Divide and Conquer: Subset Matching for Scene Graph Generation in Complex Scenes

XIN LIN, JINQUAN ZENG, XINGQUAN LI
1School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
2School of Microelectronics of Shenzhen Institute of Information Technology, Shenzhen, China

ABSTRACT The goal of scene graph generation (SGG) is to classify objects and their pair-wise relationships in a visual scene. Object occlusion is a critical challenge when generating scene graphs in complex scenes. However, this issue has rarely been explored in recent works. Accordingly, in this paper, we propose a subset matching network (SM-Net) that handles the above problem. First, we decompose SGG into two types of subset matching problems: node subset matching and edge subset matching. Each node/edge subset handles the occlusion between one node/edge pair, thereby reducing the difficulty of SGG in a "divide and conquer" manner. Second, we introduce a node subset prediction module that utilizes a subset-based message passing module to refine the node subset representation and a matching loss to supervise node subset prediction. Third, we propose an edge subset prediction module that applies a feature selection-based fusion function to obtain edge subset features and a matching loss to supervise edge subset predictions. Experiments on three popular datasets show that our model achieves state-of-the-art performance. The code of SM-Net will be released.

INDEX TERMS Scene graph generation, object occlusion, subset matching, message passing

I. INTRODUCTION

Scene graph generation (SGG) aims to provide a graph-based representation for a given image. A scene graph consists of objects as nodes and relationships as edges. As illustrated in the dotted box of Figure 1, two objects and a relationship form a triplet subject-predicate-object. Specifically, an object is denoted as a node with a category label, and the relationship between two nodes is characterized by a directed edge between them. The edge has a specific category of predicates and points from the subject towards the object. Recent works have shown the benefits of SGG for object detection [1], semantic segmentation [2] and high-level vision tasks, e.g., 3D scene understanding [3], visual question answering [4], image captioning [5].

Despite the effectiveness of existing SGG works, the application of SGG in complex scenes involving many objects remains a challenging task. Specifically, occlusion between objects is more likely to occur when their number increases. Object occlusion usually results in a large intersection over union (IoU) between the bounding boxes of nearby nodes. As shown in Figure 1(a), a large proportion of triplets in existing datasets [6]–[8] suffer from the occlusion problem.

Unfortunately, existing SGG works [9]–[14] typically make predictions for each element in the scene graph independently, without considering the challenges that arise due to occlusion. Consequently, as shown in Figure 1(b), these SGG methods [9]–[14] suffer from a performance degradation when the occlusion level increases. More specifically, this is because occlusion reduces the difference between the appearances of nearby nodes, making prediction in both the object and relationship categories more challenging. As illustrated in Figure 1(c), the bounding boxes of fork and knife are highly overlapped. It is hard to distinguish between these two nodes. Moreover, as shown in Figure 1(d), due to the bounding boxes of man and boy having large IoU, it is difficult to predict their related relationships (e.g., man-ride-bike and boy-sit-on-bike).

We discover that grouping occluded nodes/edges into a subset could aid the network in learning which types of nodes/edges frequently appear together, thus improving occluded nodes/edges prediction performance. Accordingly, we propose a subset matching network (SM-Net) that jointly predicts the categories of nodes and edges in a scene graph rather than predicting each of them individually. In more detail,
TABLE I: Ratio of triplets under different levels of occlusion.

| Occlusion | L0: 0-0.3 | L1: 0.3-0.5 | L2: 0.5-0.7 | L3: 0.7-1.0 |
|-----------|-----------|-------------|-------------|-------------|
| VG        | 98.126 (22.1%) | 138.483 (31.4%) | 113756 (25.9%) | 90907 (20.6%) |
| OI        | 59.839 (34.2%) | 42337 (24.2%) | 56761 (32.4%) | 15917 (9.0%) |
| VRD       | 12.481 (33.4%) | 12057 (32.3%) | 7873 (21.0%) | 4952 (13.3%) |

(a) The ratio of triplets under different levels of occlusion.
(b) Scene graph classification under different levels of occlusion.

In summary, the innovation of the proposed SM-Net is three-fold: (1) the SGG problem is decomposed into node subset matching and edge subset matching problems; (2) an NSP module that utilizes subset-based message passing to refine the node subset representation and a matching loss to minimize the gap between node subset prediction and ground-truth; and (3) an ESP module that applies subset-based edge fusion function to obtain the edge subset representation and a matching loss to minimize the gap between edge subset prediction and ground-truth. The effectiveness of SM-Net is systematically evaluated on three popular SGG datasets: Visual Genome (VG) [6], OpenImages (OI) [8], and Visual Relationship Detection (VRD) [7]. Experimental results demonstrate that the proposed SM-Net consistently outperforms the state-of-the-art methods.

