Geometric and Semantic Improvement for Unbiased Scene Graph Generation

Ruhui Zhang[1], Pengcheng Xu[2], Kang Kang[1] and You Yang[3*]
[1] College of Computer and Information Science, Chongqing Normal University,
Chongqing 401331, China
[e-mail: RuhuiZhang@outlook.com]
[2] College of Computer Science and Technology, Chongqing University of Posts and Telecommunications,
Chongqing 400065, China
[e-mail: pxu204303@gmail.com]
[3] National Center for Applied Mathematics in Chongqing, Chongqing Normal University,
Chongqing 401331, China
[e-mail: 20130958@cqnu.edu.cn]
*Corresponding author: You Yang

Received March 29, 2023; revised August 22, 2023; accepted September 11, 2023;
published October 31, 2023

Abstract

Scene graphs are structured representations that can clearly convey objects and the relationships between them, but are often heavily biased due to the highly skewed, long-tailed relational labeling in the dataset. Indeed, the visual world itself and its descriptions are biased. Therefore, Unbiased Scene Graph Generation (USGG) prefers to train models to eliminate long-tail effects as much as possible, rather than altering the dataset directly. To this end, we propose Geometric and Semantic Improvement (GSI) for USGG to mitigate this issue. First, to fully exploit the feature information in the images, geometric dimension and semantic dimension enhancement modules are designed. The geometric module is designed from the perspective that the position information between neighboring object pairs will affect each other, which can improve the recall rate of the overall relationship in the dataset. The semantic module further processes the embedded word vector, which can enhance the acquisition of semantic information. Then, to improve the recall rate of the tail data, the Class Balanced Seesaw Loss (CBSLoss) is designed for the tail data. The recall rate of the prediction is improved by penalizing the body or tail relations that are judged incorrectly in the dataset. The experimental findings demonstrate that the GSI method performs better than mainstream models in terms of the mean Recall@K (mR@K) metric in three tasks. The long-tailed imbalance in the Visual Genome 150 (VG150) dataset is addressed better using the GSI method than by most of the existing methods.

Keywords: unbiased scene graph generation, semantic dimension, geometric dimension, CBSLoss, long-tailed imbalance.

This work was supported in part by the Chongqing Postgraduate Joint Training Base Project (Grant No. 2019-45), the Graduate Scientific Research and Innovation Project of Chongqing Normal University (Grant No. YKC20038), and the Health Project of the National Clinical Research Center for Child Health and Disorders (Grant No. NCRCCHD-2022-HP-01).

http://doi.org/10.3837/tiis.2023.10.003 **ISSN : 1976-7277**

1. Introduction

In the field of computer vision, progress has been made toward developing machines that can understand images, videos or other forms of content as well as humans do. By analyzing the relationships between pairs of objects, scene graph generation (SGG) constructs rich semantic information for tasks such as visual question answering (VQA)[1][2].

Inspired by the ladder of causation devised in THE BOOK OF WHY[3], unbiased scene graph generation (USGG) was first proposed by Tang et al.[4] using the Total Direct Effect (TDE) method. This method aims to solve the problem of long-tailed distributions in datasets.

The severe imbalance in the distribution of the dataset has a significant impact on the model. If unconstrained in training, the model is prone to overfitting the tail class, and underfitting the head class. Many methods are effective in suppressing defects in long-tailed datasets. Yu et al.[5] proposed a novel Cognition Tree (CogTree) loss, which differentiated coarse-grained relationships first and then fine-grained relationships through the tree structure. Wang et al.[6] proposed a novel model based on memory that enriches the features of low-frequency relations. Yan et al.[7] proposed Predicate-Correlation Perception Learning (PCPL), which adaptively determines appropriate loss weights according to correlations among the predicate classes. Li et al.[8] proposed a novel model-agnostic Label Semantic Knowledge Distillation (LS-KD) method that can capture correlations between subject-object instances and different predicate categories. Most of the methods deal well with the case where the ground truth is tail relations and the predicted values are head relations. This is also the original intention of the USGG method, i.e., to transform relations from coarse to fine. We found that the prediction accuracy of the tail relationship can be improved if constrained in the case where the ground truth is the tail relationship and the predicted value is the incorrect tail relationship. However, the above methods focus on improving the predicted recall between tail predicates and other predicates and do not intentionally provide additional corrections for the case of misjudged tail predicates. To effectively suppress the long-tail problem, we designed the Class Balanced Seesaw Loss (CBSLoss), which improves the correct prediction rate of the tail samples.

Visual relations can be divided into four categories: geometric, semantic, possessive, and miscellaneous. We start with semantic and geometric dimensions to enhance the capture of model features. Specifically, at the semantic level, we improve the classification accuracy and prediction accuracy of the model by deep learning processing of the model input data. In real world applications, adjacent objects may have numerous relationships. For instance, consider the situations "person uses computer" and "person lying on table." The information about the position of the person can be used to deduce the spatial locations of the computer and table as "computer on table." Obviously, information about the proposed frame positions of neighboring object pairs for each object pair in the same image can assist in the training of that object pair. Thus, we designed a geometry module that enhances the acquisition of geometric information. By considering and processing these two dimensions together, we are able to better understand and utilize the features of the model, thus improving the performance and effectiveness of the model.

The contributions of this work can be summarized as follows:

(1) We designed a new geometric module to assist in the training of each object pair in the same image by utilizing the positional information of the bounding boxes of neighboring object pairs for that object pair. This helps the model to better understand the positional relationship between objects, thus improving the recall of scene graph generation.

(2) We devised a new semantic module that exploits these rich semantic relationships.

(3) We designed a new loss function, CBSLoss, for the tail relation, to improve the tail

relation recall by introducing a penalty factor when the predicted relation is a tail relation and the prediction is incorrect.

2. Related Work

2.1 Class Imbalance

In the actual world, some relationships are distributed in only a few classes, whereas other relationships are dispersed among most categories. This is known as class imbalance. Deep learning faces significant obstacles as a result of this universal natural occurrence. Numerous improvements to one-stage[9][10] and two-stage[11][12] object detection have addressed the issue of foreground-background class imbalance. However, SGG is subject to foreground-foreground class imbalance.

