



# Fine-Grained Scene Graph Generation with Overlap Region and Geometrical Center

Y. Q. Zhao<sup>1,2,3</sup> , Z. Jin<sup>†1,3</sup> , H. Y. Zhao<sup>1,3</sup>, F. Zhang<sup>2</sup>, Z. W. Tao<sup>1,2,3</sup>, C. F. Dou<sup>1,3</sup>, X. H. Xu<sup>2</sup> and D. H. Liu<sup>2</sup>

<sup>1</sup>School of Computer Science, Peking University, China

<sup>2</sup>Academy of Military Sciences, China

<sup>3</sup>Key Laboratory of High Confidence Software Technologies (PKU), Ministry of Education, China

## Abstract

Scene graph generation refers to the task of identifying the objects and specifically the relationships between the objects from an image. Existing scene graph generation methods generally use the bounding boxes region features of objects to identify the relationships between objects. However, we feel that the overlap region features of two objects may play an important role in fine-grained relationship identification. In fact, some fine-grained relationships can only be obtained from the overlap region features of two objects. Therefore, we propose the Multi-Branch Feature Combination (MFC) module and Overlap Region Transformer (ORT) module to comprehensively obtain the visual features contained in the overlap regions of two objects. Concretely, the MFC module uses deconvolution and multi-branch dilation convolution to obtain high-pixels and multi-receptive field features in the overlap regions. The ORT module uses the vision transformer to obtain the self-attention of the overlap regions. The joint use of these two modules achieves the mutual complementation of local connectivity properties of convolution and the global connectivity properties of attention. We also design a Geometrical Center Augmented (GCA) module to obtain the relative position information of the geometric centers between two objects, to prevent the problem that only relying on the scale of the overlap region cannot accurately capture the relationship between two objects. Experiments show that our model ORGC (Overlap Region and Geometrical Center), the combination of the MFC module, the ORT module, and the GCA module, can enhance the performance of fine-grained relation identification. On the Visual Genome dataset, our model outperforms the current state-of-the-art model by 4.4% on the R@50 evaluation metric, reaching a state-of-the-art result of 33.88.

## CCS Concepts

• **Computing methodologies** → Artificial Intelligence; Neural Networks; Computer Vision;

## 1. Introduction

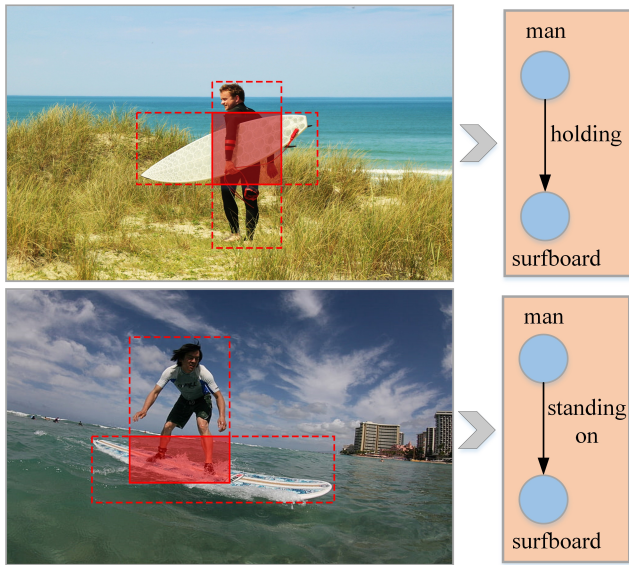
The goal of computer vision (CV) is to build intelligent systems which can extract valuable information from digital images, videos, or other modalities as humans do [CWJ\*21, CJXL21, RPG\*21], and scene graph generation is an essential part of it. Scene graph generation refers to the task of identifying the objects and specifically the relationships between these objects from an image. It contains several sub-tasks [YK21, HCO\*20, GDF\*21], such as object detection, attributes detection, and relationship identification.

Scene graph generation has been the focus of research due to the important role of scene graphs in image understanding, i.e., as a bridge for concept modeling from an image. Many substantial progresses [SMS\*21, ZKC20] have been made in recent years. The existent scene graph generation models identify the relationship between objects based on the visual information by visual feature extraction. Most of the existing methods [CRX\*21, LZL22,

TZW\*19] struggle for better feature extraction networks. Some works [YLL\*18, ZYTC18] focused on improving context aggregation modules that facilitate learning better representations, thereby improving performance. There are also works [LDZT20, ZSE\*19] that overcome shortcomings of training using existing loss and propose hand-crafted loss formulations to improve the performance.

The existing scene graph generation models [GZL\*19, TNH\*20, ZYL\*22] usually only use the respective regional features of objects to determine the relationship while ignoring the consideration of the overlap region of the two objects. However, the overlap region of two objects plays a crucial role in determining the relationship between the objects, especially in the fine-grained identification of the relationship between objects. In many cases, the fine-grained relationships are usually implied in the features of overlap regions. As shown in Figure 1, after detecting two objects, a person and a surfboard, the relation, i.e., holding or standing on, between the person and the surfboard is implied in the overlap region of these two objects, through the features of the contact region between the person (by hand or leg) and the surfboard.

† Corresponding Author



**Figure 1:** Fine-grained relationship implied by overlap region

Our ORGC (Overlap Region and Geometrical Center) model proposes to combine the high-pixel, multi-receptive field and self-attention information of the overlap region, as well as the relative position information of the geometric center between the object bounding boxes, and the original joint information of the object. First, in order to better obtain the visual features of the overlap region, we design the Multi-Branch Feature Combination (MFC) module and the Overlap Region Transformer (ORT) module. Considering that it will bring problems to the feature extraction of the overlap region when the overlap region is too large or there is no overlap region at all, we design a Geometrical Center Augmented (GCA) module to avoid this problem.

More concretely, the MFC module uses deconvolution and multi-branch dilation convolution to process the feature mapping of the overlap region to obtain more spatial structure information and receptive field information, so that the relationship identification pays more attention to the actual visual content of the overlap region. Second, inspired by the success of Transformer in NLP [KOH21, TBL\*19, WDS\*20] and CV tasks [HWC\*22, LWH\*21, ZJJ\*21], we apply vision transformer to the scene graph generation tasks for obtaining the self-attention of the overlap region. The MFC module and the ORT module are used in parallel to realize the local connection characteristics of convolution and the global connection characteristics of attention. Then, the GCA module integrates the relative geometric relationship between bounding boxes' center coordinates and bounding boxes' sizes to improve the performance of inter-object relationship identification. Experiments show that our ORGC model achieves the performance of 33.88 on the Visual Genome dataset in the R@50 evaluation metric, which is 4.4% higher than the current best methods.

In summary, the contributions of this paper are reflected in the following innovative techniques:

- We show that the overlap region and geometric center of two ob-

jects play an essential role in fine-grained relationship identification in the scene graph generation and propose a multi-attention fusion architecture ORGC for the scene graph generation.

- The multi-branch feature combination module is beneficial to obtain more spatial structure information and receptive field information. The overlap region transformer module is good at obtaining the self-attention. The combination of the two modules can achieve the mutual complementation of the local connection characteristics of convolution and the global connection characteristics of attention and make full use of the overlap region.
- The geometrical center augmented module can solve the problem that the relative position between objects cannot be accurately judged when there is a large overlap or no overlap of the bounding boxes by enhancing the geometric center relative position between considering objects.

## 2. Related Work

With the continuous development of computer vision technology, people are no longer satisfied with simply detecting and recognizing objects in images, but expect higher-level understanding and inference of images [CRX\*21, LZL22]. The scene graph has attracted the attention of many researchers as it is a bridge connecting computer vision and natural language processing, and brings a potential revolution to downstream visual inference tasks. Scene graph generation (SGG) refers to the task of automatically mapping an image into a semantic graph structure, which requires the correct labeling of detected objects and their relationships [LZWH21, CYCL19].