The remainder of this paper is organized as follows: First, related works on SGG and graph similarity learning are briefly reviewed in Section II. Second, the SM-Net problem formulation and model structure are presented in Section IV. Third, detailed experiments are conducted and analyzed in Section V. Finally, we conclude this paper in Section VI.
tri-linear model to encode edge information into an attention map that guides message passing.

However, the above message-passing methods are based on individual objects, which may not work well for objects suffering from heavy occlusion. As discussed in Section I, occluded objects usually have a similar visual appearance; accordingly, message-passing modules will encode similar contextual information for these objects, achieving only limited benefit for occluded object prediction. In a departure from existing work, we group occluded objects into subsets and then refine their features via a subset-based message-passing method. Subsequently, we jointly predict the object categories in each subset.

B. GRAPH SIMILARITY LEARNING

Over the past few decades, many techniques have been developed for studying the similarity of graphs. In the early stages, multiple graph similarity metrics were defined, including Graph Edit Distance [15], Maximum Common Subgraph [16], and Graph Isomorphism [17] to address the graph similarity search and graph matching problems. However, the computation of these metrics is generally an NP-complete problem [19]; therefore, these approaches are only feasible for graphs of relatively small sizes. More recently, graph similarity learning has also been explored for applications in computer vision. In [20], context-dependent graph kernels are proposed to measure the similarity between graphs for human action recognition in video sequences. In [21], a deep model called Neural Graph Matching Network is first introduced for the 3D action recognition problem in the few-shot learning setting. Under this approach, interaction graphs are constructed from 3D scenes; here, the nodes represent physical entities in the scene, while edges represent interactions between the entities.

In this paper, we reduce the complexity of graph similarity measurement by decomposing the entire node and edge sets in a scene graph into subsets, respectively. Besides, current works focus primarily on measuring the similarity of unlabeled graphs whose nodes and edges do not have category labels [22]. In comparison, the nodes and edges in a scene graph own their specific categories and locations. Therefore, we jointly consider the bounding box regression and the node/edge classification in the matching loss.

III. MOTIVATION AND PROBLEM FORMULATION

Existing works [9], [11], [14] typically classify each element of a scene graph independently. However, this strategy may not be suitable for generating scene graphs in complex scenes. As illustrated in Figure 1, the occlusion issue, which frequently occurs in complex scenes, may cause errors during node and edge prediction. Therefore, it is more appropriate to jointly predict their categories as a set. Accordingly, we formulate the SGG problem as two set matching problems, i.e., node set matching and edge set matching, as follows:

\[ \mathcal{L}_{SGG} = D(\tilde{O}, O) + D(\tilde{R}, R), \]  

where \( D \) represents a metric measuring the distance between the predicted and ground-truth sets. The \( O \) and \( R \) stand for the ground-truth node sets and edge sets, respectively, with \( \tilde{O} \) and \( \tilde{R} \) being their predictions. Eq. (1) aims to make predictions for the entire node set and the entire edge set in the scene graph, respectively. Attempting to solve Eq. (1), traditional set prediction methods mainly utilize the Hungarian algorithm [44] to match the predicted set with its ground-truth. However, the complexity of the Hungarian algorithm increases polynomially as the number of elements in the set increases. Moreover, as mentioned in [45], a large number of training epochs is required to optimize Eq. (1). Therefore, it is necessary to explore approximations.

One intuitive way of approximating Eq. (1) would be to decompose it into several subset matching problems. As illustrated in Figure 2, we regard two nodes whose bounding box IoU exceeds 0.5 as a crowded node subset \( \tilde{O}_c \). Specifically, according to the IoU value in descending order, each node has only one chance to be selected in building node subsets. Each of the remaining nodes that have small IoU with the others is duplicated to build a sparse node subset \( \tilde{O}_s \). Similarly, we categorize all edges with the same subject into four types of edge subsets. (1) \( \tilde{R}_c \): the edges connecting two nodes within one crowded node subset; (2) \( \tilde{R}_{cs} \): the edges connecting one crowded node subset and sparse node subset; (3) \( \tilde{R}_{sc} \): the edges connecting two crowded node subsets; (4) \( \tilde{R}_{ss} \): the edges connecting two sparse node subsets. Note that for \( \tilde{R}_{cs} \), the edges with the same object are also unified into one node subset. More details can be found in Figure 2.

Finally, using the above decomposed node and edge subsets, we can efficiently approximate Eq. (1), as follows:

\[ \mathcal{L}_{SM} = \frac{1}{N} \sum_{n=1}^{N} D(\tilde{O}_n, O_n) + \frac{1}{Q} \sum_{q=1}^{Q} D(\tilde{R}_q, R_q), \]

where \( O_n \) and \( R_q \) denote the predicted node and edge subsets, respectively; \( O_n \) and \( R_q \) represent the ground-truth node and edge subsets, respectively; \( N \) and \( Q \) stand for the number of subsets in the entire node set \( O \) and edge set \( R \), respectively.

