $e_{ij} = (o_i, r_{ij}, o_j)$, where $r_{ij} \in \mathcal{R}$ is the type of the predicate including "background", which indicates that there is no relation between the connected objects.

Scene graph generation (SGG) is the task of generating the scene graph $G$ from the image $I$. The probability model of such a task is formulated as:

$$P(G \mid I) = P(B, O, R \mid I). \tag{1}$$

### 2.2. Scene Graph Generation With biLSTM

We construct a framework based on biLSTM as a baseline. Following [9], SGG with biLSTM decomposes the whole task of generating the scene graph $G$ into three sequential steps:

$$P(G \mid I) = P(B \mid I)P(O \mid B, I)P(R \mid B, O, I). \tag{2}$$

**Bounding box prediction.** The prediction of bounding boxes ($P(B \mid I)$) is based on an off-the-shelf detection model (such as Faster R-CNN [16]), which generates region proposals $B = \{b_1, \cdots, b_n\}$. In addition, the corresponding visual features for the objects $F^o = \{f_1^o, \cdots, f_n^o\}$ are extracted and object distributions $O^d$ are predicted by the detector.

**Object classification.** Inspired by [17], the object classification model ($P(O \mid B, I)$) concatenates the position features, the visual features, and the prior object distributions of

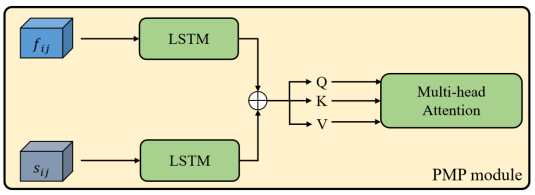


**Fig. 3**. The illustration of the proposed Progressive Message Passing module. The visual feature $f_{ij}$ and semantic feature $s_{ij}$ are fed into an two separate LSTMs to encode the context. A multi-head attention layer is then used to calculate the self-attended features to incorporate the context precisely.

the detected regions, and feeds them into a fully connected layer. The resulting object features are then refined by a biLSTM, and fed into another fully connected layer to make final predictions. Word embeddings are extracted from the results of the object classifier.

**Predicate prediction.** The predicate prediction model ($P(R \mid B, O, I)$) enumerates all possible pairs of subject and object, and extracts union box feature $u_{ij} \in R^{512}$ for each pair of subject $i$ and object $j$. The spatial masks of the union boxes are fed into a sequential of convolution layers and a global pooling layer to get representations having the same size as the features of the union boxes. The spatial and visual features of the union boxes are added to get the features of the union boxes. The visual features of $i$ and $j$ are concatenated with union box features and forwarded into an LSTM. We then get the visual features $f^r_{ij}$ of the predicate for $i$ and $j$. Similarly, the word embeddings of $i$ and $j$ are concatenated and fed into an LSTM to get the semantic features of the predicates. The visual and semantic features of the predicates are then concatenated and fed into a biLSTM followed by a Multi-Layer Perceptron (MLP) to get the predicate prediction.

**Analysis.** The simple and straight baseline achieves superior performance (see Sec. 3.3.1). This result may owe to the fact that the biLSTM-based framework utilizes the visual context in a manner of performing random walks on the fully connected scene graphs. This method incorporates more context information than incomplete estimations of the scenes, and includes less noise compared to methods using a fully connected scene graph.

### 2.3. Progressive Message Passing Network

However, the biLSTM-based framework constrains the degree of a node to be no more than 2, which is inconsistent with the nature of scene graphs. Also, random connections between nodes inevitably involve noises. To overcome the disadvantages, incorporating different types of message passing methods leads us to a better solution.

As illustrated in Fig 3, we combine the LSTMs with a multi-head attention layer to better estimate the visual context for message passing. We define this architecture as a PMP module, and use it as the message passing network for predicate prediction. The LSTMs encode context information with noise, while the multi-head attention layer models the visual context more precisely.

**Algorithm 1** The algorithm of PMP-Net

**Require:** Image set $I$
  $B, O^d, F^o = detector(I)$ ▷ Bounding box prediction
  **for** each object $i$ **do** ▷ Object classification
    $\overline{f}^o_i \leftarrow FC(W^b b_i, W^d o^d_i, f^o_i)$
    $\tilde{f}^o_i \leftarrow biLSTM(\overline{f}^o_i)$
    $o_i \leftarrow FC(\tilde{f}^o_i)$
  **end for**
  **for** each object pair $\{i, j\}$ **do** ▷ Predicate predication
    $\overline{u}_{ij} \leftarrow u_{ij} + m_{ij}$
    $s_{ij} \leftarrow cat(s_i, s_j)$
    $f_{ij} \leftarrow cat(f^o_i, f^o_j, \overline{u}_{ij})$
    $f^r_{ij} \leftarrow cat(LSTM(s_{ij}), LSTM(f_{ij}))$
    $\tilde{f}^r_{ij} \leftarrow MH(f^r_{ij})$
    $r_{ij} \leftarrow FC(\tilde{f}^r_{ij})$
  **end for**
  **Output:** $B, O = \{o_1, \cdots, o_m\}, R = \{r_1, \cdots, r_n\}$