Currently, the challenges in SGG mainly include the following four aspects [GGS\*21, SZX\*21]: First, the semantics of predicate are not mutually exclusive, and the semantic overlap is obvious, such as "on" and "walking on", "standing on", etc. Second, the naming of relationships has a serious long-tailed distribution, that is, words that can more accurately describe the relationship between objects may be difficult to obtain. The long-tailed distribution problem is very obvious in some commonly used datasets [KZG\*17, LKBFF16]. Third, the image annotations of the datasets are too sparse, with many groups of objects that appear to be related, but are not labeled. Fourth, in addition to the existing problems of long-tailed distribution and sparse annotation, the datasets have many repeated and unreasonable annotations, and the bounding boxes of objects are inaccurate.

For addressing these issues, there are many studies [LYMB21, LRC\*21, KSS21] on scene graph generation. From a perceptual point of view, researchers focus on how to get the scene graph to represent the main content of an image. For example, Wang et al. [WWSC20] considered that Visual Genome is a dense annotation, COCO caption describes the main content of the image. They considered those visual relationships mentioned by COCO captions to be "key relationships" and constructed a new dataset for the scene graph generation. Yu et al. [YWR\*20] found the noun reference in the COCO caption in the image by artificial methods and marked the relationship.

Researchers also consider introducing semantic associations between the names of the relationships. For example, PCPL [YSJ\*20] learns a vector representation for each predicate (i.e. the name of

the relationship) and use the semantic distance between predicates to decide the similarity. CogTree [YCW\*21a] considers not only the gap between the output predicate and the labeled predicate when classifying, but also the matching degree between the output predicate and the labeled predicate ancestor node.

In addition, there are some studies to assist the generation of scene graphs by introducing commonsense knowledge. For example, GB-NET [ZKC20] re-formulates scene graph generation as the reasoning that builds a bridge between the scene and commonsense graph, where each object or predicate in the scene graph must be linked to its corresponding object or predicate class in the commonsense graph. One-shot SGG [GSGS20] introduces relational knowledge and commonsense knowledge for scene graph generation, in which, relational knowledge is prior knowledge of relationships between entities extracted from visual content, and commonsense knowledge encodes "sense-making" knowledge.

In summary, among the existing models, some determine the semantics of the not-mutually exclusive relationships by introducing semantic associations between the relation names. Some focus on the problem of sparse and unreasonable dataset annotation. Some deal with the long-tail distribution of the naming of the relationship using commonsense knowledge. However, we find a new research challenge, i.e., fine-grained scene graph generation. The core of fine-grained scene graph generation is fine-grained relationship identification, that is, to more accurately and fine-grained identify the relationship between two entities, such as whether the relationship between entities is "standing on" or "walking on," rather than a coarse-grained "on." Existing models generally ignore the fine-grained features of overlap regions as well as the relative geometric central relations between objects. Thus they do not perform well in fine-grained relationship identification. This is the main focus of this paper.

### 3. Preliminary Knowledge

This section describes the background knowledge of our approach.

**Deconvolution:** Deconvolution is known as transposed convolution, but not the reverse operation of convolution [YYT\*21]. Deconvolution can be used to solve the problem that the feature map becomes smaller after a series of previous convolution operations [MB21, APS19]. It can be used to convert a small feature map into a large feature map and obtain the relative position relationship information in the feature map.

**Vision Receptive Field:** Vision receptive field is a biological concept [SIVA17]. It refers to a specific region in the retina. When a particular region of the retina is stimulated, it can activate the activity of nerve cells in all layers of the visual system connected to this region. This region in the retina is the receptive field of these nerve cells. The mechanism of the receptive field in the human visual system can be used to enhance the feature representation of computer vision [LHW18].

**Dilation Convolution:** Dilation convolution is known as atrous convolution [MRC\*18]. Unlike regular convolutions, dilation convolution introduces a hyper-parameter called the "dilation rate," which defines the spacing of values when the kernel processes the

data [YZZ\*21]. Since the adjacent pixels of the image are almost the same, if all of them participate in the convolution operation, the result will be redundant, so the dilation convolution chooses to skip  $H$  pixel values and takes a valid value, for reducing the amount of operation on the premise of increasing the receptive field [LCWZ19].

**Vision Transformer:** The standard Transformer [VSP\*17] receives as input a 1D sequence of token embeddings. To handle 2D images, Vision Transformer reshape the image into a sequence of flattened 2D patches [BMYX21, THX\*21]. Specifically, the Vision Transformer [HXW\*21, YCW\*21b, HWC\*22] first splits the image into several patches evenly, where each patch can be regarded as a word in natural language processing tasks. Next, all patches will be flattened into sequences and fed into the encoder of the original Transformer model. Finally, the images are classified through a fully connected layer. Compared with other networks, such as traditional convolutional neural networks and recurrent neural networks, vision transformers have become a hot topic in computer vision due to their superior performance and great potential. For example, Google proposed ViT [DBK\*20] architecture, which first sliced the image and then directly used the image patch sequence to transformer architecture, completing the image classification task.

## 4. Method

In this section, we first introduce the proposed architecture and then details the main components.

### 4.1. The Architecture

The proposed architecture, ORGC (Overlap Region and Geometrical Center), is illustrated in Figure 2. It contains three pipelines corresponding to the three contents input to the "Fusion Function," starting with the "Feature Map." In the first pipeline, the image uses the object detection Faster R-CNN (including convolution backbone and RPN, etc.) [RHGS15] to obtain a set of bounding boxes  $B = \{b_i | i = 1, 2, \dots, n\}$ , ROI (Region of Interest [HGDG17]) region features  $R = \{r_i | i = 1, 2, \dots, n\}$  and original object labels  $L = \{l_i | i = 1, 2, \dots, n\}$ . Next, a bidirectional LSTM (Bi-LSTM) [GS05] is used to encode visual contexts for each object:

$$Input : \{(b_i, r_i, l_i)\} \rightarrow Output : \{x_i\}. \quad (1)$$

Then, the pairwise object feature  $X$  takes value from  $\{x_p = (x_i, x_j) | i \neq j, i, j = 1, 2, \dots, n\}$  and merges into a joint representation:

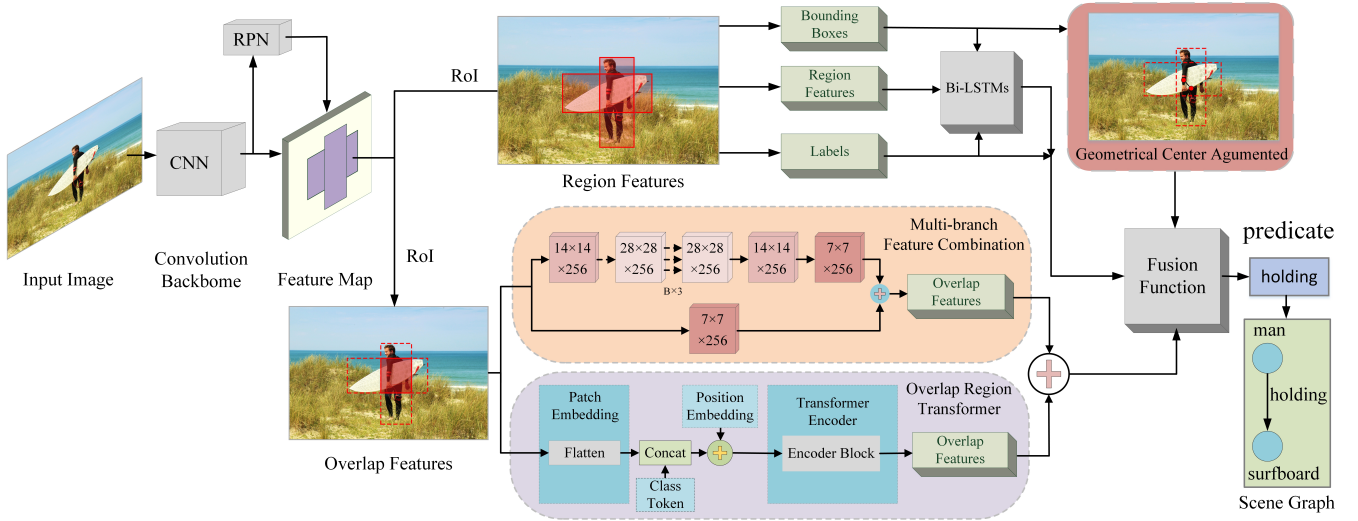
$$Input : \{x_p = (x_i, x_j)\} \rightarrow Output : \{x'_p\}. \quad (2)$$

In addition, we utilize LSTM [HS97] as decoders to fine-tune the label of each object from the corresponding  $x_i$ :

$$Input : \{x_i; l_i\} \rightarrow Output : \{l'_i\}. \quad (3)$$

The input of each LSTM cell is the concatenation of features and the previous label. We can get a pair of one-hot vectors for the object labels  $l_p = \{l'_i, l'_j\}$ , and the one-hot vectors are calculated through a joint embedding layer:

$$l'_p = W_l [l'_i \otimes l'_j], \quad (4)$$



**Figure 2:** ORGC architecture. The input images first use the Faster R-CNN to obtain the bounding box, region feature, and label of each object, then encodes them using Bi-LSTMs, and input the corresponding results into the fusion function. The overlap features are processed by the multi-branch feature combination module and the overlap region transformer module. Then, the processed overlap features are added and input into the fusion function. At the same time, the bounding boxes are subjected to geometrical center augmented module processing and input into the fusion function to identify the predicate jointly.

where  $W_i$  is learnable weight,  $\otimes$  generates the one-hot unique vector  $R^N \times N$  for the pair of  $N$ -way object labels.

