Multi-View Predicate Recognition for Solving Semantic Ambiguity Problem in Scene Graph Generation

Xuezhi Tong
xuezhitong899@gmail.com
College of Intelligence and Computing, Tianjin University
Tianjin, China

Lihua Jing*
jinglihua@iie.ac.cn
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China

Cong Zou
zoucong@iie.ac.cn
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China

Rui Wang
wangrui@iie.ac.cn
State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China

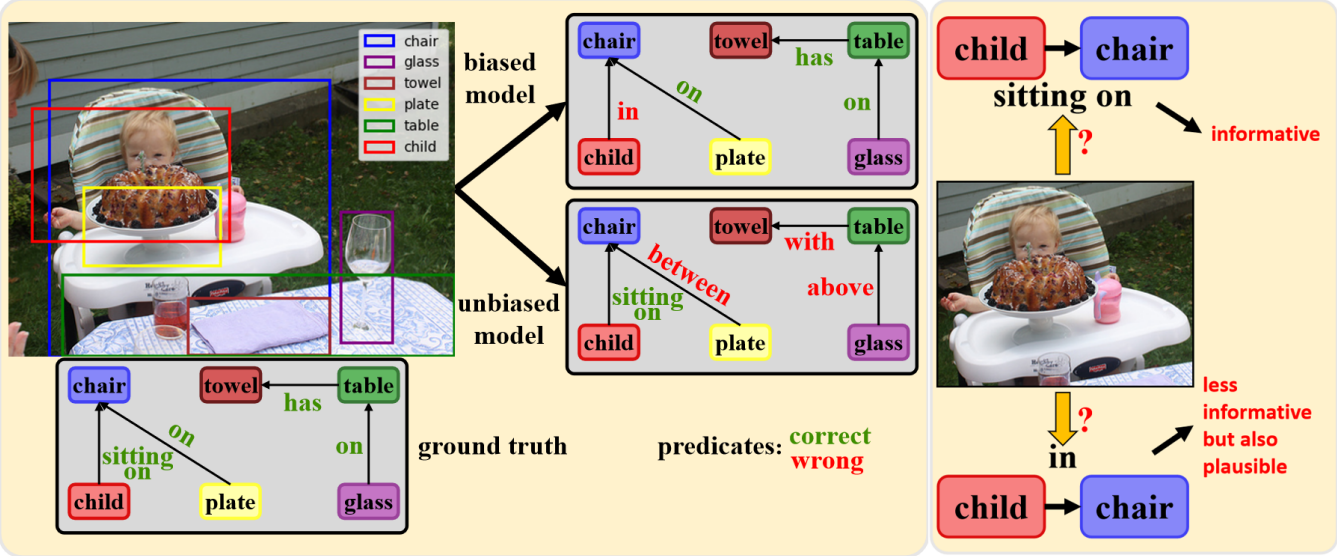

Figure 1: The left figure is the illustration of the trade-off between head and tail predicates that current unbiased SGG methods suffer from. This is an image selected from the Visual Genome dataset [14], and the biased and unbiased results are SSR-CNN and SSR-CNN with Logit Adjustment. While the ground-truth consists of head and tail predicates (e. g. *on* from head predicates and *sitting on* from the tail ones). The right figure represents the semantic ambiguity problem. For a given object pair like *child* and *chair*, there may exist multiple plausible predicates to describe their relationship.

ABSTRACT

Recent works on Scene Graph Generation (SGG) have been concentrating on solving the problem of long-tailed distribution. While these methods are making significant improvements on the tail predicate categories, they sacrifice the performance of the head ones severely. The major issue lies in the semantic ambiguity problem, which is the contradiction between the commonly used criterion and the nature of relationships in the SGG datasets. The models are evaluated with graph constraint, which allows merely one relationship between a pair of objects. However, the relationships are much more complex and can always be described from different views. For example, when a *man* is *in front of* a *computer*, we can also say he is *watching* it. Both options are plausible, describing the different aspects of the relationship. Which of them is determined to be the ground-truth is highly subjective. In this paper, we claim that the relationships should be considered from multiple views to avoid the semantic ambiguity. In other words, the model should provide all the possibilities, rather than being biased to any

*Corresponding Author.

McGE '23, October 29, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0278-5/23/10.
https://doi.org/10.1145/3607541.3616817

one of the options. To this end, we propose the **Multi-View Predicate Recognition (MVPR)**, which separates the label set into multiple views and enables the model to represent and predict in a "multi-view" style. Specifically, MVPR can be divided into three parts: **Adaptive Bounding Box for Predicate** is proposed to help the model attend to the crucial areas for the predicate categories in different views; **Multi-View Predicate Feature Learning** is designed to separate the feature space of different views of predicate categories; **Multi-View Predicate Prediction** and **Multi-View Graph Constraint** are used to allow the model to provide multi-view predictions to accurately estimate ambiguous relationships. Experimental results on the Visual Genome dataset show that our MVPR can significantly improve the model performance on the SGG task, and achieves a new state-of-the-art.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**.

KEYWORDS

Scene graph generation, long-tailed distribution, multi-view, semantic ambiguity

ACM Reference Format:

Xuezhi Tong, Lihua Jing, Cong Zou, and Rui Wang. 2023. Multi-View Predicate Recognition for Solving Semantic Ambiguity Problem in Scene Graph Generation. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3607541.3616817

1 INTRODUCTION

Scene graph generation (SGG), aiming to recognize the objects and their relationships in images, is of great importance for high-level visual scene understanding. In this task, multiple relationship triplets are produced for each image and connected to form a scene graph, where objects are regarded as nodes and predicates are treated as edges in the graph. As the graphical representation of scenes, a scene graph not only presents spatial (localization) and semantic (recognition) information of objects but also consists of an interactive formulation describing the scene.