USGG methods for resolving long-tailed data issues can be divided into three categories[13]. (1) Data augmentation on resampling for tail data. Knyazev et al.[14] proposed a data augmentation technique based on Generative Adversarial Networks (GANs). Yao et al. [15] proposed a visual distant supervision technique without applying any human-labeled data and devised a denoising framework to reduce noise. However, these methods do not perform well in the case of strong correlation between predicate labels. This phenomenon arises partly because re-balancing strategy simply utilizes the frequency of classes while ignoring their semantic relatedness[7]. For the existing re-balancing strategies fail to increase the diversity of the relation triplet features of each predicate, Li et al.[16] proposed a novel Compositional Feature Augmentation (CFA) strategy which is the first work to mitigate bias problem by increasing the diversity of triplet features in the USGG. (2) Elaborately designed training curricula or learning losses. Wei et al.[17] proposed a higher-order structure embedded network (HOSE-Net), which consists of structure-aware embedding-to-classifier (SEC) and hierarchical semantic aggregation (HSA) modules. This method incorporates structural information in the output space and reduces the number of subspaces. To reduce error propagation, Li et al.[18] devised the bipartite graph neural network (BGNN). Chiou et al.[19] proposed the dynamic label frequency estimation (DLFE) method to address the reporting bias problem. Suhail et al.[20] designed the energy-based model (EBM), which incorporates the scene graph structure into the learning framework. (3) Disentangling biased and unbiased representations. The TDE method employs counterfactual surgery to produce unbiased predictions. These methods are effective in improving the long-tail effect. However, the causal models constructed are always causally insufficient due to dataset noise that makes confounders unobservable. Sun et al.[21] proposed Two-stage Causal Modeling (TsCM). It uses long-tailed distributions and semantic confusions as confounders for structural causal modeling, and then decouples the causal intervention into two stages.

2.2 Utilization of Semantic and Geometric Features

Several models make clever use of semantic information. For example, the random intercept factor analysis (RiFa) method[22] reveals the rich semantic features of relations. This method uses the semantic distinctions between the subject and object of the same entity to prevent biased relations. The PCPL method adaptively determines the loss weights by using the correlations between the predicate classes.

In terms of geometric information, the Motifs[23] model determines the number of relationships according to the object height. He et al.[24] incorporated relative position coding

into relationship matrices. The EBM model is trained on a loss function that incorporates the structure of the output space. Obviously, semantic and geometric information are crucial for the prediction of relationships. The GSI method effectively utilizes both semantic and geometric information, and mitigates the long-tail effect.

3. Methods

The base CogTree method is inspired by the prefrontal cortex and mimics the way human intuition distinguishes between associations that are significantly different from those that are similar to them. The main idea is to first use a biased SGG model to generate subtrees based on the frequency ranking of misclassified associations and then engage the model with subtree structures to mitigate long-tail effects. Our GSI method is shown in **Fig. 1**. The improvements are shown in color, and the subtree constructed in CogTree continues.

The structure of the GSI method based on the Transformer model is shown in **Fig. 1**. In addition, the GSI method can be added to any of the SGG base models (Motifs, visual context tree model (VCTree)[25] and Transformer) and their improved models. Since our method is model independent, the Transformer model is not changed, but follows the dual Transformer model in the CogTree method.

The Transformer model is selected as an example. The dataset is first passed through the object detection model and then through the semantic and geometric modules to enhance the acquisition of image information before entering the Transformer model. After the Transformer model output, the relationship between object pairs is predicted using the loss function CBSLoss. By using the tree structure of the subtree and the result of CBSLoss, the Tree-based Class Balanced Loss (TCBLoss) of the current node is generated, and finally, the two loss functions are combined to generate the relationship between the object pairs and the final scene graphs.

The object encoder structure of the Transformer model, contains multiple object-to-object (O2O) blocks. The components of both the O2O and relation-to-object (R2O) blocks are the attention module, residual connection, layer normalization, feedforward module, residual connection, and layer normalization. The difference between the R2O and O2O blocks is the attention module. The O2O block uses self-attention while the R2O block uses cross-attention. In both the object decoder and relational decoder, the fully connected layer and softmax layer are used.

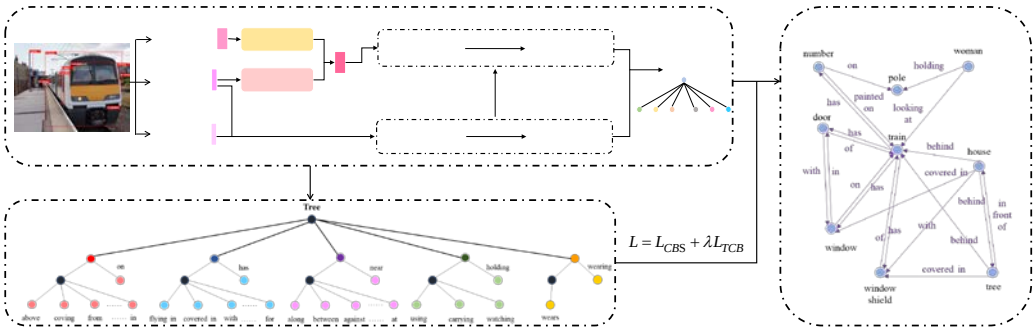


Fig. 1. The GSI method has three components: first, the Faster R-CNN object detection model is used to acquire object information; then, the double Transformer model (relation transformer and object transformer) and CBSLoss are used to predict each relationship probability; and finally, subtrees are used to generate the final scene graph.

3.1 Geometric Feature Improvement

The recall of SGG models can be improved by efficiently capturing the geometric information. Thus, this module improves on the CogTree method inability to fully exploit position information by using position embedding alone. Moreover, this module effectively addresses noise interference in the model by introducing thresholding.