IV. METHOD

A. OVERVIEW OF SM-NET

Figure 2 illustrates the framework of SM-Net. We employ Faster R-CNN [24] to obtain object proposals for each image. We adopt exactly the same method used in [10] to obtain features for each proposal: these include the appearance feature, the vector of the object class prediction scores, and the relative coordinate of the proposal’s bounding box in the image, which are represented as \( v_i \in \mathbb{R}^{512} \), \( s_i \in \mathbb{R}^{C_o} \), and \( b_i \in \mathbb{R}^4 \) for the \( i \)-th node, respectively. \( C_o \) further denotes the number of object categories (including the background, i.e., no object). Moreover, we extract the appearance feature from the union area of the two nodes \( i \) and \( j \) in the \( n \)-th node subset, denoted as \( v_{ij} \in \mathbb{R}^{512} \). Following [10], we further extract the spatial feature, denoted as \( B_{ij} \in \mathbb{R}^{512} \) for the two nodes \( i \) and \( j \) to describe their geometric relation; similarly,
the relative location features for the union area of one node. Next, we apply two convolutional layers and two FC layers to the binary maps, finally obtaining a 512-dimensional representation. $\mathcal{B}_{nm}$ is obtained in the same way as $\mathcal{B}_{ij}$, with each binary channel representing the area of a single node subset.

In this section, we will sequentially introduce the two modules of SM-Net. First, we introduce the node subset prediction (NSP) module in Section IV-B, then present the edge subset prediction (ESP) module in Section IV-C. Finally, the overall loss function and the SGG inference method are described in Section IV-D.

**B. NODE SUBSET PREDICTION**

NSP jointly predicts the two nodes in each node subset. Its predictions for the $n$-th node subset are denoted as follows:

$$\hat{O}_n = \left\{ (\hat{a}^{(1)}_n, \hat{b}^{(1)}_n), (\hat{a}^{(2)}_n, \hat{b}^{(2)}_n) \right\},$$

where $\hat{a}^{(1)}_n$ and $\hat{a}^{(2)}_n \in \mathbb{R}^{C_o}$ represent object class predictions for the two nodes, respectively. Moreover, $\hat{b}^{(1)}_n$ and $\hat{b}^{(2)}_n \in \mathbb{R}^4$ denote bounding box coordinates that associate each predicted object label with one of the nodes. NSP contains two main components: Subset-based Message Passing and Node Matching Loss.

**Subset-based Message Passing.** Existing message-passing methods aim to refine the node representation with contextual information. Taking the GCN-based SGG models [13], [14], [30], [58] as an example, the feature for the $i$-th node is refined by the surrounding nodes as follows:

$$x_i^{mp} = x_i + W_{t1} \left( \sum_{j \in \mathcal{N}_i} \alpha(x_i, x_j)W_{t2}x_j \right),$$

where $\mathcal{N}_i$ denotes the set of nodes that are neighbors of the $i$-th one. $W_{t1}$ and $W_{t2}$ are projection matrices. $\alpha(x_i, x_j)$ computes the attention weight between nodes $i$ and $j$. One popular way of computing this weight is as follows [13], [27]:

$$\alpha_{ij} = \exp \left( \frac{\mathbf{u}^T [x_i; x_j]}{\sum_{m \in \mathcal{N}_i} \exp (\mathbf{u}^T [x_i; x_m])} \right),$$

where $\mathbf{u}$ is a projection vector and $[\cdot]$ represents the concatenation operation. Due to the highly similar appearance features, two nearby nodes tend to obtain relatively high attention weights regardless of whether their object categories are related. Consequently, it may still hard to distinguish between spatially proximate nodes through the above message passing model.

To deal with this issue, we propose a subset-based message-passing (SMP) module to refine the features of node subsets. Firstly, we merge the multi-modal features of two nodes to represent their subset. Specifically, the feature for the $n$-th node subset, which includes nodes $i$ and $j$, can be obtained as follows:

$$\mathbf{u}_n = W_{u1}[\mathbf{v}_i; \mathbf{v}_j] \odot W_{u2}[\mathbf{s}_i; \mathbf{s}_j] \odot W_{u3}v_{ij},$$

where $\odot$ represents Hadamard product. $W_{u1} \in \mathbb{R}^{512 \times 1024}$, $W_{u2} \in \mathbb{R}^{512 \times 1024}$, and $W_{u3} \in \mathbb{R}^{512 \times 512}$ are the projection matrices. Secondly, for the $n$-th node subset, we aggregate all its related pairwise messages as $c_n$. More specifically, $c_n$ is obtained via the following three steps: (1) we calculate the message sent from the $m$-th node subset as follows:

$$\mathbf{z}_{nm} = W_{z1} \mathbf{u}_n \odot W_{z2} \mathbf{u}_m + W_{z3} \mathbf{B}_{nm},$$
where \( W_{z1}, W_{z2}, W_{z3} \in \mathbb{R}^{512 \times 512} \) are learnable projection matrices. The first term on the right-hand side of the equal sign aims to encode the correlations between the visual features of the two node subsets, while the second term learns their relative spatial relation; (2) We stack all \( N - 1 \) messages to create a matrix \( Z_n \in \mathbb{R}^{512 \times (N-1)} \); (3) We apply row-wise max-pooling to \( Z_n \) and obtain the overall contextual information \( c_n \). Finally, we update the feature for the \( n \)-th node subset as follows:

\[
\hat{U}_n = U_n + c_n. 
\]