The main framework of recent works is mostly based on object detection methods and can be divided into two groups: one-stage and two-stage SGG methods. For two-stage SGG methods, a pre-trained object detector is used to generate object candidates and extract the corresponding visual features. These features are refined by a message passing network, and combined in pairs to represent the predicates describing the relationships between objects. The resulting object and predicate features are then fed into classifiers to predict the corresponding object/predicate categories for constructing the scene graph prediction. For one-stage SGG methods, the prediction of the relationship triplets (subject, predicate, and object) is regarded as a set prediction problem and tackled via expanding anchor-free object detection frameworks.

However, both of the two groups of methods show a severe training bias towards naive predicate classes like preferring *on* rather than *sitting on*, making the generated scene graph less informative.

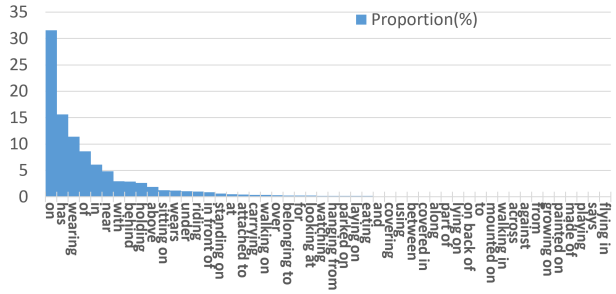

Figure 2: The long-tailed predicate distribution in the VG dataset [14].

Most de-biasing works owe such training bias to the long-tailed distribution of predicate categories, which is illustrated in Fig. 2. They either re-sample the image data to provide more training instances for tail classes [1, 7] or re-weight the predicate classes to emphasize tail classes, based on the frequency of predicate classes[9, 19]. However, while these methods have achieved state-of-the-art performance on mean Recall@K (mR@K), they have also severely degraded the model performance on Recall@K (R@K), which is dominated by the head predicates. That means as a cost of correcting a relatively small number of tail predicates, a large number of head predicates have been recognized wrongly, this trade-off dilemma is shown in Fig. 1.

In fact, the trade-off of head predicates and tail predicates is due to the diversity of predicate labels in the real-world. As shown in Fig. 1 for a certain object pair, there may exist multiple options for the annotators. As the existing criterion of the SGG task asks the model to predict a single relationship for each pair of objects, instances that are similar visually may have different labels. To conclude, there exists a contradiction between the criterion and the nature of relationships in the SGG datasets. We call it the semantic ambiguity problem for SGG. Although some recent works [2, 12, 22] have paid some attention to the problem, current SGG methods still severely suffer from it.

After investigating the label set of the SGG datasets, we find that the categories of the predicate are actually not mutual-exclusive. Moreover, there exist different views[1] of the categories. We divide them into two major views: spatial and semantic predicates. The spatial ones describe simpler relationships that can be inferred by only considering the relative position of two objects, like *on*, *in front of*, and *above*. The semantic ones describe the relationships that involve less spatial information and more high-level semantic information, like *belonging to*, *wearing*, and *playing*. Either for SGG models or the annotators, it is not reasonable to claim that a pair of objects can only have spatial or semantic relationships.

To this end, we propose ***Multi-View Predicate Recognition (MVPR)*** to grant SGG model the ability to recognize the relationships in different views. MVPR concentrates on refining predicate features by adaptively re-positioning the bounding boxes for the

[1]In this paper, we regard predicates that describe different aspects of the relationship between a pair of object and can exist simultaneously, such as *near* and *watching*, to be in different views.

predicates and learning separate representations in spatial and semantic views, respectively. More specifically, rather than using the fixed union box of the subject and object, MVPR learns to regress the bounding box for the predicate, enabling the model to attend to different visual information for different predicate categories. For multi-view feature learning, the model uses two separate branches for spatial and semantic predicates, respectively. The two branches consider the background category separately, allowing for the situation that only one of the branches has a reasonable prediction result to describe the relationship. Moreover, we propose a novel scene graph prediction constraint, dubbed Multi-View Graph Constraint (MVGC), which softens the common graph constraint that a pair of objects can only have one relationship at the same time. This new constraint allows the model to produce one predicate classification result for each view and is a better choice compared to both using the graph constraint or not using it. While the graph constraint can help construct a clear scene graph and prevent the model from guessing the ground-truth by listing all possibilities, our MVGC avoids the concern of guessing the categories by adding an inductive bias. Our MVGC also fits the data in SGG better than the graph constraint, because it is always the case that two objects have multiple relationships from different views. Although allowing multiple predictions for an object pair, applying our MVGC is different from omitting the graph constraint. For example, our Multi-View Recall (MV-R) is different from Recall without graph constraint (ng-R), because the possible number of relationships between two objects is limited to be no more than K (let K be the number of views) compared to ng-R. More importantly, we give different predicate classification results by adding an inductive bias to the model, rather than listing all possibilities of one classifier. Practically, we use a certain number of views for MVGC, dubbed K-VGC (K equals 2 in this paper).

To conclude, the contribution of MVPR is three-fold:
(1) MVPR adaptively attends to different crucial areas for different predicate categories, thus providing more accurate representations for the classifiers;
(2) MVPR proposes to solve the semantic ambiguity problem by separately approximating the feature spaces for K views of predicate categories, defined as the spatial and semantic views in this paper so that the model can avoid getting confused when being trained with object pairs that are similar in visual but have different predicate labels;
(3) MVPR softens the contradiction between the commonly used criterion and the nature of relationships in the SGG datasets by allowing for multiple predicate predictions for a single object pair, which is no more than the number of views.