The two improvement modules for geometric and semantic dimension features are shown in **Fig. 2**. First, a series of convolution operations are performed on the semantic dimension features; then, splicing operations are performed on the semantic features, the new geometric features obtained from the geometric module, and the initial union region features; finally, a full join operation is performed on the spliced features to obtain the new union features. If the geometry module or semantic module is used separately in the ablation experiment, only the original joint region features are fully connected with the corresponding parts. $S_b \in ¡^{1\times m\times 4}$ and $O_b \in ¡^{1\times m\times 4}$ represent the position information in the batch image object pair <subject, object> and m is the total number of relational pairs in the batch sheet. Moreover, the number of columns (in this case, 4) corresponds to the bottom-left coordinate values x_1 and y_1 and the top-right coordinate values x_2 and y_2. The equations for processing S_b and O_b are shown in (1):

$$X^{'} = G_{drop}(\sigma(G_{FC}(G_{drop}(\sigma(G_{FC}(X)))))), \tag{1}$$

where X represents the input to S_b or O_b and $X^{'}$ represents the $S_b^{'} \in ¡^{1\times m\times 4}$ or $O_b^{'} \in ¡^{1\times m\times 4}$ obtained after processing. σ is the rectified linear unit (ReLU) function, G_{FC} is a linear layer, and G_{drop} is a dropout layer.

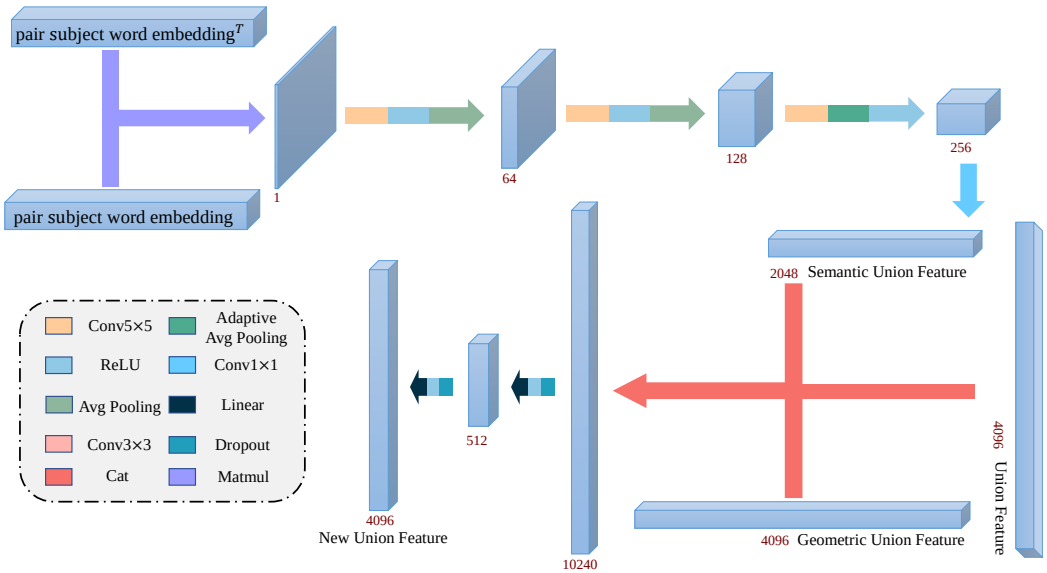


Fig. 2. The improvements in the geometric and semantic features are shown schematically. Convolutions are performed on the semantic features in the first line; then, the semantic features, geometric union features, and initial geometric union features are concatenated in the second line.

The initial union region feature vector $F \in ¡^{4096\times1\times1}$ corresponds to the object pair information in each batch of images. The first step of the improved geometric feature module is to determine the union region features $U \in ¡^{1\times k\times4}$ in object pairs in each picture, where k represents the number of object pairs in each image. In the second step, the position information of the union region between object pairs and object pairs in U is calculated to obtain the union region feature position matrix $C\_U \in ¡^{1\times k\times k\times4}$. In the third step, the $IoU \in ¡^{1\times k\times k\times1}$ of the union region location is calculated. The inputs of this step are C_U_{ij} and U_i, where i represents the current object pair and j represents another object pair in the same image.

The maximum value of each row in IoU_{ij} is compared with the threshold s ($s \in (0, 1)$), and if the value is larger than s, subscripts i and j are obtained, and the corresponding value of the union region feature vector F_j is put into M_i according to the subscript j, as shown in (2). Since the values of IoU_{ij} in the experiments are mostly between 0.75 and 0.95, s is set to 0.80 in this paper.

$$M_i = \begin{cases} F_j, \max(IoU_i) > s \\ 0, else \end{cases} \tag{2}$$

Finally, the matrices F and M are fully connected to obtain the new union feature $F' \in ¡^{4096\times1\times1}$ as (3):

$$F' = G_{drop}(\sigma(G_{FC}(G_{drop}(\sigma(G_{FC}(F, M)))))), \tag{3}$$

where F' is the improved union feature that is used as the input to the Transformer model to ensure that the model acquires rich geometric features.

3.2 Semantic Feature Improvement

The effective acquisition of semantic features in SGG models is critical for increasing SGG model accuracy. In the base method, the word embeddings from the GloVe model[26] and the object names are used as inputs to the transformer model.

Fig. 2 shows a schematic diagram of the enhanced semantic and geometric features. Assume that the subject and object labels in the pair <subject, object> are obtained; then, the word embeddings obtained by feeding these labels into the GloVe model are C_i and C_j, and the union word embedding $X_{ij} \in ¡^{1\times200\times200}$ can be formulated as (4):

$$X_{ij} = C_i^T \cdot C_j, \tag{4}$$

where $\cdot$ is the matrix dot product. To obtain richer information, this module performs a series of convolution operations on the obtained X_{ij}. First, 5×5 convolutions, the ReLU function, average pooling (avg pooling), and two 3×3 convolutions are applied. Then, the module performs adaptive average pooling. Finally, the module performs two 1×1 convolutions and applies the ReLU function to obtain $X_{ij}^{'} \in ¡^{4096\times1\times1}$. $X_{ij}^{'}$ is shown in (5).

$$X_{ij}^{'} = G_{conv}^{1\times1}(\sigma(G_{pool2}(G_{conv}^{3\times3}(...(G_{pool1}(X_{ij}))...)))), \tag{5}$$

where G_{pool1} represents avg pooling, G_{pool2} denotes adaptive avg pooling, and $G_{conv}^{n\times n}$ indicates an $n\times n$ convolution (where n is set to 1, 3, or 5).