In the second pipeline, we use a multi-branch feature combination module and the overlap region transformer module to extract the fine-grained overlap region features  $v_{overlap}$  of the two objects to better obtain the visual feature information of the overlap region. The details of these two modules are explained in Section 4.3 and Section 4.4, respectively.

In the third pipeline, we use the geometrical center augmented module to extract the relative geometric center feature  $R_p$  of the bounding boxes between the two objects. The details of the module are explained in Section 4.5.

Finally, the predicate  $Y$  that takes inputs from the three pipelines is then generated by using a fusion function:

$$SUM : y = W_x x'_p + l'_p + W_o v'_{overlap} + R'_p, \quad (5)$$

where  $W_x$  and  $W_o$  are trainable fusion weights of the corresponding pipeline. All models are trained by using the cross-entropy losses of object labels and predicate labels.

In the following sub-sections, we detail each of the components.

## 4.2. Overlap Region Calculation

When there is an overlap region between the two objects, the coordinate of the bounding box of the overlap region as follows:

$$(x_{overlap1}, y_{overlap1}, x_{overlap2}, y_{overlap2}) = (\max(h(x_1), t(x_1)), \max(h(y_1), t(y_1)), \min(h(x_2), t(x_2)), \min(h(y_2), t(y_2))), \quad (6)$$

$$bbox_{overlap} = (x_{overlap1}, y_{overlap1}, x_{overlap2}, y_{overlap2}), \quad (7)$$

where  $(x_{overlap1}, y_{overlap1})$  and  $(x_{overlap2}, y_{overlap2})$  represent the coordinates of the upper left corner and lower right corner of the overlap region bounding box  $bbox_{overlap}$ ,  $h$  and  $t$  represent the bounding boxes corresponding to the two objects,  $(x_1, y_1)$  and  $(x_2, y_2)$  represent the coordinates of the upper left corner and the lower right corner of the bounding boxes of the two objects.

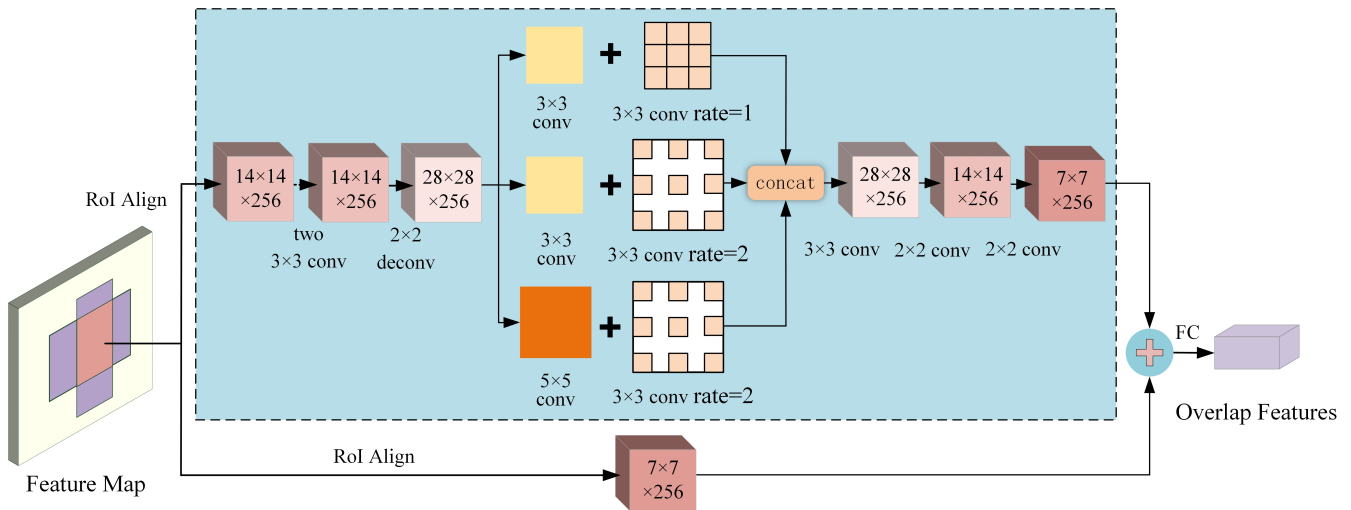
RoI Align converts the overlap region into a fixed-size feature map ( $14 \times 14 \times 256$ ), and then the multi-branch feature combination module and the overlap region transformer module process the overlap region feature map to obtain a more fine-grained relationship identification. If there is no overlap region between the two bounding boxes, the above processing is skipped. More details on feature map size selection can be found in Appendix A.1 of the additional material.

## 4.3. Multi-Branch Feature Combination Module

The existent scene graph generation models identify the relationship between objects based on the visual information input by visual feature extraction programs. However, traditional object detections [RHGS15, ZHCY21] have performed multiple convolution and pooling operations before obtaining the features of ROI (Region of Interest) Align [HGDG17], which significantly reduces the spatial structure information contained in the feature map and limits the receptive field of the feature map.

The multi-branch feature combination (MFC) module uses deconvolution and multi-branch dilation convolution to process the feature map of the overlap region. This module deeply mines the visual information in the overlap region so that the relationship identification pays more attention to the actual visual content of





**Figure 3:** Detailed illustration of the multi-branch feature combination module. The red region in the feature map indicates the overlap region of the two objects.  $14 \times 14 \times 256$ ,  $28 \times 28 \times 256$  and  $7 \times 7 \times 256$  are all represent the size of the features maps.  $2 \times 2$ ,  $3 \times 3$ , and  $5 \times 5$  represent the kernel sizes of different convolution, deconvolution, and dilation convolution operations, respectively. FC is a fully connected layer with activation function.

the overlap region, thereby improving the fine-grained relationship identification ability.

More concretely, the overlap region of the two objects is first transformed into a  $14 \times 14 \times 256$  feature map through RoI Align. Then, the MFC module processes the feature map using convolution, deconvolution, and multi-branch dilation convolution. Next, the feature maps of three different dilation convolution branches are concatenated to obtain an  $28 \times 28 \times 768$  feature map. Finally, a series of convolutions are used to convert the feature map from  $28 \times 28 \times 768$  to  $7 \times 7 \times 256$  and combined with the  $7 \times 7 \times 256$  feature map obtained by the original overlap region through RoI Align, as shown in Figure 3. In this way, the spatial structure information captured by the image and the receptive field contained in the feature map can be significantly improved, thereby improving the fine-grained relationship identification ability in the object overlap region. We provide more details about the MFC module in Appendix A.2 of the additional material.