2 RELATED WORK

Regarding the relatedness to this paper and the target problems, we divide existing works of SGG into three groups: Contextual Embedded, Unbiased, and Disambiguated methods.

Contextual Embedded Scene Graph Generation Method. As an extension of object detection, the SGG task differs from it in that the SGG task explores how objects are correlated with each other. Therefore, it is natural for early works to research the way of exploiting contextual information, so as to refine the visual features extracted from the object detector. As a pioneer work, [4] propose to use a message passing network to iteratively refine the feature of objects and predicates, based on Gated Recurrent Units (GRU). As the efficacy of contextual embedding has been proved, Motifs [15] explore using Bi-LSTM as the structure of message passing network, and taking better advantage of the contextual information. They have also proposed FREQ [15], revealing the fact that simply choosing the most frequent predicate for a given pair of objects can result in state-of-the-art performance. However, both randomly connecting the input nodes in a sequence or fully connecting them as a graph are not ideal options. It is obvious that they introduce noise to the contextual information by connecting unrelated nodes. To this end, Graph R-CNN [6] and VCTREE [9] propose to add a separate MLP to estimate the connectivity between two objects according to their visual features, but this estimation is far from perfect. In addition, due to the long-tailed distribution of predicate categories, the predictions are severely biased to the frequent predicates, which are relatively less informative in semantics.

Unbiased Scene Graph Generation Method. Inspired by the observation in Motifs [15], the following works use the frequency of predicate as a cue to emphasize the less frequent (tail) predicate categories against the frequent (head) ones. TDE [10] subtracts a biased prediction with the estimated bias, which is learned by the biased model and dominated by the frequent categories. BGNN [7] resamples the data from both image-level and instance-level, making the model trained with more data containing the tail predicates. SQUAT [8] further optimizes the message-passing network of BGNN [7] and achieves better performance. PPDL[18] re-weights the training loss according to the similarity SSR-CNN introduces Logit Adjustment (LA) into SGG, and finds that simply post-processing the predicted logits can also improve the performance on tail predicate categories significantly. Similarly, RTPB also proposes to add the frequency information to the predicted logits. More recently, GCL [5] divides the label set into 5 subsets that are relatively balanced, and makes 5 classifiers collaboratively learn on the dataset based on the idea of class-incremental learning. However, they are still limited to the long-tailed distribution of the predicate, which is only a phenomenon of the core problem. As a result, the so-called unbiased methods sacrifice the performance on Recall@K to improve performance on mean Recall@K, and produce predictions that are biased to the tail predicates in turn.

Disambiguated Scene Graph Generation Method. The origin of the issue lies in the contradiction between the criterion and the nature of the relationship in the SGG datasets. In recent years, attention has arisen to solving this problem. [2] impute the ground-truth labels with the predictions of a biased model, so that one object pair can have multiple predicate labels. NICE [12] clean the dataset by automatically detecting noisy samples and correcting them, so as to give more informative predicate labels to all the object pairs. PSG [22] reconstructs a dataset and carefully chose 56 categories that are as independent as possible, to avoid the semantic ambiguity problem. All these methods try to change the labels to conduct disambiguating, overlooking the nature of SGG data that one pair of objects can have multiple types of relationships simultaneously.