To better use the joint word embedding X_{ij}, this feature is integrated into the union feature $F \in ¡^{4096\times1\times1}$, yielding $F' \in ¡^{4096\times1\times1}$ as follows (6):

$$F' = G_{drop}(\sigma(G_{FC}(G_{drop}(\sigma(G_{FC}(X_{ij}', F)))))) . \tag{6}$$

3.3 Learning with Tree Loss

3.3.1 CBSLoss

Cui et al.[27] proposed Class Balanced Loss (CBLoss) to address the effects of long-tailed data distributions on image segmentation tasks.

However, CBLoss ignores the issue that when the situation arises where the ground truth is the tail predicate and the predicted value is not the correct tail predicate, it is not possible to constrain the situation effectively. This is because the hyperparameter of the weights n_i takes the value of the number of ground-truth categories, and the formula cannot effectively "penalize" such cases when the number is relatively small (tail predicate). In response to this situation, a new loss function, the CBSLoss, which is inspired by the seesaw loss function[28], is created. When the ground truth label T_i is a tail predicate and the predicted probability T_j is also a tail predicate ($T_i \neq T_j$), the weight W_i is defined as (7):

$$W_i = (\frac{1-\beta}{1-\beta^{n_i}})(\frac{T_j}{T_i})^q , \tag{7}$$

where the hyperparameter β is set to 0.999 ($\beta \in [0,1)$), as in the literature[5]. Moreover, the hyperparameter q is set to 3.0. The experimental results are shown in **Table 1**. Here, n_i represents the number of true values of category i.

The novel CBSLoss loss function is defined in (8), where P_{pred} is the predicted probability and g_i is the ground truth label corresponding to node i. This loss function includes the softmax function and corresponding class-balanced weight W_{g_i}.

$$L_{CBS} = -W_{g_i} \log(\frac{\exp(p_{g_i})}{\sum_{p_j \in p_{pred}} \exp(p_j)}) . \tag{8}$$

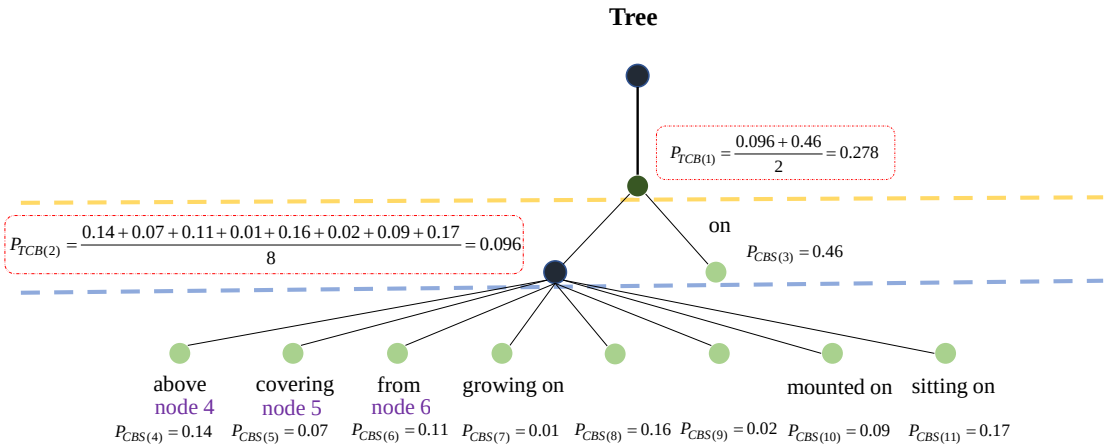


Fig. 3. The TCBLoss of the parent node can be calculated from the CBSLoss of the leaf node.

3.3.2 TCBLoss

In the structure of the subtree, the GSI method still uses the tree-based class balanced loss (TCBLoss), which is used in the base method [5]. TCBLoss is a loss function constructed for leaf nodes that consists of weight and a softmax function, as shown in (9). **Fig. 3** shows an example of the calculation of TCBLoss.

$$L_{TCB} = \frac{1}{k}\sum_{k=1}^{K} -W_{S_k} \log\left(\frac{\exp(Z_{S_k})}{\sum_{z_j \in Z_{(S_{k-1})}} \exp(z_j)}\right), \tag{9}$$

where W_{s_k} is calculated as shown in (7) and K represents the number of leaves except for the number of nodes. The inputs to the softmax function are the average of the sum of the probabilities of all brother nodes of the predicate node as the probability of its parent node $Z_{s_{(k-1)}}$, which is the probability Z_{s_k} of the node obtained by CBSLoss. The total loss function of the GSI method is the weighted sum of CBSLoss and TCBLoss, and the equation is shown in (10), where the hyperparameter λ is 0.7, which is consistent with the literature [5].

$$L = L_{CBS} + \lambda L_{TCB}. \tag{10}$$

4. Experiments

4.1 Dataset, Tasks and Metrics

We use the Visual Genome 150 (VG150)[29] dataset, which is based on the Visual Genome dataset (VGD)[30], for experiments. This dataset includes 50 relationships and the 150 most common object categories in the VGD.

In this paper, the mean recall@K (mR@K) metric is examined in three tasks: predicate classification (PredCls), scene graph classification (SGCls), and scene graph detection (SGDet). The evaluation metrics for K predictions (@20, @50, and @100) are compared.

The mR@K statistic represents the average the predicate category R@K values. This metric provides a good indication of the model's influence on the dataset's unbalanced distribution.