#### 4.4. Overlap Region Transformer Module

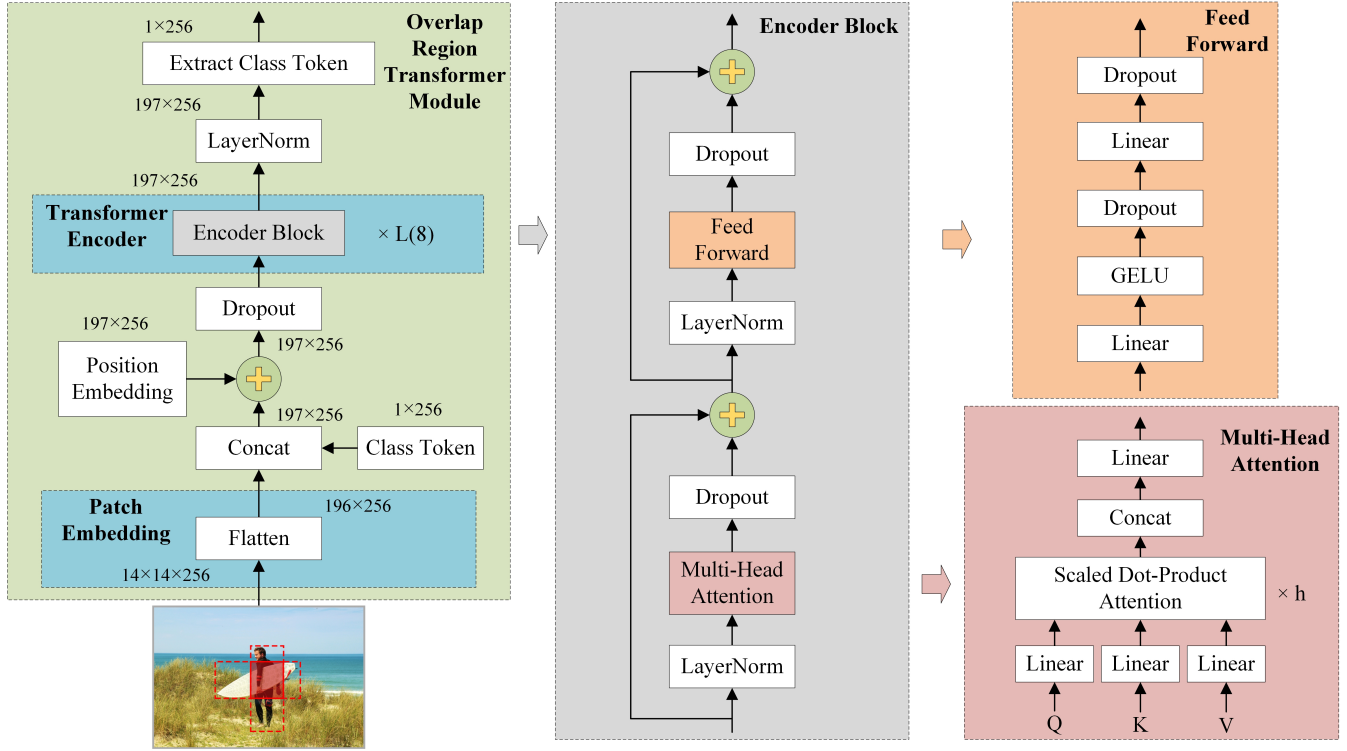
The MFC module uses operations such as convolution, deconvolution, and dilation convolution to obtain the local visual features of the overlap region while ignoring the global visual information of the overlap region. Therefore, we propose the overlap region transformer (ORT) module to use a vision transformer to obtain the global visual features of the overlap regions of two objects. The parallel use of the ORT module and the MFC module realizes the mutual complementation of the global connection characteristics of attention and the local connection characteristics of convolution, which can better improve the performance of fine-grained relationship identification in the process of scene graph generation.

Specifically, RoI Align first converts the overlap region into a

fixed-size feature map ( $14 \times 14 \times 256$ ) and then uses the ORT module to obtain the self-attention of the overlap region. The feature extraction structure in the ORT module is similar to the feature extraction structure in Transformer. As shown in Figure 4, the  $14 \times 14 \times 256$  overlap region features are first flattened into  $196 \times 256$ . Then a class token representing the global visual features of the overlap region is added, and the position information corresponding to each token is added. And then, the features are input to the encoder block to obtain the global visual features of the overlap region. Finally, obtain the global visual features of the overlap region, which is the input to identify the fine-grained relationship between objects. More details of the size of each feature map are provided in Appendix A.3 of the additional material.

Each encoder block includes two sub-layers, as shown in Figure 4. The first sub-layer is multi-head attention, which is used to calculate the self-attention of the overlap region. The second sub-layer is a fully connected feed forward neural network, using the GELU [HG16] activation function. The normalization operation LayerNorm is applied before each sub-layer, and the residual network is applied after each sub-layer. To ensure connection, all sub-layers and embedding layers output the same dimension.

Multi-head attention can improve the performance of the general self-attention layer. Specifically, this paper sets heads  $h$  in multi-head attention to 4 and transforms the input overlap region feature into three groups of different vectors, namely query group, key group, and value group. In each group, there are  $h$  vectors with dimension  $d_{q'} = d_{k'} = d_{v'} = d_{model}/h = 64$ , where  $d_{q'}$ ,  $d_{k'}$ , and  $d_{v'}$  are the dimensions of the vectors in the three different groups, and  $d_{model}$  is the dimension of the input feature. As shown in Figure 4, different heads in multi-head attention use different query, key, and value matrices, namely,  $Q$ ,  $K$ , and  $V$ , which are randomly initial-



**Figure 4:** Detailed illustration of the overlap region transformer module. The red solid line in the image represents the overlap region of the two objects. The detailed architecture of the overlap region transformer module (in green) is shown on the left subgraph. The detailed architecture of the encoder block (in gray) is shown in the middle subgraph. The detailed architecture of the feed forward (in orange) and multi-head attention (in red) contained in the encoder block are shown in the two subgraphs on the right, respectively.

ized. The multi-head attention process is shown as follows:

$$\text{MultiHead}(Q', K', V') = \text{Concat}(\text{head}_1, \dots, \text{head}_4)W^o, \quad (8)$$

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i), \quad (9)$$

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_{\text{model}}/h}}\right)V_i. \quad (10)$$

Here,  $Q'$ ,  $K'$  and  $V'$  are the concatenation of  $\{Q_i\}_{i=1}^h$ ,  $\{K_i\}_{i=1}^h$  and  $\{V_i\}_{i=1}^h$ ,  $W^o \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$  is the projection weight. More details of Attention can be seen in Appendix A.3 of the additional material.

#### 4.5. Geometrical Center Augmented Module

In order to prevent the occurrence of the problem of missing the spatial information of the input image when the overlap region of the bounding boxes of the two objects is too large or when there is no overlap region, we propose the geometrical center augmented (GCA) module. The GCA module uses the relative geometric center position information between objects to assist the relationship identification, which improves the performance of fine-grained relationship identification between objects.

The GCA module integrates the relative geometric relationship

between the bounding boxes' center coordinates and sizes. Specifically, the two-dimensional relative coordinate position of each bounding box is  $\{(x_i^{\min}, y_i^{\min}), (x_i^{\max}, y_i^{\max})\}$ , where  $(x_i^{\min}, y_i^{\min})$  is the relative position coordinate of the upper left corner of the bounding box,  $(x_i^{\max}, y_i^{\max})$  is the lower right corner of the bounding box. Therefore, the relative geometric center coordinate  $(x_i, y_i)$  and relative width  $w_i$ , relative height  $h_i$  of the bounding box  $i$  are:

$$(x_i, y_i) = \left(\frac{x_i^{\min} + x_i^{\max}}{2}, \frac{y_i^{\min} + y_i^{\max}}{2}\right), \quad (11)$$

$$w_i = (x_i^{\max} - x_i^{\min}) + 1, h_i = (y_i^{\max} - y_i^{\min}) + 1. \quad (12)$$

The relative features of the geometric center relationship between the bounding boxes of two objects  $i$  and  $j$  are:

$$r_p = \begin{pmatrix} \log(1 + |x_i - x_j|) \\ \log(1 + |y_i - y_j|) \\ \log(1 + \frac{w_i}{w_j}) \\ \log(1 + \frac{h_i}{h_j}) \end{pmatrix}, \quad (13)$$

$$R_p = FC(r_p), \quad (14)$$

$$R'_p = \text{ReLU}(W_R^T R_p), \quad (15)$$

where  $r_p$  is the relative geometric relationship between the bounding boxes of two objects,  $R_p$  is the high-dimensional representation of  $r_p$ ,  $FC$  is the fully connected layer,  $ReLU$  is the activation function,  $W_R^T$  is the weight parameter to be learned,  $R'_p$  is the relative geometric feature of the bounding boxes.