As described in Algorithm 1, the proposed PMP-Net is conducted as follows. **(a) Bounding box prediction.** An object detector is used to extract bounding boxes $B$, prior object distributions $O^d$, and object features $F^o$. **(b) Object classification.** The bounding boxes and prior object distributions are transformed by linear matrices $\boldsymbol{W^b}$ and $\boldsymbol{W^d}$, respectively. A fully connected layer $FC$ is used to encode the concatenated features of object $i$. Then, a biLSTM is used to model the visual context for objects and generate representations for object classification. **(c) Predicate prediction.** The mask and visual feature of the union box for $(i, j)$ are $m_{ij}$ and $u_{ij}$, respectively. $m_{ij}$ is fed into a sequence of convolution layers $Conv$ and added with $u_{ij}$. The visual features $f^o_i$, $f^o_j$ and word embedding $s_i$, $s_j$ are encoded by two different LSTMs respectively. The concatenation of visual feature, word embedding, and union box feature are then further encoded by a multi-head attention layer $MH$ for more precise visual context modeling. The resulting features are used for predicate prediction. Finally, we get the scene graph generation result $\{B, O, R\}$ by concluding the aforementioned steps.

## 3. EXPERIMENTS

### 3.1. Dataset and Settings

Visual Genome (VG) dataset [14] is used to evaluate the proposed method on the scene graph generation task. We follow [9] to get a subset of the VG dataset (VG150), which has 75,651 images in training set and 32,422 images in test set.

| Method | SGGen | | | SGCls | | | PredCls | | |
|---|---|---|---|---|---|---|---|---|---|
| | R@20 | R@50 | R@100 | R@20 | R@50 | R@100 | R@20 | R@50 | R@100 |
| IMP [8] | 14.6 | 20.7 | 24.6 | 31.7 | 34.6 | 35.4 | 52.7 | 59.3 | 61.3 |
| Graph R-CNN [10] | - | 11.4 | 13.7 | - | 29.6 | 31.6 | - | 54.2 | 59.1 |
| MotifNet [9] | 21.4 | 27.2 | 30.3 | 32.9 | 35.8 | 36.5 | 58.5 | 65.2 | 67.1 |
| MotifNet-Freq [9] | 20.1 | 26.2 | 30.1 | 29.3 | 32.3 | 32.6 | 53.6 | 60.6 | 62.2 |
| VCTREE [11] | 22.0 | 27.9 | **31.3** | 35.2 | 38.1 | 38.8 | 60.1 | **66.4** | **68.1** |
| KERN [18] | - | 27.1 | 29.8 | - | 36.7 | 37.4 | - | 65.8 | 67.6 |
| CMAT [19] | 22.1 | 27.9 | **31.2** | 35.9 | 39.0 | 39.8 | **60.2** | **66.4** | **68.1** |
| NODIS [17] | 21.6 | 27.7 | 31.0 | 37.7 | 41.7 | 42.9 | 58.9 | 66.0 | 67.9 |
| biLSTM | 21.7 | 27.5 | 30.9 | 37.8 | 41.9 | 42.9 | 58.3 | 65.5 | 67.6 |
| **PMP-Net** | **22.3** | **28.0** | **31.2** | **38.5** | **42.5** | **43.5** | 58.6 | 65.7 | 67.6 |

**Table 1**. **Comparison on the test set of VG150 [9] with graph constraint.** For a fair comparison, all the compared methods share the same object detector [16] with VGG16 [20] as the backbone. "biLSTM" is the biLSTM-based baseline described in Sec. 2.2. The bold numbers represent the best results.

The most frequent 150 object categories and 50 predicate categories are selected.

Following [9], the task of scene graph generation is evaluated on three sub-tasks:

1) **Predicate classification** (PredCls): predict the types of predicates for the object pairs given ground truth bounding boxes and object labels;

2) **Scene graph classification** (SGCls): predict object labels and predicate labels given ground truth bounding boxes;

3) **Scene graph generation** (SGGen): predict the bounding boxes and labels of the objects, and classify the relationships for the object pairs.

**Evaluation.** The bounding boxes of a subject and an object should have more than $50\%$ IoU with the ground truth. A relationship prediction is regarded as correct only when the corresponding object pair and the predicate are classified correctly at the same time. All the relationship prediction candidates are ordered according to the prediction confidences of the objects and predicates. For each sub-task, recall@K (R@K) is used as the evaluation metric.