The GSI method uses ResNeXt101-FPN as its backbone and is trained in the Ubuntu operating system with an Intel Core i7-10700KF CPU with 32 GB RAM and an M40 GPU. Moreover, the initial learning rate is set to 0.001, the weight decay is set to 0.0001, the optimizer is the stochastic gradient descent (SGD) optimizer, the total number of training iterations is set to 50000, and the batch size is set to 4. The other hyperparameters are consistent with those used in the literature [5].

The experimental results for various values of the hyperparameter q in CBSLoss are shown in **Table 1**. The experimental results show that when q is set to 3.0, the method achieves the best effect. A q value of 0.0 indicates that CBSLoss degenerates to CBLoss.

The threshold s used in the geometry enhancement module is selected because the maximum value of IOU between object pairs in the same image is typically between 0.75 and 0.95. Thus, to better capture the features of the most adjacent object pairs, we set the threshold s to 0.80.

Table 1. q values

q	**mR@20**	**mR@50**	**mR@100**
0.00	23.62	28.03	29.86
0.50	0.59	1.00	1.38
1.00	20.84	26.19	28.50
1.50	0.59	1.00	1.38
2.00	23.92	28.17	30.44
3.00	22.43	**28.70**	**30.54**
4.00	**23.95**	27.75	29.55
5.00	23.45	28.33	29.48

4.2 Comparison with state-of-the-art methods

The GSI method is compared with three baseline models, Motifs, VCTree, and the Transformer, as well as their corresponding improved methods: TDE, STL[31], PCPL, CogTree, NARE[32], CAME[33] and LS-KD. In addition, the mR@K values of the IMP[34], KERN[35], GPS-Net[36], BGNN and NLS[37] models are listed in **Table 2**.

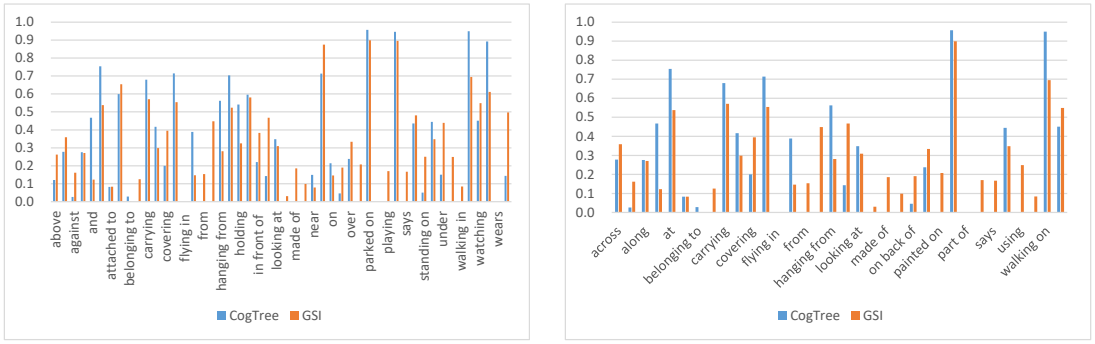
Table 2 demonstrates the following findings: (1) Among the three baseline models, the Transformer performs better than Motifs and VCTree due to its advantages in discriminating objects and generating relationship representations. (2) The GSI method can be added to the baseline models, and the results of the combined models are better than the results of the models alone for all three tasks. Thus, the combined methods are stable. (3) In the PredCls task, the Transformer-based GSI method clearly outperforms the VCTree-based and Motifs-based GSI methods. In the SGCls task, the VCTree-based GSI method significantly outperforms the Motifs-based and Transformer-based GSI methods. Finally, in the SGDet task, the Motifs-based GSI method performs better than the VCTree-based and Transformer-based GSI methods. (4) Even when compared to state-of-the-art models such as BGNN, NLS, and the VCTree-based NARE, LS-KD method, the GSI method still shows advantages in the three tasks. (5) The Motifs-based GSI method shows a significant improvement over the Motifs-based CogTree method on all three tasks. The experimental results show that the Motifs-based GSI method is a more significant upgrade than the VCTree-based and Transformer-based GSI methods. Thus, our method effectively mitigates the impact of long-tailed distributions in the dataset.

Table 2. Comparison with state-of-the-art methods

Model	Method	PredCls	SGCls	SGDet
		mR@20/50/100	**mR@20/50/100**	**mR@20/50/100**
IMP[34]	-	-/9.80/10.50	-/5.80/6.00	-/3.80/4.80
KERN[35]	-	-/17.70/19.20	-/9.40/10.00	-/6.40/7.30
GPS-Net[36]	-	17.40/21.3/22.8	10.00/11.80/12.60	6.90/8.70/9.80
BGNN[18]	-	-/30.40/32.90	-/14.30/16.50	-/10.70/12.60
NLS[37]	-	13.30/17.70/19.50	8.30/10.40/11.10	5.30/7.30/8.70
Motifs[23]	-	10.80/14.00/15.30	6.30/7.70/8.20	4.20/5.70/6.60
	+TDE[4]	18.50/25.50/29.10	9.80/13.10/14.90	5.80/8.20/9.80
	+PCPL[7]	19.30/24.30/26.10	9.90/12.00/12.70	8.00/10.70/12.60
	+CogTree[5]	20.90/26.40/29.00	12.10/14.90/16.10	7.90/10.40/11.80
	+CAME[33]	18.10/26.20/32.00	10.50/15.10/18.00	6.70/9.30/12.10
	+LS-KD[8]	-/24.10/27.40	-/13.80/15.20	-/9.70/11.50
	+**Ours**	22.76/28.66/30.91	13.91/16.75/17.74	**8.58**/**11.67**/**13.88**
VCTree[25]	-	11.70/14.90/16.10	6.20/7.50/7.90	4.20/5.70/6.90
	+TDE[4]	18.40/25.40/28.70	8.90/12.20/14.00	6.90/9.30/11.10
	+STL[30]	14.30/21.40/23.50	10.50/14.60/16.60	5.10/7.10/8.40
	+PCPL[7]	18.70/22.80/24.50	12.70/15.20/16.10	8.10/10.80/12.60
	+CogTree[5]	21.80/25.49/26.97	15.40/18.80/19.90	7.80/10.40/12.10
	+NARE[31]	18.00/21.70/23.10	11.90/14.10/15.20	7.10/8.20/8.70
	+CAME[33]	18.90/26.60/32.00	11.70/17.00/20.50	5.90/8.70/10.80
	+LS-KD[8]	-/24.20/27.10	-/17.30/19.10	-/9.70/11.30
	+**Ours**	23.23/29.25/31.57	**17.50**/**20.68**/**21.93**	8.17/11.34/13.47
Transformer	-	14.40/18.50/20.20	8.60/11.50/12.30	5.60/7.70/9.00
	+TDE[4]	18.90/25.30/28.40	9.80/13.10/14.70	6.00/8.50/10.20
	+CogTree[5]	23.62/28.03/29.86	13.00/15.70/16.60	8.35/10.34/12.02
	+**Ours**	**24.86**/**29.78**/**32.90**	13.29/15.81/16.70	8.35/10.69/12.33