## 5. Experiments

This section will detail our ORGC model's implementation details, experimental results, and limitations.

### 5.1. Implementation Details

**Dataset.** We utilize the Visual Genome (VG) [KZG\*17] dataset to train and evaluate our model. The VG dataset consists of 75k object categories and 37k predicate categories, with 108k images. Since 92% of the predicates in the VG dataset do not have more than ten instances, we follow the widely used VG dataset splitting method [CZX\*19] to split the dataset. The split dataset contains the most frequently occurring 150 object categories and 50 predicate categories. At the same time, the split dataset only has a training set and test set, of which the training set accounts for 70% and the test set accounts for 30%, and there is no validation set. We follow the Neural Motifs [ZYT18] method and sample 5k images from the training set as a validation set for parameter tuning.

**Evaluation Metrics.** We use two evaluation metrics, i.e., Recall@K (R@K) [MLZMG16] and mean Recall@K (mR@K) [CYCL19], to evaluate the performance of our fine-grained scene graph generation model ORGC. R@K is the earliest and most widely accepted evaluation metric in scene graph generation. Since the ground truth of the VG dataset is not complete in the annotation of the relationship, the accuracy rate cannot sufficiently reflect the effect of scene graph generation. Therefore, researchers regard the metric Recall in the retrieval field as an evaluation metric for scene graph generation, which not only requires accurate recognition but also requires a better ability to eliminate unrelated object pairs. R@K is defined as a ratio of the number of relevant results within the top-k ranked results to the total number of relevant results [PTM22]. mR@K calculates the Recall of all predicate categories separately and then calculates their mean so that all categories are equally important, which can also be a great quantitative evaluation metric of the fine-grained scene graph generation. The evaluation of our model is carried out in scene graph generation (SGG): generating scene graphs directly from images, the most realistic application scenario for scene graph generative models.

**Baseline.** The baseline model Causal-TDE [TNH\*20], trained with standard cross entropy loss, as well as our proposed ORGC architecture, are trained using an identical setup. On the VG dataset, the experimental results of the baseline model on the three evaluation metrics of R@20, R@50, and R@100 can reach 25.42, 32.45, and 37.26, which are the best results available.

**Object Detector.** Following the previous work [TZW\*19], we pretrain Faster R-CNN [RHGS15] and freeze it as the underlying object detector for our scene graph generation model. We choose ResNeXt-101-FPN as the backbone network of Faster R-CNN and scale the long side of the input image to 1k pixels. The object detector is trained on the VG dataset using one 3090, where the batch

size is set to 8, and the initial learning rate is set to  $8 \times 10^{-3}$ , which decays at 30k and 40k iterations to 1/10 of the existing learning rate. The final detector achieves a detection accuracy of 28.14 mAP (mean Average Precision) on the VG dataset.

**Scene Graph Generation.** Before our scene graph generation model ORGC training, the object detector Faster R-CNN is trained in advance, and the parameters are frozen. When the scene graph generation model is trained, the batch size is set to 12, and the initial learning rate is set to  $12 \times 10^{-3}$ . We apply linear learning rate warmup over the first 1K steps and cosine learning rate decay afterward. After the verification performance is stable, the learning rate will decay to 1/10 of the existing learning rate. All the experiments are carried out on the Ubuntu system with 128 GB RAM, a RTX 3090 (24 GB) GPU, and AMD EPYC 7H12 64-Core Processor.

### 5.2. Quantitative Studies

We compare the experimental results of various existing state-of-the-art models with our proposed ORGC model on the VG dataset, as shown in Table 1. From the results of the table, we can observe that compared with existing models, our ORGC achieves extensive improvements on all evaluation metrics (R@20, R@50, R@100, mR@20, mR@50, and mR@100) and significantly improves the performance of fine-grained scene graph generation.

More concretely, it can be found that our ORGC model achieves 26.89, 33.88, and 38.32 in the evaluation metrics R@20, R@50, and R@100, which are the best experiment results on these evaluation metrics. Our model results outperform the previous state-of-the-art model results by 5.7%, 4.4%, and 2.8% on the evaluation metrics R@20, R@50, and R@100, which fully reflects the performance improvement of our model for the scene graph generation.

When our ORGC model achieves the current best experimental results on the evaluation metrics R@20, R@50, and R@100, it is also better than the corresponding models on the three evaluation metrics of mR@20, mR@50, and mR@100. Under the same model settings, our model can achieve experimental results of 6.22, 8.38, and 9.45 on the relevant evaluation metrics, which are 22.6%, 18.0%, and 9.9% higher than the results of baseline model. This further shows that our ORGC model is effective for all categories of predicates.

### 5.3. Ablation Studies

To investigate the contribution of each component of our ORGC model, we conducted extensive ablation studies. Our ablation studies consist of various modules proposed in this paper and different combinations of multiple modules. The experimental results are shown in Table 2. In the following, we will provide a detailed analysis of the experimental results of the ablation studies. We analyze all ablation studies on the core evaluation metric R@50.

Through the ablation studies, it can be found that the multi-branch feature combination (MFC) module can improve the model's performance from 32.45 to 33.29, which is 2.6% higher than the baseline model. The overlap region transformer (ORT) module can improve the model's performance from 32.45 to 32.92, which is 1.4% higher than the baseline model. The geometrical

Model	R@20	R@50	R@100	mR@20	mR@50	mR@100
GB-NET	-	29.40	35.10	-	7.10	8.50
Neural Motifs	25.7	30.50	35.80	5.07	6.91	8.12
VCTree	-	31.80	36.10	-	6.60	7.70
KRE	24.6	30.90	35.80	-	6.40	7.30
MSDN	-	31.90	36.66	-	6.10	7.20
GPS-Net	-	31.10	35.90	-	6.70	8.60
Causal-TDE	25.42	32.45	37.26	4.36	5.83	7.08
Ours (ORGC)	<b>26.89</b>	<b>33.88</b>	<b>38.32</b>	<b>6.22</b>	<b>8.38</b>	<b>9.45</b>

**Table 1: Quantitative Results.** The table shows the performance comparison between our proposed ORGC model and various existing state-of-the-art models on six evaluation metrics (R@K (20, 50, 100) and mR@K (20, 50, 100)) for the fine-grained scene graph generation.

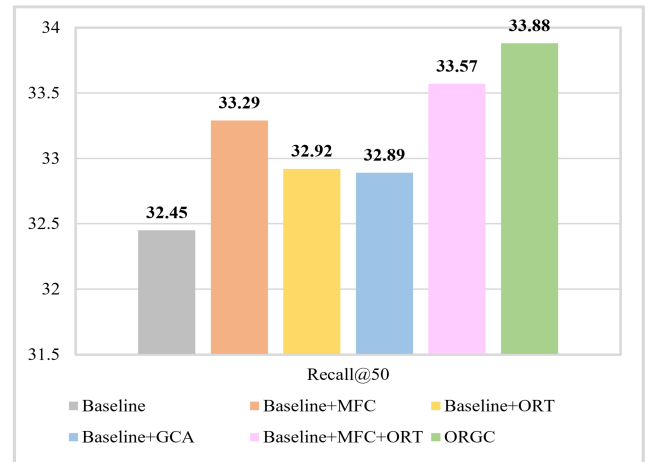
Model	R@20	R@50	R@100	mR@20	mR@50	mR@100
Baseline	25.42	32.45	37.26	4.36	5.83	7.08
Baseline+MFC	26.23	33.29	37.91	5.62	7.37	8.49
Baseline+ORT	25.91	32.92	37.62	5.18	6.92	7.96
Baseline+GCA	25.85	32.89	37.48	4.95	6.65	7.81
Baseline+MFC+ORT	26.58	33.57	38.19	6.05	8.09	9.05
ORGC (Baseline+MFC+ORT+GCA)	<b>26.89</b>	<b>33.88</b>	<b>38.32</b>	<b>6.22</b>	<b>8.38</b>	<b>9.45</b>

**Table 2: Ablation Studies.** The table shows the experimental results corresponding to various modules and combinations of modules proposed in this paper. The experimental results show that each module significantly improves fine-grained scene graph generation.

center augmented (GCA) module can improve the model's performance from 32.45 to 32.89, which is a 1.4% improvement over the baseline model. The combined use of the MFC and ORT modules can improve the model's performance from 32.45 to 33.57, which is 3.5% higher than the baseline model. The combined use of the MFC, ORT, and GCA modules (ORGC) can improve the model's performance from 32.45 to 33.88, which is 4.4% higher than the baseline model. An intuitive comparison of the relevant ablation experimental results is shown in Figure 5. The ablation studies have proved that the modules proposed in our ORGC model can significantly improve the performance of the scene graph generation.