Compared with existing message passing mechanisms [13], [14], [30] defined in Eq. (4), our SMP has three main advantages. First, two nearby nodes with highly similar appearance features are unified into one node subset and will not receive messages from each other. Second, \( z_{nm} \) is able to store more correlation information than a scalar weight. Third, as revealed in [62], the row-wise max-pooling on \( Z_n \) could reduce the noise that is induced by assuming that all node subsets interact with each other.

**Node Matching Loss.** Having obtained the refined representation for each node subset using SMP, we utilize two FC layers to make predictions for the two nodes in each subset as follows:

\[
\begin{align*}
\hat{\sigma}_q^{(1)}; \hat{\sigma}_q^{(2)} = \mathcal{W}_a \hat{U}_n, \\
\hat{b}_q^{(1)}, \hat{b}_q^{(2)} = \mathcal{W}_b \hat{U}_n,
\end{align*}
\]

where \( \mathcal{W}_a \) and \( \mathcal{W}_b \in \mathbb{R}^{(C + 4) \times 512} \) denote the weights of the FC layers. As two types of node permutations exist in each node subset, we further design a loss to penalize the gap between the predicted subset \( \hat{O}_n \) and the ground-truth subset \( O_n \). Inspired by the Wasserstein distance measurement [46], we introduce the following matching cost function to minimize the Earth Mover’s Distance between \( \hat{O}_n \) and \( O_n \):

\[
\mathcal{L}_{emd}(\hat{O}_n, O_n) = \min_{\pi} \sum_{k=1}^{2} \left[ \mathcal{L}_c(\hat{\sigma}_q^{(k)}, o_n^{(k)}) + \mathcal{L}_r(b_n^{(k)}, B_{\pi k}) \right],
\]

where \( \pi \) represents a certain node permutation, while \( o_n^{(k)} \) denotes the \( k \)-th item of \( \pi \). Accordingly, \( (\hat{\sigma}_q^{(k)}, b_n^{(k)}) \in O_n \) refers to the ground-truth object label and bounding box coordinates of the \( \pi_k \)-th node. \( \mathcal{L}_c(\cdot) \) and \( \mathcal{L}_r(\cdot) \) further represent the cross-entropy and \( l_1 \) regression loss functions, respectively. Intuitively, Eq. (10) first looks for the “best match” between elements in the predicted and ground-truth node subsets and then computes the final loss.

### C. Edge Subset Prediction

ESP jointly predicts the two edges in each edge subset. The prediction for the \( q \)-th edge subset can be denoted as follows:

\[
\hat{R}_q = \left\{ (\hat{r}_q^{(1)}, \hat{b}_q^{(1)}), (\hat{r}_q^{(2)}, \hat{b}_q^{(2)}) \right\},
\]

where \( \hat{r}_q^{(1)} \) and \( \hat{r}_q^{(2)} \in \mathbb{R}^C \) represent two relationship class predictions for the \( q \)-th edge subset, while \( C_r \) refers to the number of relationship categories (including the non-relationship). Moreover, \( \hat{b}_q^{(1)} \) and \( \hat{b}_q^{(2)} \in \mathbb{R}^4 \) denote the coordinates of the union bounding box for the two nodes connected by the edge. ESP module comprises two components: Subset-based Reasoning and Edge Matching Loss.

**Subset-based Reasoning.** Existing works [10], [11], [14], [25], [28] tend to merge multimodal features, e.g., the appearance feature, the vector of object class prediction scores, and the spatial features of two nodes, to represent their edges. More formally, the edge feature from the \( j \)-th to the \( k \)-th nodes can be represented as follows:

\[
f_{jk} = \mathcal{F}(v_{ijk}, v_{j,k}^{(1)}),
\]

where nodes \( i \) and \( j \) belong to the \( n \)-th node subset, while node \( k \) is from \( m \)-th node subset. \( \mathcal{F} \) denotes a fusion function defined in [42]:

\[
x \ast y = ReLU(W_x x + W_y y) - (W_x x - W_y y) \in \mathbb{R}^{512 \times 512}.
\]

Using the obtained edge feature, existing works typically make predictions for each edge independently. However, Eq. (12) may result in indistinguishable representations for the crowded node subset-related edges. This is because these edges may connect nodes with high IoU, and therefore very similar features.