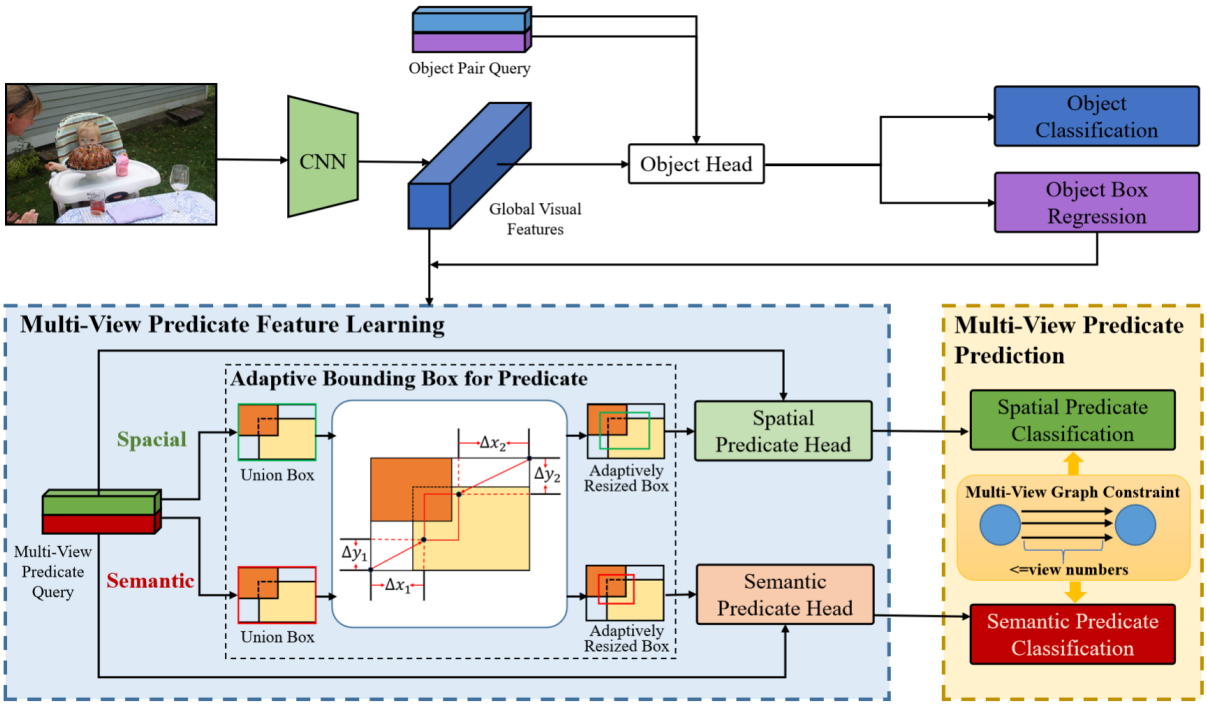

Figure 3: The main process of our MVPR. A triplet query is divided into four parts, two of which are learned for object localization and classification. The other two parts are used for the spatial and semantic views. The bounding boxes of the object pairs are regressed and used to find the union box. Each branch of the Multi-View Predicate Feature Learning module exploits the input predicate features to resize that union box and get the Adaptive Bounding Box for Predicate. Finally, Multi-View Predicate Prediction is adopted to produce results that are robust to the semantic ambiguity problem.

3 METHOD

Overview. The framework of our Multi-View Predicate Recognition is illustrated in Fig. 3. We use Structured Sparse R-CNN (SSR-CNN) [21] as the base message passing network, and extend it to have the multi-view inductive bias and bounding box regression for the predicates. The framework first extracts the feature map of the whole image and then decodes it using a series of triplet queries. Each triplet query represents a relationship triplet, which consists of a subject, object, and predicate. This query is then divided to be three major parts of features: the bounding box for the object pair, visual features for the object pair, and the features for the predicate. Among them, the predicate features can be further divided into the features for the spatial and semantic predicates. All these features are then processed by the following message passing network, and corresponding to each part of the outputs: bounding box regression results for objects, object classification results, and the predicate predictions for spatial and semantic categories. We describe the network architecture in detail in this section.

3.1 Problem Formulation

A *scene graph* $G = (U, E)$, is a graphical representation of the visual contents in a scene. The objects in the scene are represented by the node set $U = (B, O)$ of the scene graph, where B and O are the set of bounding box and label for the objects, respectively. The edge set E consists of predicates describing the relationships of connected objects. Object i, j, and the predicate between them forms a relationship triplet $e_{ij} = (o_i, r_{ij}, o_j)$, where $r_{ij} \in \mathcal{R}$ means the class of predicate.

3.2 Structured Sparse R-CNN

Before describing our Multi-View Predicate Recognition, we briefly recap the content of SSR-CNN to make our contribution clear, note we omit Pair Fusion, which is a less related detail in SSR-CNN. In SSR-CNN, the basic building block is a Sparse R-CNN [16]. It consists of a dynamic convolution layer, a feed-forward network (FFN), a classifier, and a regression head. The bounding box part of the triplet query is used to extract the ROI-Align features for the objects from the whole image feature map. Before fed into the Sparse R-CNN, a multi-head attention layer is used to embed contextual information to the object query features. The dynamic convolution layer uses the query features as convolution kernels to refine the object features extracted via ROI-Align. Finally, the refined object features are fed into the classifier and the regression head to get the object detection results. Similarly, the predicate recognition is also conducted via a Sparse R-CNN. While the object part uses object bounding boxes to extract ROI-Align features, the predicate part uses the union box of the object pair. Noted that after the dynamic convolution layer of the predicate head, a bottom-up connection is used to combine the object-level features with the predicate feature vectors, dubbed entities to relation fusion (E2R).

This operation is formulated as:

$$\boldsymbol{Q} = \boldsymbol{W}_x ReLU(LN(\boldsymbol{W}_r^s \boldsymbol{V}_s)) + \boldsymbol{W}_y ReLU(LN(\boldsymbol{W}_r^o \boldsymbol{V}_o)), \quad (1)$$

$$\boldsymbol{P'} = LN(\boldsymbol{P} + \boldsymbol{Q} + \boldsymbol{W}_r^p ReLU(\boldsymbol{W}_p^s Posi(s) + \boldsymbol{W}_p^o Posi(o))), \quad (2)$$

where $\boldsymbol{V}$ and $\boldsymbol{P}$ are the object and predicate features, respectively. LN represents the Layer Normalization function [3], and ReLu is the activation function. While $\boldsymbol{Q}$ is an intermediate variable representing the information from the object pairs, Posi is the position embedding function. $bmW_x, \boldsymbol{W}_y, \boldsymbol{W}_r^s, \boldsymbol{W}_r^o, \boldsymbol{W}_r^p, \boldsymbol{W}_p^s$ and $\boldsymbol{W}_p^o$ are the weight matrices of the linear layers.

Finally, the object and predicate feature refining process is iterated for M times, and the model takes the output of the current time step as the input for the next time step.

3.3 Multi-View Predicate Recognition (MVPR)

However, the SSR-CNN overlooks the semantic ambiguity problem and struggles at balancing between the head and tail predicate. Although experiments have been conducted to find the best hyper-parameter for the Logit Adjustment (LA), R@K still decreases significantly as a cost for improving mR@K [21]. To alleviate this problem, we propose Multi-View Predicate Recognition (MVPR). We describe the details of MVPR by the order of Adaptive Bounding Box for Predicate, Multi-View Predicate Feature Learning, and Multi-view Predicate Prediction.