The MFC module and the ORT module act on the overlap region of two objects. The visualization experimental result of the two modules used together is shown in Figure 6 (up). The orange box in the figure represents an example of the predicate identification result of the baseline model, and the green box represents an example of the predicate identification result after using the MFC module and the ORT module. The common use of the two modules can generate a new triple <woman, walking on, sidewalk> instead of the triple <woman, on, sidewalk> generated by the baseline model. Experiments show that using the MFC module and the ORT module for the overlap region can obtain the fine-grained visual feature information contained in the overlap region of the objects and improve the accuracy of predicate identification.

The GCA module acts on the bounding boxes of two objects, and the visualization experimental result is shown in Figure 6 (down). The orange box in the figure represents an example of the predicate identification result of the baseline model, and the green box represents an example of the predicate identification result after using the GCA module. We can find that a new triple <car, parked on, street> can be generated by using our GCA module instead of



**Figure 5: Comparison of the results of different methods (i.e., Baseline, Baseline+MFC, Baseline+ORT, Baseline+GCA, Baseline+MFC+ORT, ORGC) on the evaluation metric Recall@50.**

the triple <car, on, street> generated by the baseline model. Experiments show that the GCA module is used to obtain the relative position information of geometric centers between objects, thereby obtaining more fine-grained predicate identification results.

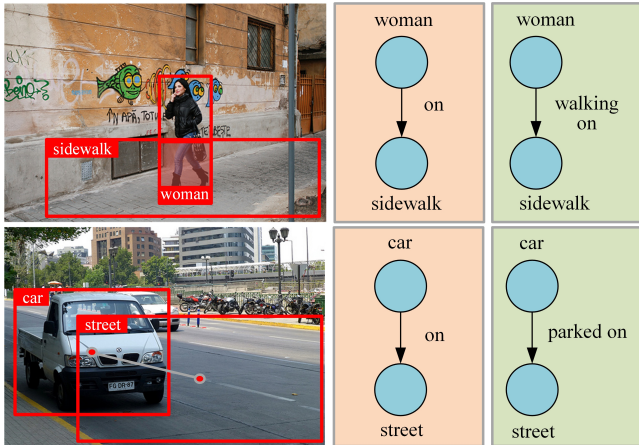
#### 5.4. Overlap Region vs Union Region

We use the same model to conduct comparative experiments on the overlap region and the union region, and the experimental results are shown in Table 3. The Overlap Region represents the overlap

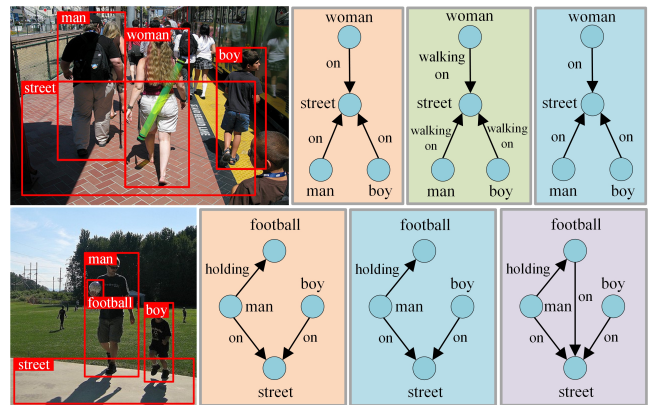


Model	R@20	R@50	R@100	mR@20	mR@50	mR@100
Baseline	25.42	32.45	37.26	4.36	5.83	7.08
Overlap Region (Baseline+MFC+ORT)	<b>26.58</b>	<b>33.57</b>	<b>38.19</b>	<b>6.05</b>	<b>8.09</b>	<b>9.05</b>
Union Region1 (Baseline+MFC+ORT)	25.71	32.86	37.33	4.88	6.57	7.73
Union Region2 (Baseline+MFC+ORT)	25.08	32.04	36.82	4.13	5.49	6.88

**Table 3:** Comparative experiments of overlap region and union region. Overlap Region represents the overlap region of two objects. Union Region1 and Union Region2 represent the union region of two objects under two different conditions. The model used for Overlap Region, Union Region1, and Union Region2 is the Baseline+MFC+ORT model proposed in this paper.



**Figure 6:** The visualization results of scene graphs generated from the Baseline model (in orange), Baseline+MFC+ORT model (in green, up), and Baseline+GCA model (in green, down).



**Figure 7:** The visualization results of scene graphs generated from the Baseline model (in orange), Overlap Region (in green), Union Region1 (in blue), and Union Region2 (in purple).

region of two objects, and the Union Region1 represents the union region of the two objects when they have an overlap region. The Union Region2 represents the union region of the two objects, including two cases where the two objects have an overlap region and do not have an overlap region. The Baseline represents the baseline model, MFC represents the multi-branch feature combination module, and ORT represents the overlap region transformer module.

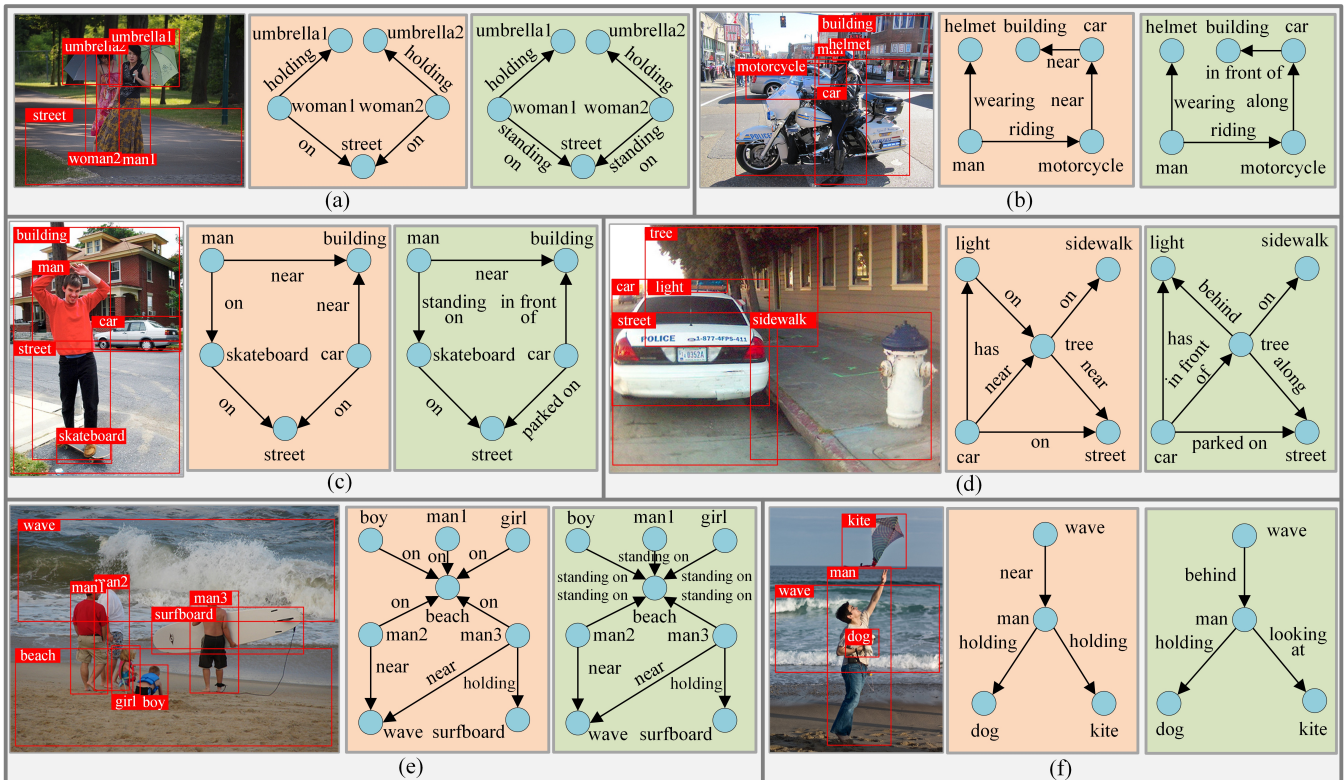
By comparing the experimental results of the Overlap Region and Union Region1, it can be found that when two objects have an overlap region, the performance improvement of the overlap region of the two objects for predicate identification is better than that of the union region, which not only proves that the overlap region is unique for predicate identification, but proves that the effect of our proposed module on the overlap region is better than that on the union region. By comparing the experimental results of Union Region1 and Union Region2, it can be found that when there is no overlap region between the two objects, too much attention to the union region of the two objects will affect the performance of model predicate identification.