To address the above problem, we propose a new fusion function to obtain the edge subset representation. In the below, we take the \( q \)-th edge subset, which is composed of \( R_{cs} \) edges that connect the \( n \)-th to \( m \)-th node subsets, as an example to describe the fusion function:

\[
f_q = \sigma(v_{ij}^{(1)} \ast v_{jk}^{(2)}) \ast (v_{ij}^{(1)} + B_{ij,k}),
\]

where \( \sigma \) is from Eq. (9).

**Edge Matching Loss.** After obtaining the edge subset representation by subset-based edge fusion, we adopt two FC layers to predict the two edges in each edge subset as follows:

\[
\begin{align*}
\hat{p}_q^{(1)}; \hat{p}_q^{(2)} = \mathcal{W}_c f_q, \\
\hat{r}_q^{(1)}; \hat{r}_q^{(2)} = \mathcal{W}_d f_q,
\end{align*}
\]

where \( \mathcal{W}_c \) and \( \mathcal{W}_d \in \mathbb{R}^{(C_r + 4) \times 512} \) denote two projection matrices. We utilize the following matching cost function to penalize the difference between the \( q \)-th predicted and the ground-truth edge subsets:

\[
\mathcal{L}_{emd}(\hat{R}_q, R_q) = \min_{\phi} \sum_{k=1}^{2} \left[ \mathcal{L}_c(\hat{\phi}_q^{(k)}, R_q^{(k)}) + \mathcal{L}_r(b_q^{(k)}, R_q^{(k)}) \right],
\]

where \( \phi \) denotes a specific edge permutation in the subset, while \( R_q^{(k)} \) denotes the \( k \)-th item of \( \phi \) and \( (\hat{p}_q^{(k)}, b_q^{(k)}) \in \mathbb{R}^4 \).
D. SGG BY SM-NET

During training, the overall loss function $\mathcal{L}$ of SM-Net can be written as follows:

$$
\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}_{emd}(\hat{O}_n, O_n) + \frac{1}{Q} \sum_{q=1}^{Q} \mathcal{L}_{emd}(\hat{R}_q, R_q).
$$

During testing, the object categories of nodes $\{\hat{o}_n^{(1)}, \hat{o}_n^{(2)}\}$ for the $n$-th node subset are predicted by:

$$
\hat{o}_n^{(1)} = \arg \max_{c_o \in C_o} (\hat{q}_n^{(1)}(c_o)), \quad \hat{o}_n^{(2)} = \arg \max_{c_o \in C_o} (\hat{q}_n^{(2)}(c_o)),
$$

where $C_o$ represents the set of object categories. The predicted bounding boxes $\hat{b}_n^{(1)}$ and $\hat{b}_n^{(2)}$ are utilized to assign the predicted object categories to a single node, according to the IoU value between the bounding box of one node and the predicted bounding box. The relationship categories of edges $\{\hat{r}_q^{(1)}, \hat{r}_q^{(2)}\}$ for the $q$-th edge subset are obtained as follows:

$$
\hat{r}_q^{(1)} = \arg \max_{c_r \in C_r} (\hat{p}_q^{(1)}(c_r)), \quad \hat{r}_q^{(2)} = \arg \max_{c_r \in C_r} (\hat{p}_q^{(2)}(c_r)),
$$

where $C_r$ represents the set of relationship categories. The predicted bounding boxes $\hat{b}_q^{(1)}$ and $\hat{b}_q^{(2)}$ are utilized to assign the corresponding edge category to one node pair that is relevant to the edge subset, according to the IoU value between the union box of the node pair and the predicted bounding box. Moreover, we enforce two predictions for all types of node and edge subsets to facilitate the form unity in Eq. (3) and Eq. (11). However, for the $O_s$ node subset, as well as $R_{ss}$ and $R_e$ edge subsets, there is essentially only one prediction required. In this case, we retain only the most confident prediction for these subsets.

V. EXPERIMENTS

To demonstrate the effectiveness of SM-Net, we conduct exhaustive experiments on three datasets: Visual Genome (VG) [6], Visual Relationship Detection (VRD) [7], and Open Images (OI) [8]. In this section, we report the evaluation settings and implementation details, after which we conduct ablation studies and comparisons with state-of-the-art methods.

A. EVALUATION SETTINGS

Visual Genome: For this dataset, we follow the same data cleaning strategy [9] that has been widely adopted in recent works: specifically, the most frequently occurring 150 object categories and 50 relationship categories are utilized for evaluation. The scene graph for each image consists of 11.6 objects and 6.2 relationships on average. The data is then split into one training set and one testing set: the training set contains 75,651 images, with 5,000 images used as a validation subset, while the testing set comprises 32,422 images. Following existing works [13], [18], [23], [27], [28], [31], performance evaluation is conducted under the following three standard settings: 1) Scene Graph Detection (SGDET). Given an image, the model detects object bounding boxes and predicts both the object category for each bounding box and the relationship category for each bounding box pair. 2) Scene Graph Classification (SGCLS). Beginning from the ground-truth location of object bounding boxes, the model predicts both object and relationship categories. 3) Predicate Classification (PREDCLS). Using the ground-truth object bounding boxes and their object categories, the model predicts only the relationship categories. All three settings are evaluated according to Recall@K ($R@K$) metric, where $K$ is set to 20, 50, and 100, respectively.