Adaptive Bounding Box for Predicate. The bounding box for a predicate has long been defined as the union box of the two related objects [15]. This is a nature extending of the bounding box for object and includes adequate information for predicate recognition. However, as the data in SGG show severe intra-class variance and inter-class similarity [25], complete visual information may not be a good choice. As shown in Fig. 3, different categories of predicates may have different crucial areas, while the rest areas introduce severe noises. To this end, we propose an Adaptive Bounding Box for Predicate (ABBP) module to resize the union box of the two objects, so that for different predicate categories, the model attends to different areas. We adopt the network of the bounding box regression head for object and use the predicate features in the query as the input. Because these features are optimized for predicate classification and are not in the same space with object features, they fit well to the regression task. Specifically, we resize the union box u_{ij} for a relationship r_{ij} by moving the left-bottom point and the right-top points of u_{ij} horizontally and vertically inside u_{ij}. The network predicts the four resizing factors ranged in $(0, 1)$, which describe percentages for the four sides of u_{ij} that the two points go along the x-axis and y-axis. We get the resizing factors from the ABBP module as follows:

$$(\triangle x_1, \triangle y_1, \triangle x_2, \triangle y_2) = \sigma(f(L(\boldsymbol{p_{ij}}))), \quad (3)$$

where $\triangle$ represents the resizing factor, σ is a sigmoid function for restricting the range of the regressor. f is a linear layer for adjusting the dimension of features to 4, and L is a stacked network consisting of linear layers, activation functions, and the Layer Normalization functions [3]. $\boldsymbol{p_{ij}}$ represents the predicate features for object i and j. According to the experimental results, the ROI-Align features for the predicate is underweight because of the object information introduced by the E2R module. To this end, we add am empirical weight to the ROI-Align predicate features, which change the formulation of the E2R module to be:

$$\boldsymbol{P'} = LN(\beta \boldsymbol{P} + \boldsymbol{Q} + \boldsymbol{W}_r^p ReLU(\boldsymbol{W}_p^s Posi(s) + \boldsymbol{W}_p^o Posi(o))), \quad (4)$$

where β is the weight factor used to enhance the ROI-Align predicate features.

Multi-View Predicate Feature Learning. Considering the different aspects of information involved in spatial and semantic predicates, we propose to separate the features of these two views of the predicate. Practically, we use two branches of predicate heads, both learn a 256-d predicate query separately. The network architectures are kept the same for these two branches, while the parameters are not shared. The predicate features, whose dimensions are kept to 256 during the training process, are fed into a multi-attention layer to introduce contextual information. The output features are then refined by a dynamic convolution layer, with the resized bounding box from the aforementioned bounding box regression module for the predicate. The predicate features for the next time step are then produced by the following feed-forward network. Finally, the predicate predictions at this time step are made by the Multi-View Predicate Prediction module. In order to guarantee the separation between the feature space of spatial and semantic predicates, we use two focal losses to restrict the two branches:

$$\mathcal{L}_{sp}(t) = -H_t(\alpha)(1 - \hat{Pr}_t^{sp})^{\mu} \log(H_t(\sigma(\hat{Pr}_t^{sp}))), \quad (5)$$

$$\mathcal{L}_{se}(t) = -H_t(\alpha)(1 - \hat{Pr}_t^{se})^{\mu} \log(H_t(\sigma(\hat{Pr}_t^{se} t))), \quad (6)$$

where $\mathcal{L}_{sp}$ and $\mathcal{L}_{se}$ are the losses for the spatial predicates and semantic predicates, respectively. t is the current category index ranging from 1 to the number of predicate categories in the dataset. Pr means the output logits from the classifiers, and σ is a sigmoid function. α and μ are the hyper-parameters used to control the weight between the predicate categories and the background category. Function $H_t(x)$ is defined as:

$$H_t(x) = \begin{cases} x & y_t = 1 \\ 1 - x & \text{otherwise} \end{cases}, \quad (7)$$

where $y_t = 1$ means the current category is the ground-truth.

Multi-View Predicate Prediction. In this paper, we use the popular VG-150 dataset [15], which is a subset of the Visual Genome [14] and consists of 50 predicate categories and the background category. We select 13 of them as the spatial categories, and the rest 37 are assigned to the semantic view. For both branches we add the background category, therefore allowing any of them to predict that a pair of objects have no relationship in the corresponding view. As for the detailed division, please refer to Sec. 4.2. After refining the features for the predicates with the network for multi-view predicate feature learning, we classify them with two MLP classifiers. Between them, the spatial predicate classifier is responsible for 13 spatial predicate categories plus the background category, while the semantic one corresponds to the rest of the predicate categories plus the background category. We then reorder the logits of the two branches to keep the prediction consistent on the category ordering with the ground-truth. We also keep the category indexing of the two views to calculate the final multi-view predicate predictions, and the model performance with our multi-view graph constraint

Method	SGDet					
	Recall@20	Recall@50	Recall@100	mRecall@100	mRecall@50	mRecall@100
IMP [4]	18.1	25.9	31.2	2.8	4.2	5.3
Graph R-CNN [6]	-	29.7	32.8	-	5.8	6.6
VTransE [23]	24.5	31.3	35.5	5.1	6.8	8.0
RelDN [24]	-	31.4	35.9	-	6.0	7.3
GPS-Net [19]	-	31.1	35.9	-	7.0	8.6
MOTIFS [15]	25.1	32.1	36.9	4.1	5.5	6.8
VCTREE [9]	24.5	31.9	36.2	5.4	7.4	8.7
Transformer [17]	25.6	33.0	37.4	6.0	8.1	9.6
BGNN [7]	-	31.0	35.8	-	10.7	12.6
PE-Net [25]	-	30.70	35.2	-	12.4	14.5
SGTR	-	24.6	28.4	-	12.0	<u>15.2</u>
SSR-CNN [21]	26.1	33.5	38.4	6.2	8.6	10.3
SSR-CNN$^{\diamond}$ [21]	**30.7**	<u>37.8</u>	<u>41.3</u>	<u>10.4</u>	<u>13.4</u>	15.0
MVPR	25.8	33.4	38.2	6.4	9.0	10.9
MVPR*	<u>29.9</u>	**38.6**	**44.1**	**10.9**	**15.1**	**17.9**

Table 1: SGDet performance of on VG dataset [14] compared with state-of-the-art methods. All the methods are divided into two-stage and one-stage SGG methods via a horizontal line. $*$ means the methods that are evaluated with the Multi-View Graph Constraint. For a fair comparison, we also design an approximation of our MVGC for single view methods and compare MVPR with the baseline method in Sec. 4.4, which is noted as $\diamond$. Methods are divided into two-stage and one-stage ones. Best results are marked in bold and suboptimal ones are underlined.