The visualization experiment results of the corresponding examples are shown in Figure 7. From the first example, it can be found that, compared with the Baseline model, our Overlap Region method can fine-grained identify the relationship between person and street as "walking on," while the Union Region1 method pays

attention to the irrelevant region, resulting in coarse-grained experimental results. The second example shows that the Union Region2 method generates a new triple <football, on, street> compared to the Baseline and Union Region1 methods. In fact, the relationship between football and the street is not "on" because the two entities are not in contact. This is because the Union Region2 method pays attention to many irrelevant regions and introduces noise.

### 5.5. Qualitative Studies

We visualize several examples of scene graph generation via the baseline model and our proposed ORGC model, as shown in Figure 8. From the comparison, it can be found that the scene graph generated by our ORGC model is more discriminative than the baseline model. The baseline model often identifies some coarse-grained predicates, such as "on" and "near," while our model can identify a series of more fine-grained predicates, such as "standing on," "parked on," "in front of," etc. For example, in the first image (a), the baseline model generates the triplet <woman, on, street>, while our ORGC model can identify a fine-grained predicate "standing on" instead of the original predicate "on" and generate a new triple <woman, standing on, street>. In the second image (b), the relationship between the car and building identified by the baseline model is "near," while our model identified "in front of." The experiments show that the results identified by our ORGC model are more fine-grained and more in line with objective facts.



**Figure 8:** The visualization results of scene graph generation from the baseline model (in orange) and our ORGC model (in green).

In addition to identifying more fine-grained relationships, our ORGC model can also identify more accurate relationships in some complex scenarios. For example, in the fourth image (d), the baseline model identified the relationship between the light and the tree as "on," while our model identified "behind." The last image (f) shows that when the man is not touching the kite, the baseline model produces "holding." In contrast, our ORGC model successfully identifies "looking at." All visualization results show a clear trend that our ORGC model is more effective and accurate for fine-grained scene graph generation.

### 5.6. Limitations

Through related experiments, we find that our ORGC model can significantly improve the performance of fine-grained scene graph generation, but it also has limitations. Firstly, when there is no overlap region between the two objects, the multi-branch feature combination module and the overlap region transformer module cannot be used, and the ability to realize fine-grained relationship recognition can only be achieved through the geometrical center augmented module, and the ability of the ORGC model is limited. Secondly, when too many objects and relationships exist in the overlap region, it is easy to introduce noise and affect the recognition of the fine-grained relationship between the considering objects. We leave all these in future work.

### 6. Conclusion

In this work, we explore the important role of the overlap region and geometric center of the bounding boxes between two objects for the fine-grained scene graph generation task. We use the multi-branch feature combination module to obtain the high-pixel and multi-receptive field information of the overlap region and use the overlap region transformer module to obtain the self-attention of the overlap region. Combining the two modules enables the mutual complementation of the local connectivity properties of convolution and the global connectivity properties of attention. At the same time, we design the geometrical center augmented module to obtain the relative position information of the geometric centers between objects to prevent the problem that the relative relationship between objects cannot be accurately judged when the proportion of the overlap region is too large or there is no overlap region. Experiments demonstrate that our ORGC model can significantly enhance the ability of fine-grained relationship identification during scene graph generation. Finally, the methods proposed in this paper do not make any assumptions about the underlying generation model and can be easily used with any model.

While these results are encouraging, many challenges remain. One is to improve the object detector's performance in scene graph generation, which can refer to the methods [CXD\*21, FFF\*21] in object detection, such as new network structure, feature fusion, and training method. Another challenge is to build a large dataset with fine-grained labels and accurate annotations [SZDL20], which is

necessary and significant. It needs to contain as many scenes as possible, preferably constructed by computer vision experts. Finally, applying the fine-grained scene graph generation model to other computer vision and multi-media tasks [WS22, DJX\*21], such as content-based image search, image captioning, visual question answering, and multi-modal knowledge graph construction. The fine-grained scene graph generation model can provide a better representation for these scene understanding-related tasks and can significantly improve the model performance of these tasks.

## 7. Acknowledgements

Our work is supported by the National Key Research and Development Program of China (Project Number: 2020AAA0109400).

## References

- [APS19] ALJADAANY R., PAL D. K., SAVVIDES M.: Douglas-rachford networks: Learning both the image prior and data fidelity terms for blind image deconvolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10235–10244. 3
- [BMYX21] BAI Y., MEI J., YUILLE A. L., XIE C.: Are transformers more robust than cnns? *Advances in Neural Information Processing Systems* 34 (2021). 3
- [CJXL21] CHEN L., JIANG Z., XIAO J., LIU W.: Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 16846–16856. 1
- [CRX\*21] CHANG X., REN P., XU P., LI Z., CHEN X., HAUPTMANN A. G.: A comprehensive survey of scene graphs: Generation and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021). 1, 2
- [CWJ\*21] CAO Y., WANG J., JIN Y., WU T., CHEN K., LIU Z., LIN D.: Few-shot object detection via association and discrimination. *Advances in Neural Information Processing Systems* 34 (2021). 1
- [CXD\*21] CHEN X., XU C., DONG M., XU C., WANG Y.: An empirical study of adder neural networks for object detection. *Advances in Neural Information Processing Systems* 34 (2021). 10
- [CYCL19] CHEN T., YU W., CHEN R., LIN L.: Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 6163–6171. 2, 7
- [CZX\*19] CHEN L., ZHANG H., XIAO J., HE X., PU S., CHANG S.-F.: Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 4613–4623. 7
- [DBK\*20] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBERN D., ZHAI X., UNTERTHINER T., DEGHANI M., MINDERER M., HEIGOLD G., GELLY S., ET AL.: An image is worth 16×16 words: Transformers for image recognition at scale. 3
- [DJX\*21] DENG C., JIA Y., XU H., ZHANG C., TANG J., FU L., ZHANG W., ZHANG H., WANG X., ZHOU C.: Gakg: A multimodal geoscience academic knowledge graph. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (2021), pp. 4445–4454. 11
- [FFF\*21] FAN Q., FAN D.-P., FU H., TANG C.-K., SHAO L., TAI Y.-W.: Group collaborative learning for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 12288–12298. 10
- [GDF\*21] GARG S., DHAMO H., FARSHAD A., MUSATIAN S., NAVAB N., TOMBARI F.: Unconditional scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 16362–16371. 1
- [GGS\*21] GUO Y., GAO L., SONG J., WANG P., SEBE N., SHEN H. T., LI X.: Relation regularized scene graph generation. *IEEE Transactions on Cybernetics* (2021). 2
- [GS05] GRAVES A., SCHMIDHUBER J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks* 18, 5-6 (2005), 602–610. 3
- [GSGS20] GUO Y., SONG J., GAO L., SHEN H. T.: One-shot scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 3090–3098. 3
- [GZL\*19] GU J., ZHAO H., LIN Z., LI S., CAI J., LING M.: Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 1969–1978. 1
- [HCO\*20] HUANG Y., CHEN J., OUYANG W., WAN W., XUE Y.: Image captioning with end-to-end attribute detection and subsequent attributes prediction. *IEEE Transactions on Image processing* 29 (2020), 4013–4026. 1
- [HG16] HENDRYCKS D., GIMPEL K.: Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016). 5
- [HGDG17] HE K., GKIOXARI G., DOLLAR P., GIRSHICK R.: Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2961–2969. 3, 4
- [HS97] HOCHREITER S., SCHMIDHUBER J.: Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780. 3
- [HWC\*22] HAN K., WANG Y., CHEN H., CHEN X., GUO J., LIU Z., TANG Y., XIAO A., XU C., XU Y., ET AL.: A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). 2, 3
- [HXW\*21] HAN K., XIAO A., WU E., GUO J., XU C., WANG Y.: Transformer in transformer. *Advances in Neural Information Processing Systems* 34 (2021). 3
- [KOH21] KIM J., OH S., HONG S.: Transformers generalize deepsets and can be extended to graphs & hypergraphs. *Advances in Neural Information Processing Systems* 34 (2021). 2
- [KSS21] KHANDELWAL S., SUHAIL M., SIGAL L.: Segmentation-grounded scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 15879–15889. 2
- [KZG\*17] KRISHNA R., ZHU Y., GROTH O., JOHNSON J., HATA K., KRAVITZ J., CHEN S., KALANTIDIS Y., LI L.-J., SHAMMA D. A., ET AL.: Visual genome: Connecting language and vision using crowd-sourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73. 2, 7
- [LCWZ19] LI Y., CHEN Y., WANG N., ZHANG Z.: Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 6054–6063. 3
- [LDZT20] LIN X., DING C., ZENG J., TAO D.: Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3746–3753. 1
- [LHW18] LIU S., HUANG D., WANG A.: Receptive field block net for accurate and fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)* (September 2018). 3
- [LKBFF16] LU C., KRISHNA R., BERNSTEIN M., FEI-FEI L.: Visual relationship detection with language priors. In *European conference on computer vision* (2016), Springer, pp. 852–869. 2
- [LRC\*21] LU Y., RAI H., CHANG J., KNYAZEV B., YU G., SHEKHAR S., TAYLOR G. W., VOLKOV M.: Context-aware scene graph generation with seq2seq transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 15931–15941. 2
- [LWH\*21] LIU Z., WANG Y., HAN K., ZHANG W., MA S., GAO W.: Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems* 34 (2021). 2