Visual Relationship Detection: This dataset contains 5,000 images with 100 object categories and 70 relationship categories. It further has around 30,000 relationship annotations, with an average of eight relationships per image. We use the same train/test data split protocol as in [7], i.e., 4,000 images for training and 1,000 images for testing. Moreover, following [7], we evaluate the Relationship Detection and Phrase Detection tasks. The Relationship Detection task requires both object detection and relationship prediction between each pair of object bounding boxes. The Phrase Detection task detects the subject-object union boxes and predicts one triplet for each bounding box. We subsequently rank all retained predictions and report the R@50 and R@100 performance, respectively.

Open Images: The complete training and validation sets of OI contain 53,953 and 3,234 images, respectively. The dataset includes 57 object categories and 10 relationship categories. To conduct evaluation on the OI dataset, we follow the same evaluation metrics as RelDN [25]: namely, Recall@50, the weighted mean AP of relationships (wmAP$_{rel}$), and the weighted mean AP of phrases (wmAP$_{phr}$). More specifically, to address the extreme predicate class imbalance issue in OI, RelDN [25] scales the impact of each predicate category on the performance with reference to their ratios in the validation set, which is referred to as the weighted mAP (wmAP). wmAP$_{rel}$ evaluates the AP of the predicted triplets where both the subject and object boxes have an IoU above 0.5 with their ground-truth, respectively. wmAP$_{phr}$ is similar except that the IoU threshold is applied to the union box of one subject-object pair rather than their respective boxes. The final score is given by score$_{wtd} = 0.2 \times R@50 + 0.4 \times$ wmAP$_{rel} + 0.4 \times$ wmAP$_{phr}$.
TABLE 1: Performance comparisons with state-of-the-art methods on the VG dataset. We compute the mean over all tasks on Descent (SGD) with momentum, with an initial learning rate of $10^{-3}$ and batch size of 6. The learning rate is reduced by multiplying it by 0.1 if the performance on the validation set does not increase for two successive epochs. The total number of training epochs is set to 20. For the SGDET task on the VG dataset, we adopt the same two settings as in [10]. First, we only predict relationships for proposal pairs where the two boxes have overlap. Second, the top 64 object proposals in each image are selected following per-class non-maximum suppression (NMS) with an IoU of 0.3. For all protocols in the VRD and OI databases, we select the top 100 object proposals in each image following per-class NMS with an IoU threshold of 0.5 and then predict the relationships for all obtained proposal pairs. Moreover, the ratio between pairs without any relationship and those with relationships during training on all three databases is set to 3:1.

C. COMPARISONS WITH STATE-OF-THE-ART METHODS

In the following, we compare the performance of SM-Net with state-of-the-art methods on the VG, VRD, and OI databases, respectively.

**Visual Genome**: Performance comparisons on the VG dataset are summarized in Table 1. It is clear from the results that SM-Net achieves state-of-the-art performance under all evaluation metrics. Using the same detector as [25], SM-Net outperforms one very recent model, i.e., Seq2Seq-RL [53], by 1.8% on average at R@50 and R@100 over the three protocols. Specifically, it outperforms SGGNLs by 0.5%, 3.2%, and 1.7% at R@100 on SGDET, SGCLS, and PREDCLS, respectively. When compared with the same ResNeXt-101-FPN backbone, the performance advantage secured by SM-Net is even more dramatic: for example, it outperforms the MOTIFS by 2.2% at R@100 on the SGCLS protocol and 35.5% on average at R@20, R@50, and R@100 over the three protocols.

**Table 2**: Performance comparisons on mean recall (%) across all 50 relationship categories in the VG dataset.
TABLE 3: Performance comparisons with state-of-the-art methods on the VRD dataset.