Method	R@50	R@100	mR@50	mR@100
SSR-CNN w/o gc [21]	36.8	43.6	16.7	22.4
SSR-CNN$^{\diamond}$ [21]	37.8	41.3	14.7	17.7
MVPR w/o gc	36.4	43.2	16.2	22.2
MVPR*	38.6	44.1	15.1	17.9

Table 2: Comparison of different criteria.

(MVGC). Therefore, the inference process of our MVPR and the calculation of MV-R@K can be formulated as:

$$\widetilde{r}_{sp} = \delta_{sp}(argmax(\boldsymbol{Pr_{sp}})), \tag{8}$$

$$\widetilde{r}_{se} = \delta_{se}(argmax(\boldsymbol{Pr_{se}})), \tag{9}$$

$$\boldsymbol{c_{ij}} = TopK(\boldsymbol{s_{ij}}) \tag{10}$$

$$MVR@K = Recall([S^{sp}_{\boldsymbol{c_{ij}}}, S^{se}_{\boldsymbol{c_{ij}}}], G_t), \tag{11}$$

where $\widetilde{r}$ is the model prediction of the predicate, $TopK$ is a ranking function that maps a vector to the index list of its topk values. δ is the permutation function that maps the category inside the spatial and semantic to the index inside the whole label set. $\boldsymbol{S}$ is the triplet confidence vector, which calculating by multiplying the confidences of the subject, object, and predicate. While $\boldsymbol{c_{ij}}$ is the top K candidate triplets, MV-R is evaluated by concatenating the spatial and semantic predicate predictions of the candidates and calculating the Recall of the re-ranked triplets based on the ground-truth G_t.

4 EXPERIMENTS

4.1 Dataset and Settings

Visual Genome (VG) dataset [14] is the most popular dataset for SGG. We follow [15] to get a subset of the VG dataset (VG150), which has 150 object categories and 50 predicate categories. Limited by the structure of the one-stage model [21], the proposed method is only evaluated on SGDet. Following [15], 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/ Scene Graph Detection** (SGGen/SGDet): predict the bounding boxes and labels of the objects, and classify the relationships for the object pairs.

However, as predictions of object localization, and object classification is conducted based on the triplet queries in one-stage SGG methods, it is not suitable to simply replace the prediction results without adapting the object features. Therefore, we only evaluate our MVPR on the most challenging task, *i.e.*, SGDet.

Evaluation. The bounding boxes of the subject and object of a relationship triplet should have more than 50% IoU with the ground truth boxes. A prediction for the relationship triplet is regarded as a correct one only when the corresponding subject, object, and the predicate are classified correctly at the same time. All the relationship prediction candidates are ordered according to the prediction confidences, which is calculated by multiplying the classification confidence of the objects and predicates. Recall@K

counts the number of relationship triplet predictions that hit the ground-truths and are ranked as the top K ones. Mean Recall@K calculates the Recall@K for each predicate category separately and averages them for the whole label set, which gives much more weight to the less frequent predicate categories.

4.2 Implementation Details

All the compared methods use ResNeXt-101-FPN [11, 20] as the CNN backbone. Following [21], we optimize the network by AdamW [13] and set the initial learning rate and batchsize to be 3.2 × 10-5 and 4, respectively. The model is trained for 160K iterations, and the learning rate is decayed by a factor of 10 at iterations 94K and 128K. For the two focal losses, we fix the hyper-parameters α and μ to be 0.25 and 2, respectively. The number of the triplet queries is set to 800 so as to capture all the possible relationships, considering the common setting of the number of queries in object detection. It is also worth noting that because we adopt a one-stage SGG method and assign one query to a separate relationship triplet, the NMS used to filter out duplicate relationship proposals is not needed. For the weight of ROI-Align features used in the E2R Fusion, we fix it to be 4. Because of the time limit, we have not yet conducted an ablation study on this hyper-parameter, and leave it to later optimization. For the code environment, all experiments are implemented with the ML framework Pytorch, and trained with 4 NVIDIA RTX A5000 GPUs. For the division of predicate views, we follow [2] and set *above, across, against, along, at, behind, between, in, in front of, near, on, over, under* as the spatial predicate categories, and the rest ones are semantic predicate categories.

4.3 Comparison with the State of the Art

We compare our MVPR to the results of the state-of-the art methods on the VG dataset[14]. As shown in Tab. 1, experimental results prove that our Multi-View Predicate Recognition achieves a new state-of-the-art in the area of biased SGG methods. Specifically, our proposed model shows significant improvement compared to the baseline method on both Recall@K and mRecall@K. Specifically, our MVPR increases the performance baseline method with MEGC by 6.8% on R@100, and the vanilla baseline method by 14.8%. As for the mR@K, our MVPR with MVGC also achieves 4.8%/12.7%/19.3% improvement over the baseline method with MEGC.