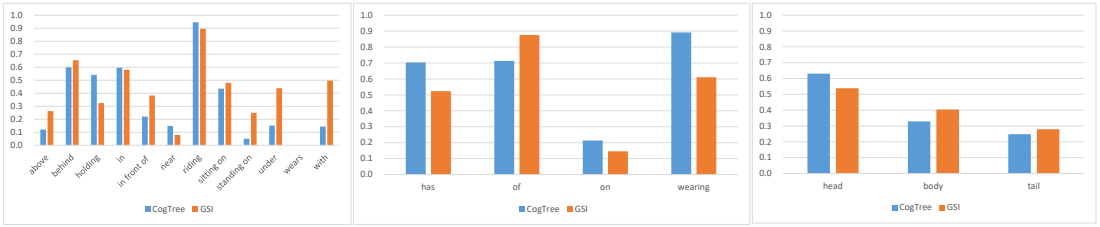
The R@100 performance of each class is shown in **Fig. 4** (a), and the head, body and tail distributions in the dataset and their respective averages are shown in **Fig. 4** (d), (c), (b) and (e). We note the following findings: (1) The GSI method outperforms the CogTree method in terms of the R@100 metric. This result shows that the GSI method significantly improves both the mR@100 metric and the R@100 metric. (2) **Fig. 4** (e) and (b) show that the GSI method performs better than the CogTree method, even in the body of the data. (3) **Fig. 4** (a) and (b) demonstrate that the distribution between the classes is relatively even.

Fig. 5 shows four pairs of visualization examples. In the first picture, the GSI method detects the relationship between a skier and a person, and the relationship between the skier and person is more specific than the ground-truth label ("standing on" is more specific than "on"). In the second, third and fourth images, the GSI method detects more object pairs and their relationships than the other methods.



(a) 50 relationships distribution

(b) tail relationships distribution



(c) body relationships distribution

(d) head relationships distribution

(e) average distribution

Fig. 4. The prediction recall distribution of each relation in the PredCls task.

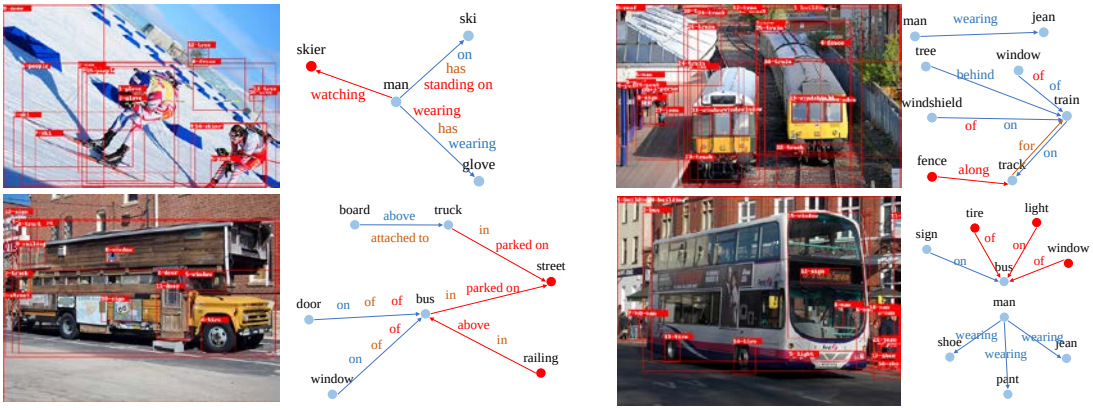


Fig. 5. Visualization of SGG by the ground truth (blue), Transformer + CogTree (orange) and Transformer-based GSI (red) methods. A blue line between two nodes indicates that all three models predict the same relationship; if only orange and red lines are visible, the ground truth and Transformer-based GSI model predict the same relationship; and if only a red line is visible, only the Transformer-based GSI model predicts the relationship between the object pairs.

4.3 Ablation Experiments

The Transformer-based GSI method was investigated through four ablation studies, and the experimental results are shown in **Table 3**. In the table, the following letters are used to represent the four improved parts: "S" represents the semantic enhancement module, "G" represents the geometric enhancement module, and L_{CBS} and L_{CB} represent the improved CBSLoss function and the original CBLoss function, respectively.