| Model             | Relation Detection R@50 | Relation Detection R@100 | Phrase Detection R@50 | Phrase Detection R@100 |
|-------------------|-------------------------|--------------------------|-----------------------|------------------------|
| VRD-Full [7]      | 15.2                    | 17.0                     | 13.9                  | 14.7                   |
| VTransE [34]      | 19.4                    | 22.4                     | 14.1                  | 15.2                   |
| VRL [36]          | 18.1                    | 20.7                     | 21.3                  | 22.6                   |
| KL-distillation [39] | 19.2                | 21.3                     | 23.1                  | 24.0                   |
| Zoom-Net [38]     | 18.9                    | 21.4                     | 24.8                  | 28.1                   |
| CAL + SCA-M [38]  | 19.5                    | 22.4                     | 25.2                  | 28.9                   |
| GPS-Net [14]      | 21.5                    | 24.3                     | 28.9                  | 34.0                   |
| ReIDN [25]        | 24.0                    | 27.5                     | 30.1                  | 35.2                   |
| MF-URLN [40]      | 23.9                    | 26.8                     | 31.5                  | 36.1                   |
| MNCN [43]         | 24.5                    | 28.0                     | 31.8                  | 37.1                   |
| HeH [61]          | 22.4                    | 24.8                     | 30.6                  | 35.5                   |
| Seq2Seq-RL [53]   | 26.1                    | 30.2                     | 33.4                  | 39.1                   |
| SM-Net            | 27.5                    | 31.6                     | 34.1                  | 39.7                   |

TABLE 4: Comparisons with state-of-the-art methods on OI. We adopt the same evaluation metric as in [25].

| Dataset | Model | R@50 | wmAP rel phr | score wsd |
|---------|-------|------|--------------|-----------|
| V4      | ReIDN [25] | 74.9 | 35.5 | 38.5 | 44.6 |
|         | BGN [60]   | 75.5 | 37.8 | 41.7 | 46.9 |
|         | SM-Net     | 78.6 | 39.2 | 42.1 | 48.0 |
| V6      | ReIDN [25] | 73.1 | 32.2 | 33.4 | 40.8 |
|         | VCTREE [11] | 74.1 | 34.2 | 33.1 | 40.2 |
|         | G-RCNN [13] | 74.5 | 33.2 | 34.2 | 41.8 |
|         | MOTIFS [10] | 71.6 | 29.9 | 31.6 | 38.9 |
|         | GPS-Net [14] | 74.8 | 32.9 | 34.0 | 41.7 |
|         | BGN [60]   | 75.0 | 33.5 | 34.2 | 42.1 |
|         | SM-Net     | 77.1 | 35.7 | 34.7 | 43.8 |

TABLE 5: Ablation studies of the proposed modules. We use the same object detection backbone as in [42].

D. ABLATION STUDY

In the following, we systematically investigate the effectiveness of each key component in SM-Net. Experiments are conducted on the most popular VG dataset. Results are summarized in Table 5. Besides, the qualitative comparisons between SM-Net and one very recent model, named SGGNLS [52] are shown in Figure 4.

To justify the advantages of set prediction, we adopt two baselines in Table 5: a “BI” baseline that predicts each element in the scene graph independently and a “BS” baseline that jointly predicts elements in each subset. Specifically, “BI” and “BS” baselines utilize the concatenation fusion function to obtain the node/edge feature and node/edge subset feature, respectively. Furthermore, “BI” baseline utilizes the cross-entropy loss to obtain the node/edge prediction. “BS” baseline applies the loss defined in Eq. (16) to obtain the node/edge subset prediction. From the results, “BS” baseline outperforms “BI” baseline in both SGCLS and PREDCLS tasks, demonstrating the set prediction’s effectiveness. Note that Exp. 1-3 are all based on the “BS” baseline. Effectiveness of the Proposed Modules. We first perform an ablation study to justify the effectiveness of NSF and ESP modules. The results are summarized in Table 5. Exps 1-3 show that each module helps to promote the performance of SGG. The best performance is achieved when both mod-

FIGURE 4: Qualitative comparisons between SM-Net and SGGNLS [52]. Specifically, we show the comparisons at R@100 in the SGCLS setting in the first row. In the second row, we show the comparisons at R@100 in the PREDCLS setting. The green color indicates correctly classified objects or predicates; the red indicates those that have been misclassified. Best viewed in color.
E. FURTHER ANALYSIS AND DISCUSSIONS

We conduct additional experiments to further verify the design choices for NSP and ESP modules. Results are summarized in Table 6 and Table 7, respectively.

Comparisons between Three MP Modules. In Table 6(a), we compare the performance of SMP with another two MP modules. The first one is implemented according to Eq. (4) and Eq. (5); we refer to it as Global Context MP (GCMP) [13], [30]. The second one is the Direction-aware MP (DMP) module [14]. These three modules utilize the node subset feature as input and produce refined node subset features with reference to the contextual information. As shown in Table 6(a), SMP outperforms GCMP and DMP by 0.5% and 0.2% at R@100 on the SGCLS task, respectively. We give this mainly credit to the SMP that formulates the contextual information as a vector, storing more complex correlation information than the simple scalar weights used in GCMP and DMP.

Comparisons between Three pooling Strategies. In Table 6(b), we test the performance of SM-Net with different pooling strategies for the SMP. “SUM”, “Aver.”, and “Max” represent the sum pooling, average pooling, and max pooling, respectively. As shown in Table 6, SMP achieves the best performance when applying max pooling to obtain the contextual information. This maybe because max pooling operation could reduce the noise that can be induced by average or sum poolings, which oblige all the subsets to interact with each other.