4.4 Discussions on Multi-View Graph Constraint

In order to provide a fair comparison and show the superior efficacy of our MVGC clearly, we design a baseline constraint for other SGG methods without the capability of Multi-View Predicate Prediction. We call it Multi-Edge Graph Constraint (MEGC), which is actually a degradation of the condition of no graph constraint. MEGC allows the model to predict multiple predicates for a pair of objects, based on the top K categories selected from the ranked logits. For example, similarly with Eq. 11, the Recall@K with the 2-edge MEGC is calculated as:

$$\boldsymbol{c_{ij}} = TopK(\boldsymbol{s_{ij}}) \tag{12}$$

$$R@K = Recall([S^{top1}_{\boldsymbol{c_{ij}}}, S^{top2}_{\boldsymbol{c_{ij}}}], G_t), \tag{13}$$

where $S^{top1}_{\boldsymbol{c_{ij}}}$ and $S^{top2}_{\boldsymbol{c_{ij}}}$ means the corresponding categories of the rank 1 and rank 2 logits in the outputs of $\boldsymbol{c_{ij}}$. In this way, we calculate the results with MEGC for SSR-CNN-LA [21] to show the difference between choosing top 2 logits and predicting from the two views.

Moreover, we discuss the different criteria used in the SGG task with our MVPR and its baseline method to evaluate the need for using MVGC. As shown in Tab. 2, MVPR with MVGC can produce results that are comparable with those without graph constraints. The experimental results indicate that by conducting the predicate predictions from multiple views, the model is enabled to provide accurate predicate predictions that approximate the results of listing all possibilities.

Another concern about the MVGC is that it is not fair to compare the results with the graph constraint. However, the improvement shows that by predicting the predicates from multiple views, misclassified relationship triplets due to the semantic ambiguity problem have been corrected. Except for providing more options for each triplet, only labeling each instance with all the possible predicate categories can help solve the semantic ambiguity problem. This is less feasible and may cost tremendous time and effort because of the huge search space, which is $M \times N^2$ (M for the number of views and N for the number of nodes).

4.5 Visualization Results

Adaptive Bounding Box for Predicate. To verify that our proposed ABBP is capable of attending to the crucial areas for different predicate categories, we select several predicate categories and visualize the corresponding regressed bounding boxes for the predicate. As illustrated in Fig. 5, our ABBP shows plausible bounding box regression results that concentrate on the areas that contain the high-level semantic information of the predicate categories. For the first row, we have two instances labeled with the predicate category *eating*. We observe that judging the spatial predicates always involve less area than the semantic ones. For the top-right image, although the ground-truth is *eating*, the model predicts the predicates as *on* and *sitting on* according to the attended area, which is also plausible. In the second row, we demonstrate a failure case. The first instance is labeled with *in front of*, but the model struggle at judging the relative position between the fence and the man. This may be a difficult example, which needs further improvement in introducing high-level semantic information to the model. As for the second one, Therefore, our ABBP is capable of removing noisy parts in the union box of the subject and object, thus alleviating the severe intra-class variance and inter-class similarity.

Qualitative Analysis. In order to investigate the ability of our MVPR to predict multiple plausible predicates for an object pair from different views, we visualize the scene graph generation results and compare them (in green) with our baseline method SSR-CNN [21] (in blue). As illustrated in Fig. 4, our model can provide multiple plausible predicate predictions, although only arbitrary one of them is annotated as the ground-truth. For the first row, both *on* and *sitting on* are annotated to describe the relationship between a person and the bench. While the baseline method provides the more informative option *sitting on*, it falls to correctly match the ground-truth. However, our model is able to handle the problem by

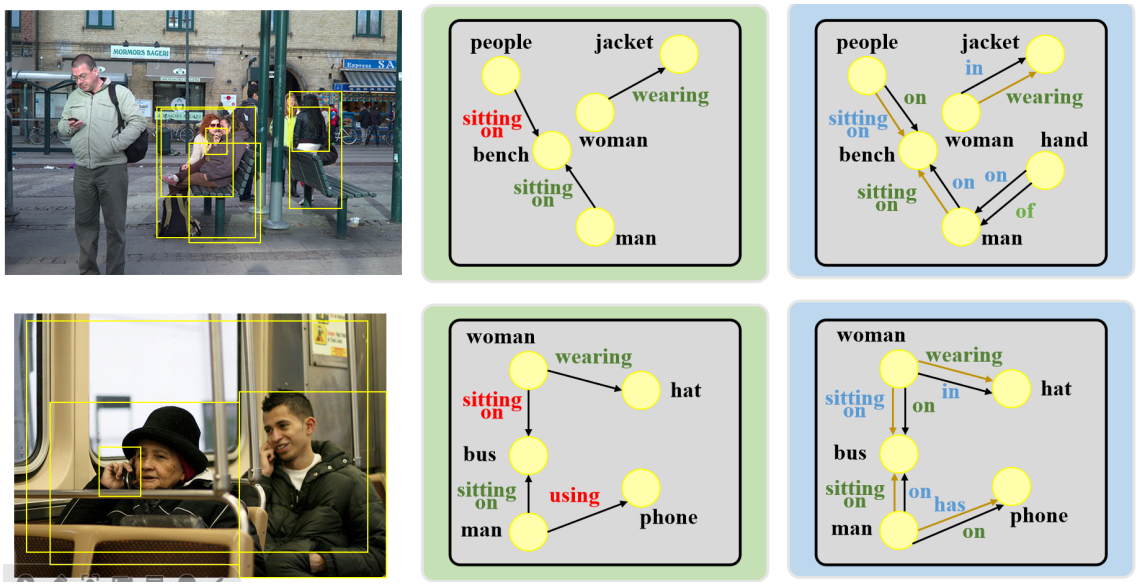

Figure 4: Results of the top 100 object pairs from our model and our baseline model. The results from our MVPR are in the last column, and those from the baseline method are in the second column. The two rows in the second column correspond to SSR-CNN and SSR-CNN-LA, respectively. Predictions that fail to match the ground-truth from the baseline method are marked in red. The blue-marked predicates are those spatial/semantic options predicted by our model that are plausible but fail to match the ground-truth.

providing both options. For the triplet *woman wearing jacket*, our model also predicts it correctly and gives another possible prediction. For the second row, the "unbiased" baseline model always tries to predict a tail predicate, but *on* is used to label the relationship for both *(woman, bus)* and *man, phone*. The results from our model match the ground-truth, while also giving plausible tail predictions. These results demonstrate that our method has significant efficacy in alleviating the ambiguity problem.