Table 3. Ablation Experiments

Model	PredCls	SGCls	SGDet
	mR@20/50/100	**mR@20/50/100**	**mR@20/50/100**
Transformer + CogTree	21.26/27.14/29.68	13.00/15.70/16.60	8.35/10.34/12.02
Transformer + CogTree + S + L_{CB}	22.14/27.70/30.55	12.95/15.44/16.55	**8.45**/10.60/12.18
Transformer + CogTree + G + L_{CB}	23.69/29.22/31.75	12.24/15.80/16.64	8.07/**10.69**/12.32
Transformer + CogTree + L_{CBS}	22.43/28.70/30.54	13.20/15.73/**16.72**	7.95/10.32/12.24
Transformer + Ours	**24.86**/**29.78**/**32.90**	**13.29**/**15.81**/16.70	8.35/**10.69**/**12.33**

Table 3 shows the results of the ablation experiments indicating the effect of each module on the model. We obtain several conclusions. (1) In the PredCls task, the geometric module improves the model most significantly, followed by the loss function module and the semantic module. (2) In the SGDet task, the geometric module improves the model most significantly, followed by the loss function module. (3) The semantic and geometric modules have the largest enhancement effect in the PredCls task due to the use of the ground-truth bounding box and labels in this task. The word embedding vector generated by the semantic module according to the labels relies heavily on the correctness of the labels. Moreover, the geometric module relies on the correctness of the bounding box to determine the degree of overlap among adjacent objects.

5. Conclusions

In this paper, we present the GSI method, a novel method for improving the mR@K metric on base model. This method can be applied to the basic SGG models and their improved models such as Motifs, VCTree and Transformer, and is model-independent. First, a geometric enhancement module is designed based on the idea that adjacent objects are related to the position relation. Second, a semantic enhancement module is designed to further enhance the semantic information. Finally, L_{CBS} is designed to punish the incorrect tail relation, which also effectively improves the accuracy of tail relation prediction. In addition, this paper has some limitations, e.g., not introducing too much external information. In future work, we will further improve the long tail distribution of the data. The next step will be to consider incorporating more a priori, common-sense information into the model to improve model prediction performance.

References

[1] V. Damodaran, S. Chakravarthy, A. Kumar, et al, "Understanding the role of scene graphs in visual question answering," *arXiv:2101.05479v2*, pp. 1-12, Jan. 2021. Article (CrossRef Link)

[2] M. Hildebrandt, H. Li, R. Koner, et al, "Scene graph reasoning for visual question answering," *arXiv:2007.01072v1*, pp. 1-5, Jul. 2020. Article (CrossRef Link)

[3] J. Pearl, D. Mackenzie, "The book of why: the new science of cause and effect," in *the ladder of causation*, New York, USA: Basic books, 2018, pp. 27-58.

[4] K. Tang, Y. Niu, J. Huang, et al, "Unbiased Scene Graph Generation from Biased Training," in *Proc. of CVPR 2020: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 3716-3725, June 14-19, 2020. Article (CrossRef Link)
[5] J. Yu, Y. Chai, Y. Wang, et al, "Cogtree: Cognition tree loss for unbiased scene graph generation.," *arXiv:2009.07526v2*, pp.1-7, Sep. 2020. Article (CrossRef Link)
[6] W. Wang, R. Liu, M. Wang, et al, "Memory-Based Network for Scene Graph with Unbalanced Relations," in *Proc. of ACMM 2020: Proceedings of the ACM Multimedia Conference on Multimedia Conference*, Seattle, USA, pp. 2400-2408, Oct 12-16. 2020. Article (CrossRef Link)
[7] S. Yan, C. Shen, Z. Jin, et al, "PCPL: Predicate-Correlation Perception Learning for Unbiased Scene Graph Generation," in *Proc. of ACMM 2020: Proceedings of the ACM Multimedia Conference on Multimedia Conference*, Seattle, USA, pp. 265-273, Oct 12-16. 2020. Article (CrossRef Link)
[8] L. Li, J. Xiao, H. Shi, et al, "Label semantic knowledge distillation for unbiased scene graph generation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1-11, 2023. Article (CrossRef Link)
[9] Z. Ge, S. Liu, F. Wang, et al, "Yolox: Exceeding yolo series in 2021," *arXiv:2107.08430v2*, pp.1-7, Jul. 2021. Article (CrossRef Link)
[10] H. Law, J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proc. of ECCV 2018: Proceedings of the European Conference on Computer Vision*, Munich, Germany, pp. 734-750, Sep 8-14. 2018. Article (CrossRef Link)
[11] S. Ren, K. He, R, Girshick, et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *Advances in neural information processing systems*, pp. 28-34, 2015. Article (CrossRef Link)
[12] K. He, G. Gkioxari, P. Dollar, et al, "Mask R-CNN," in *Proc. of ICCV 2017: Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, pp. 2961-2969, Oct 22-29. 2017. Article (CrossRef Link)
[13] Y. Zhang, B. Kang, B. Hooi, et al, "Deep long-tailed learning: A survey," *arXiv:2110.04596v1*, pp. 1-20, Oct. 2021. Article (CrossRef Link)
[14] B. Knyazev, H. de Vries, C. Cangea, et al, "Generative compositional augmentations for scene graph prediction," in *Proc. of ICCV 2021: Proceedings of the IEEE International Conference on Computer Vision*, Montreal, Canada, pp. 15827-15837, Oct 11-17. 2021. Article (CrossRef Link)
[15] Y. Yao, A. Zhang, X. Han, et al, "Visual distant supervision for scene graph generation," in *Proc. of ICCV 2021: Proceedings of the IEEE International Conference on Computer Vision, Montreal*, Canada, pp. 15816-15826, Oct 11-17. 2021. Article (CrossRef Link)
[16] L. Li, G. Chen, J. Xiao, et al, "Compositional Feature Augmentation for Unbiased Scene Graph Generation," *arXiv: 2308.06712v1*, pp.1-11, Aug. 2023. Article (CrossRef Link)
[17] M. Wei, C. Yuan, X. Yue, et al, "HOSE-Net: Higher Order Structure Embedded Network for Scene Graph Generation," in *Proc. of ACMM 2020: Proceedings of the ACM Multimedia Conference on Multimedia Conference*, Seattle, USA, pp. 1846-1854, Oct 12-16. 2020. Article (CrossRef Link)
[18] R. Li, S. Zhang, B. Wan, et al, "Bipartite Graph Network With Adaptive Message Passing for Unbiased Scene Graph Generation," in *Proc. of CVPR 2021: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Nashville, Tennessee, USA, pp. 11109-11119, Jun 19-25. 2021. Article (CrossRef Link)
[19] M. J. Chiou, H. Ding, H. Yan, et al, "Recovering the Unbiased Scene Graphs from the Biased Ones," in *Proc. of ACMM 2021: Proceedings of the ACM Multimedia Conference on Multimedia Conference*, Chengdu, China, pp. 1581-1590, Oct 20-24. 2021. Article (CrossRef Link)
[20] M. Suhail, A. Mittal, B. Siddiquie, et al, "Energy-based learning for scene graph generation," in *Proc. of CVPR 2021: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Nashville, Tennessee, USA, pp. 13936-13945, Jun 19-25. 2021. Article (CrossRef Link)