Evaluation on the number of SMP layers. In Table 6(c), we show the performance of SM-Net with different numbers of SMP layers, ranging from two to five. The model performance improves consistently as the number of SMP layers increases. However, due to limitations on GPU memory size, we only conduct experiments up to five SMP layers.

Design Choices for ESP Module. In Table 7, we compare the performance of different fusion functions utilized in the ESP module. “Concat.” denotes the concatenation fusion function utilized in “BS” baseline. “MFB” represents the multi-modal factorized bilinear fusion function applied in [62]. “SUM” and “GATE” denote the two fusion functions proposed in [42]. For comparison fairness, these fusion functions are all utilized to obtain the subset-level edge representation. As illustrated in Table 7, our SR fusion function achieves the best performance. This may be because the feature selection-based fusion function suppresses the redundant features in the union area.

VI. CONCLUSION

In this work, we study the occlusion problem in SGG and propose a novel framework named SM-Net. Specifically, we decompose SGG into node subset matching and edge subset matching problems. Unlike existing works which classify each element in scene graph independently, SM-Net jointly predicts the categories of the nodes and edges within the same node and edge subsets, respectively. With the subset matching strategy, the correlation between nearby nodes and the correlation between their related edges are both considered; therefore, SM-Net can alleviate the occlusion problem and is more robust for generating scene graphs in complex scenes. Extensive experimental results on three popular SGG datasets justify the effectiveness of SM-Net.
[50] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740-755.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248-255.

[52] Y. Zhong, J. Shi, J. Yang, C. Xu, and Y. Li, “Learning to generate scene graph from natural language supervision,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 1823-1834.

[53] Y. Lu, H. Rai, J. Chang, B. Knayazev, G. Yu, S. Shekhar, G. Taylor, and M. Volkovs, “Context-aware scene graph generation with seq2seq transformers,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 15931-15941.

[54] X. Chen, W. Chen, T. Chen, Y. Yuan, Chen Gong, K. Chen, Z. Wang, “SelfPU: Self Boosted and Calibrated Positive-Unlabeled Training” in Proc. Int. Conf. Mach. Learn., 2020, pp. 1510-1519.

[55] D. Arpit et al., “Closer look at memorization in deep networks,” Proc. Int. Conf. Mach. Learn., 2017, pp. 233-242.

[56] H. Ben-younes, R. Cadene, N. Thome, M. Cord, “BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 8102-8109.

[57] Z. Yu, J. Yu, J. Fan, D. Tao, “Multi-modal factorized bilinear pooling with co-attention learning for visual question answering,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 1821–1830.

[58] G. Yang, J. Zhang, Y. Zhang, B. Wu, and Y. Yang, “Probabilistic modeling of semantic ambiguity for scene graph generation,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 12527–12536.

[59] Y. Guo, L. Gao, X. Wang, Y. Hu, X. Xu, X. Lu, H. Shen, and J. Song, “From general to specific: Informativene scene graph generation via balance adjustment,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 16383-16392.

[60] R. Li, S. Zhang, B. Wan, and X. He, “Bipartite graph network with adaptive message passing for unbiased scene graph generation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 11109-11119

[61] W. Wang, R. Wang, S. Shan, and X. Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. in Proc. Eur. Conf. Comput. Vis., 2020, pp. 222-239.

[62] R. Cadene, H. Ben-Younes, M. Cord, and N. Thome, “Murel: Multimodal relational reasoning for visual question answering,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1989-1998.

**XIN LIN** received the master degree in information and communication engineering from Shenzhen University, Shenzhen, China, in 2018. He currently pursues the Ph.D. degree in the College of Electronics and Information Engineering, South China University of Technology, GuangZhou, China. His research interests include scene graph generation and computer vision.

**JINQUAN ZENG** received the bachelor’s degree from the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, in 2019, where he is currently pursuing the master’s degree in signal and information processing. His current research interests include scene graph generation and computer vision.

**XINGQUAN LI** received his Ph.D. degree in College of Electronics and Information Engineering from the Shenzhen University, Shenzhen, China, in 2019. He was a postdoctoral researcher with the College of Electronics and Information Engineering, Shenzhen University. He joined communication group of the Queen Mary University of London, UK, in 2019, as a visiting scholar, where he researched the green communications and resource allocation. He is currently a lecturer with the School of Microelectronics at Shenzhen Institute of Information Technology. His research interests include cooperative communications, green communications and resource allocation, scene graph generation.

---

**Note:** The above text is a sample of the content that could be extracted from the image. It includes references and brief biographies of some authors. The text is formatted to match the style of the IEEE Access journal. The image content is not the focus and has been disregarded in the conversion to plain text.