### 3.2. Implementation Details

For a fair comparison, the object detector used in this paper is Faster R-CNN [16] with VGG16 [20] as the backbone by default. We use the vanilla cross entropy loss for training the models. Following [9], we resize the input images to $592 \times 592$. A ROIAlign layer followed by a global average pooling layer is used to extract visual features for object proposals and their union boxes. For all three tasks, models are warmed up with a learning rate of $10^{-3}$ for 5 epochs, and further trained with a learning rate of $10^{-4}$. When the model plateaus for the first time, the learning rate will be further divided by 0.1.

### 3.3. Quantitative Results and Comparison

The comparison results with the proposed method are shown in Table 1. We compare PMP-Net with previous methods that share the same detector backbone (VGG16 [20] for Faster R-CNN [16]) with it. Considering the graph constraint, PMP-Net shows a state-of-the-art on SGGen and PredCls, whose results have not been improved significantly since [9]. For SGCls, PMP-Net relatively improves state-of-the-art (NODIS [17]) 2.1%(SGCls-R@20), 1.4% (SGCls-R@50) and 1.4% (SGCls-R@100).

#### *3.3.1. Ablation Studies*

In order to certificate the effectiveness of the proposed PMP-Net, we analyze the impact of using different message passing networks for predicate prediction. In Tab. 1, "biLSTM" means the biLSTM-based framework described in Sec. 2.2. We can find that replacing the biLSTM with a multi-head attention layer has positive impacts on all three tasks. The good results suggest that the two steps of context modeling are able to complement each other. Also, the simple biLSTM-based baseline shows a similar performance to a prior state-of-the-art method [17].

## 4. CONCLUSION

This paper revisits the concept of visual context in scene graph generation and contributes to understanding how the visual context is utilized. Motivated by the analysis, we introduced a strong baseline and further improved it by exploring the visual context from coarse to fine. Experimental results show that while the baseline achieves superior results, the proposed method provides additional gains over it and gets up to 2.1% relative improvements over state-of-the-art.

## 5. REFERENCES


[1] Marino K., Salakhutdinov R., and Gupta A., "The more you know: Using knowledge graphs for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 2673–2681.

[2] Johnson J., Krishna R., Stark M., Li L.J., Shamma D., Bernstein M., and Fei-Fei L., "Image retrieval using scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2015, pp. 3668–3678.

[3] Yang X., Tang K., Zhang H., and Cai J., "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 10685–10694.

[4] Teney D., Liu L., and van den Hengel A., "Graph-structured representations for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, p. 3233–3241.

[5] Shi J., Zhang H., and Li J., "Explainable and explicit visual reasoning over scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, p. 8376–8384.

[6] Dai B., Zhang Y., and Lin D., "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, p. 3076–3086.

[7] Liang X., Lee L., and Xing E.P., "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, p. 848–857.

[8] Xu D., Zhu Y., Choy C.B., and Fei-Fei L., "Scene graph generation by iterative message passing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, p. 5410–5419.

[9] Zellers R., Yatskar M., Thomson S., and Choi Y., "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, p. 5831–5840.

[10] Yang J., Lu J., Lee S., Batra D., and Parikh D., "Graph r-cnn for scene graph generation," in *Proceedings of the European Conference on Computer Vision*, 2018, p. 690–706.

[11] Tang K., Zhang H., Wu B., Luo W., and Liu W., "Learning to compose dynamic tree structures for visual contexts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 6619–6628.

[12] Tang K., Niu Y., Huang J., Shi J., and Zhang H., "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2020, p. 3716–3725.

[13] Yan S., Shen C., Jin Z., Huang J., Jiang R., Chen Y., and Hua X., "Pcpl: Predicate-correlation perception learning for unbiased scene graph generation," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, p. 265–273.

[14] Krishna R., Zhu Y., Groth O., Johnson J., Hata K., Kravitz J., Chen S., Kalantidis Y., Li L.-J., Shamma D. A., and et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," in *International Journal on Computer Vision*, 2017, vol. 123, pp. 32–73.

[15] Li R., Zhang S., Wan B., and He X., "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 11109–11119.

[16] Ren S., He K., Girshick R., and Sun J., "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[17] Cong Y., Ackermann H., Liao W., Yang M. Y., and Rosenhahn B., "Nodis: Neural ordinary differential scene understanding," in *Computer Vision–ECCV 2020: 16th European Conference*, 2020, Part XX 16, pp. 636–653.

[18] Chen L., Zhang H., Xiao J., He X., Pu S., and Chang S.F., "Knowledge-embedded routing network for scene graph generation," in *Proceedings of the IEEE International Conferenceon Computer Vision*. IEEE, 2019, pp. 6163–6171.

[19] Chen T., Yu W., Chen R., and Lin L., "Counterfactual criticmulti-agent training for scene graph generation," in *Proceedings of the IEEE International Conferenceon Computer Vision*. IEEE, 2019, p. 4613–4623.

[20] Simonyan K. and Zisserman A., "Very deep convolutional networks for large-scale image recognition," in *arXiv preprint arXiv:1409.1556*, 2014.