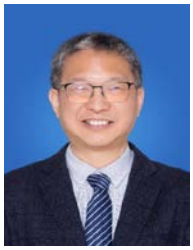
[21] S. Sun, S. Zhi, Q. Liao, et al, "Unbiased Scene Graph Generation via Two-stage Causal Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12562-12580, Oct. 2023. Article (CrossRef Link)
[22] B. Wen, J. Luo, X. Liu, et al, "Unbiased scene graph generation via rich and fair semantic extraction," *arXiv:2002.00176*, pp.1-9, Feb. 2020. Article (CrossRef Link)
[23] R. Zellers, M. Yatskar, S. Thomson, et al, "Neural Motifs: Scene Graph Parsing With Global Context," in *Proc. of CVPR 2018: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 5831-5840, Jun 18-22. 2018. Article (CrossRef Link)
[24] T. He, L. Gao, J. Song, et al, "Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation," *arXiv:2006.07585v1*, pp.1-7, Jun. 2020. Article (CrossRef Link)
[25] K. Tang, H. Zhang, B. Wu, et al, "Learning to compose dynamic tree structures for visual contexts," in *Proc. of CVPR 2019: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 6619-6628, Jun 16-20. 2019. Article (CrossRef Link)
[26] Y. Y. Lee, H. Ke, H. H. Huang, et al, "Less is more: Filtering abnormal dimensions in glove," in *Proc. of WWW 2016: Proceedings of the 25th International Conference Companion on World Wide Web*, Montreal, Canada, pp. 71-72, Apr 11-15. 2016. Article (CrossRef Link)
[27] Y. Cui, M. Jia, T. Y. Lin, et al, "Class-balanced loss based on effective number of samples," in *Proc. of CVPR 2019: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 9268-9277, Jun 16-20. 2019. Article (CrossRef Link)
[28] J. Wang, W. Zhang, Y. Zang, et al, "Seesaw loss for long-tailed instance segmentation," in *Proc. of CVPR 2021: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville*, Tennessee, USA, pp. 9695-9704, Jun 19-25. 2021. Article (CrossRef Link)
[29] D. Xu, Y. Zhu, C. B. Choy, et al, "Scene graph generation by iterative message passing," in *Proc. of CVPR 2017: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 5410-5419, Jul 21-26. 2017. Article (CrossRef Link)
[30] R. Krishna, Y. Zhu, O. Groth, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol.123, no.1, pp. 32-73, Feb. 2017. Article (CrossRef Link)
[31] D. Chen, X. Liang, Y. Wang, et al, "Soft transfer learning via gradient diagnosis for visual relationship detection," in *Proc. of WACV 2019: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, pp. 1118-1126, Jan 7-11. 2019. Article (CrossRef Link)
[32] A. Goel, B. Fernando, F. Keller, et al, "Not All Relations are Equal: Mining Informative Labels for Scene Graph Generation," in *Proc. of CVPR 2022: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, Louisiana, pp. 15596-15606, Jun 19-24. 2022. Article (CrossRef Link)
[33] L. Zhou, Y. Zhou, T. L. Lam, et al, "Context-aware Mixture-of-Experts for Unbiased Scene Graph Generation," *arXiv:2208.07109v2*, pp.1-11, Aug. 2022. Article (CrossRef Link)
[34] D. Xu, Y. Zhu, C.B. Choy, et al, "Scene graph generation by iterative message passing," in *Proc. of CVPR 2017: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, Hawaii, USA, pp. 5410-5419, July 21-26. 2017. Article (CrossRef Link)
[35] T. Chen, W. Yu, R. Chen, et al, "Knowledge-embedded routing network for scene graph generation," in *Proc. of CVPR 2019: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Los Angeles, CA, pp. 6163-6171, Jun 15-21. 2019. Article (CrossRef Link)
[36] X. Lin, C. Ding, J. Zeng, et al, "GPS-Net: Graph Property Sensing Network for Scene Graph Generation," in *Proc. of CVPR 2020: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, pp. 3746-3753, June 14-19. 2020. Article (CrossRef Link)

[37] Y. Zhong, J. Shi, J. Yang, et al, "Learning to generate scene graph from natural language supervision," in *Proc. of ICCV 2021: Proceedings of the IEEE International Conference on Computer Vision*, Montreal, Canada, pp. 1823-1834, Oct 11-17. 2021. Article (CrossRef Link)

Ruhui Zhang received the bachelor's degree in software engineering from Nanjing Xiaozhuang University, Nanjing, China, in 2020. She received the M.S. degree in software engineering from Chongqing Normal University, Chongqing, China, in 2023. Her research interests include computer vision and scene graph generation, etc.

Pengcheng Xu received the M.S. degree in computer technology from Chongqing Normal University, Chongqing, China, in 2022. He is currently a Ph.D. student in computer science and technology from Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include computer vision and single image super-resolution.

Kang Kang received the bachelor's degree in information management and information system from Rongzhi college of Chongqing Technology and Business University, Chongqing, China, in 2020. He received the M.S. degree in computer technology from Chongqing Normal University, Chongqing, China, in 2023. His research interests include deep learning, scene graph generation, etc.



You Yang received the Ph.D. degree in computer application technology from Beihang University, Beijing, China, in 2010. He is currently a professor of National Center for Applied Mathematics in Chongqing. His research interests include computer vision and digital image processing.