- [LYMB21] LIU H., YAN N., MORTAZAVI M., BHANU B.: Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 11546–11556. 2
- [LZL22] LIN B., ZHU Y., LIANG X.: Atom correlation based graph propagation for scene graph generation. *Pattern Recognition* 122 (2022), 108300. 1, 2
- [LZWH21] LI R., ZHANG S., WAN B., HE X.: Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 11109–11119. 2
- [MB21] MAKARKIN M., BRATASHOV D.: State-of-the-art approaches for image deconvolution problems, including modern deep learning architectures. *Micromachines* 12, 12 (2021), 1558. 3
- [MLZMG16] MISRA I., LAWRENCE ZITNICK C., MITCHELL M., GIRSHICK R.: Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2930–2939. 7
- [MRC\*18] MEHTA S., RASTEGARI M., CASPI A., SHAPIRO L., HAJISHIRZI H.: Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *Proceedings of the european conference on computer vision (ECCV)* (2018), pp. 552–568. 3
- [PTM22] PATEL Y., TOLIAS G., MATAS J.: Recall@k surrogate loss with large batches and similarity mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 7502–7511. 7
- [RHGS15] REN S., HE K., GIRSHICK R., SUN J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015). 3, 4, 7
- [RPG\*21] RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M., SUTSKEVER I.: Zero-shot text-to-image generation. In *International Conference on Machine Learning* (2021), PMLR, pp. 8821–8831. 1
- [SIVA17] SZEGEDY C., IOFFE S., VANHOUCHE V., ALEMI A. A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence* (2017). 3
- [SMS\*21] SUHAIL M., MITTAL A., SIDDIQUIE B., BROADDUS C., ELEDATH J., MEDIONI G., SIGAL L.: Energy-based learning for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 13936–13945. 1
- [SZDL20] SHAO D., ZHAO Y., DAI B., LIN D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 2616–2625. 10
- [SZX\*21] SHI J., ZHONG Y., XU N., LI Y., XU C.: A simple baseline for weakly-supervised scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 16393–16402. 2
- [TBL\*19] TSAI Y.-H. H., BAI S., LIANG P. P., KOLTER J. Z., MORENCY L.-P., SALAKHUTDINOV R.: Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (2019), vol. 2019, NIH Public Access, p. 6558. 2
- [THX\*21] TANG Y., HAN K., XU C., XIAO A., DENG Y., XU C., WANG Y.: Augmented shortcuts for vision transformers. *Advances in Neural Information Processing Systems* 34 (2021). 3
- [TNH\*20] TANG K., NIU Y., HUANG J., SHI J., ZHANG H.: Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 3716–3725. 1, 7
- [TZW\*19] TANG K., ZHANG H., WU B., LUO W., LIU W.: Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 6619–6628. 1, 7
- [VSP\*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. *Advances in neural information processing systems* 30 (2017). 3
- [WDS\*20] WOLFF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., ET AL.: Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (2020), pp. 38–45. 2
- [WS22] WAGHMARE P. M., SHINDE S. V.: Image caption generation using neural network models and lstm hierarchical structure. In *Computational Intelligence in Pattern Recognition*. Springer, 2022, pp. 109–117. 11
- [WWSC20] WANG W., WANG R., SHAN S., CHEN X.: Sketching image gist: Human-mimetic hierarchical scene graph generation. In *European Conference on Computer Vision* (2020), Springer, pp. 222–239. 2
- [YCW\*21a] YU J., CHAI Y., WANG Y., HU Y., WU Q.: Cogtree: Cognition tree loss for unbiased scene graph generation. In *International Joint Conference on Artificial Intelligence* (2021). 3
- [YCW\*21b] YUAN L., CHEN Y., WANG T., YU W., SHI Y., JIANG Z.-H., TAY F. E., FENG J., YAN S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 558–567. 3
- [YK21] YE K., KOVASHKA A.: Linguistic structures as weak supervision for visual scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 8289–8299. 1
- [YLL\*18] YANG J., LU J., LEE S., BATRA D., PARIKH D.: Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 670–685. 1
- [YSJ\*20] YAN S., SHEN C., JIN Z., HUANG J., JIANG R., CHEN Y., HUA X.-S.: Pcp: Predicate-correlation perception learning for unbiased scene graph generation. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 265–273. 2
- [YWR\*20] YU F., WANG H., REN T., TANG J., WU G.: Visual relation of interest detection. In *ACM International Conference on Multimedia (MM'20)* (2020), pp. 1386–1394. 2
- [YYT\*21] YOU H., YU L., TIAN S., MA X., XING Y., XIN N., CAI W.: Mc-net: Multiple max-pooling integration module and cross multi-scale deconvolution network. *Knowledge-Based Systems* 231 (2021), 107456. 3
- [YZZ\*21] YAN Z., ZHANG R., ZHANG H., ZHANG Q., ZUO W.: Crowd counting via perspective-guided fractional-dilation convolution. *IEEE Transactions on Multimedia* (2021). 3
- [ZHCY21] ZHANG D., HAN J., CHENG G., YANG M.-H.: Weakly supervised object localization and detection: a survey. *IEEE transactions on pattern analysis and machine intelligence* (2021). 4
- [ZJJ\*21] ZHAO H., JIANG L., JIA J., TORR P. H., KOLTUN V.: Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 16259–16268. 2
- [ZKC20] ZAREIAN A., KARAMAN S., CHANG S.-F.: Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision* (2020), Springer, pp. 606–623. 1, 3
- [ZSE\*19] ZHANG J., SHIH K. J., ELGAMMAL A., TAO A., CATANZARO B.: Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11535–11543. 1
- [ZYL\*22] ZHOU H., YANG Y., LUO T., ZHANG J., LI S.: A unified deep sparse graph attention network for scene graph generation. *Pattern Recognition* 123 (2022), 108367. 1
- [ZYT18] ZELLERS R., YATSKAR M., THOMSON S., CHOI Y.: Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 5831–5840. 1, 7