5 CONCLUSION

In this work, we propose a novel framework, dubbed Multi-View Predicate Recognition (MVPR), which separately learns to represent different views of predicates and gives multiple possible predicate predictions to alleviate the semantic ambiguity problem. To this end, we propose Adaptive Bounding Box for Predicate and Multi-View Predicate Feature Learning for efficiently separating the feature space of the different views. Moreover, we propose Multi-View Predicate Prediction and Multi-View Graph Constraint to enforce the model to consider the prediction of predicate from different views. Finally, we evaluate our MVPR on the Visual Genome dataset and achieve a new state-of-the-art performance, which proves the efficacy of our method.

Limitation and future work. Our work alleviates the semantic problem by softening the contradiction between the common criterion and the nature of relationships in the SGG datasets. The proposed MVPR achieves a new state-of-the-art by providing predicate predictions that include the most possible options, and shed light on the new aspect for solving the semantic ambiguity problem. However, the division of views is still rough, leaving the ambiguity

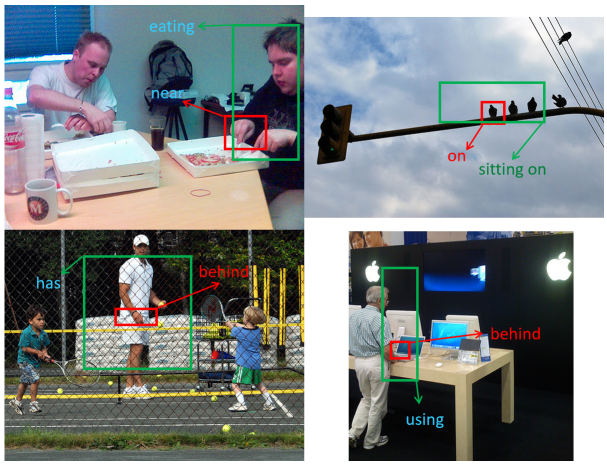

Figure 5: The visualization of the intermediate results in ABBP, which are the learned bounding boxes for the predicate. The results from the spatial view are marked in red and the semantic view is marked in green.

inside the semantic view unsolved. In future work, a more precise view division is needed for introducing a stronger inductive bias and solving the semantic ambiguity problem.

6 ACKNOWLEDGEMENTS

This work is supported in part by the National Natural Science Foundation of China Under Grants No.62176253 and No.U20B2066.

REFERENCES

[1] Desai A., Wu T. Y., Tripathi S., and Vasconcelos N. 2021. Learning of visual relations: The devil is in the tails. In *Proceedings of the IEEE International Conferenceon Computer Vision*. IEEE, 6163–6171.

[2] Goel A., Fernando B., Keller F., and H. Bilen. 2022. Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 15596–15606.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. In *Advances in Neural Information Processing Systems 29*.

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

[5] Xingning Dong, Tian Gan, Xuemeng Song, Jianlong Wu, Yuan Cheng, and Liqiang Nie. 2022. Stacked hybrid-attention andgroup collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 19427–19436.

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

[7] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. 2021. 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, 11109–11119.

[8] Deunsol Jung, Sanghyun Kim, Won Hwa Kim, and Minsu Cho. 2023. Devil's on the Edges: Selective Quad Attention for Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

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

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

[11] Tsung-Yi L., Piotr D., Ross B. G., Kaiming H., Bharath H., and Serge J. B. 2017. Feature pyramidnetworks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2117–2125.

[12] Lin Li, Long Chen, Yifeng Huang, Zhimeng Zhang, Songyang Zhang, and Jun Xiao. 2023. The Devil Is in the Labels: Noisy Label Correction for Robust Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 18869–18878.

[13] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *The International Conference on Learning Representations*. IEEE.

[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. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *International Journal on Computer Vision*. 32–73.

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

[16] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, ChenfengXu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. 2021. Sparse R-CNN: end-to-endobject detection with learnable proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 14454–14463.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and IlliaPolosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*. 5998–6008.

[18] Li W., Zhang H., Bai Q., Zhao G., Jiang N., and Yuan X. 2022. PPDL: Predicate Probability Distribution Based Loss for Unbiased Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 19447–19456.

[19] Lin X., Ding C., Zeng J., and Tao D. 2020. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3746–3753.

[20] Saining X, Ross B. G., Piotr D., Zhuowen T., and Kaiming H. 2017. Aggregated residual transformations fordeep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 5987–5995.

[21] Teng Y. and Wang L. 2022. Structured sparse r-cnn for direct scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 19437–19446.

[22] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. 2022. Panoptic Scene Graph Generation. In *Proceedings of the European Conference on Computer Vision (Part XXVII)*. 178–196.

[23] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and TatSeng Chua. 2017. Visual translation embedding network for visualrelation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3107–3115.

[24] Ji Zhang, Kevin J. Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. 2019. Graphical contrastive losses for scenegraph parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 11535–11543.

[25] Chaofan Zheng, Xinyu Lyu, Lianli Gao, and Bo Dai. 2023. Prototype-based Embedding Network for Scene Graph Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.