Visual Relation of Interest Detection

Fan Yu[1,3], Haonan Wang[1], Tongwei Ren[1,3,*]
Jinhui Tang[2], Gangshan Wu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2]School of Computer Science, Nanjing University of Science and Technology, Nanjing, China
[3]Shenzhen Research Institute of Nanjing University, Shenzhen, China
yf@smail.nju.edu.cn,haonan.wang@smail.nju.edu.cn,rentw@nju.edu.cn,jinhuitang@njust.edu.cn,gswu@nju.edu.cn

ABSTRACT

In this paper, we propose a novel Visual Relation of Interest Detection (VROID) task, which aims to detect visual relations that are important for conveying the main content of an image, motivated from the intuition that not all correctly detected relations are really "interesting" in semantics and only a fraction of them really make sense for representing the image main content. Such relations are named Visual Relations of Interest (VROIs). VROID can be deemed as an evolution over the traditional Visual Relation Detection (VRD) task that tries to discover all visual relations in an image. We construct a new dataset to facilitate research on this new task, named ViROI, which contains 30,120 images each with VROIs annotated. Furthermore, we develop an Interest Propagation Network (IPNet) to solve VROID. IPNet contains a Panoptic Object Detection (POD) module, a Pair Interest Prediction (PaIP) module and a Predicate Interest Prediction (PrIP) module. The POD module extracts instances from the input image and also generates corresponding instance features and union features. The PaIP module then predicts the interest score of each instance pair while the PrIP module predicts that of each predicate for each instance pair. Then the interest scores of instance pairs are combined with those of the corresponding predicates as the final interest scores. All VROI candidates are sorted by final interest scores and the highest ones are taken as final results. We conduct extensive experiments to test effectiveness of our method, and the results show that IPNet achieves the best performance compared with the baselines on visual relation detection, scene graph generation and image captioning.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Visual relation of interest; visual relation detection; interest propagation network; interest estimation

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413566

ACM Reference Format:
Fan Yu[1,3], Haonan Wang[1], Tongwei Ren[1,3,] and Jinhui Tang[2], Gangshan Wu[1]. 2020. Visual Relation of Interest Detection. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413566

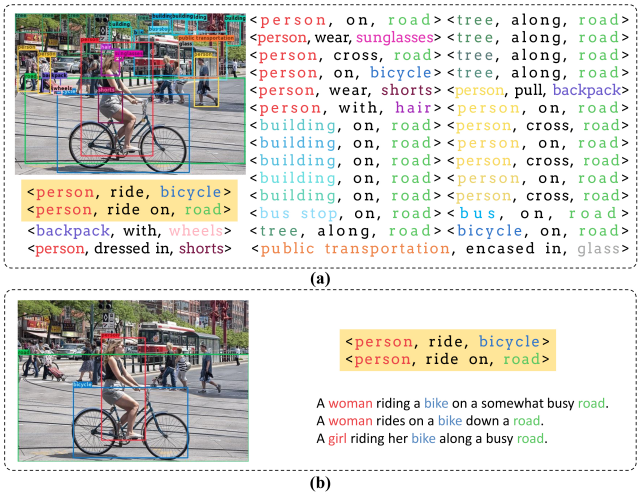

Figure 1: Comparison between visual relation detection and visual relation of interest detection. (a) Visual relation detection on an image from VisualGenome dataset [13]. (b) Visual relation of interest detection on the same image from our new ViROI dataset. (c) Image captioning on the same image from MSCOCO dataset [18].

1 INTRODUCTION

As a bridge between vision and natural language, visual relation detection (VRD) aims to describe the instances in an image and their interactions with relationship triplets in the form of <subject, predicate, object> [19]. Due to the explosive combination possibility of subject, predicate and object, there actually exist abundant visual relations that on one hand provide a comprehensive description of the image content, and on the other hand may mislead the prediction of the main content with an overwhelming amount of detail. As shown in Figure 1 (a), 24 visual relations are detected from the given image, while only five of them are used for describing the image main content in visual captioning as in Figure 1 (b) and (c). We are therefore inspired to pursue those visual relations that are more semantically important than others among all detected ones for describing the main content of an image. We call such a relation "visual relation of interest" (VROI).

Most existing VRD methods [19, 44, 45] do not distinguish VROIs from other visual relations explicitly, though ground truths of some VRD datasets are extracted from image captions, such as VisualGenome [13]. Recently, a new dataset named VrR-VG [15] has been built based on the VisualGenome dataset [13], whereas it focuses on balanced data distribution and visually relevant relations that are not necessarily "interesting" in semantics. Some VRD related tasks also seek to generate more accurate and reasonable relations. For example, [21] measures the salient weights of different relations with an attention module for salient visual relationship detection, but it does not annotate the semantic importance of relations; [46] focuses on the salient interaction regions by combining multi-level features to accurately recognize the predicate of a subject-object pair; [25] estimates the relevance of relations with prior potentials but does not measure the semantic importance of relations. They are not targeted at extracting VROIs that are meaningful in terms of describing main image content. Scene graph generation (SGG) [27, 35, 40] restricts the number of detected visual relations in a given image with the goal of generating a meaningful graph for representing visual relations, but it cannot distinguish VROIs except for ground truth annotation.

Image captioning [1, 9, 33, 41] aims to represent the main content of an image with a complete sentence that might imply VROIs. However, we argue that VROIs can provide more detailed and structural descriptions of the image main content, bringing more benefits to many image understanding applications like visual question answering [2, 11]. Similarly, instance of interest detection (IOID) [39] also tries to extract the elements that are crucial for representing the image main content. However, it only focuses on the individual instances with higher essentiality, while VROIs are able to demonstrate interactions between two instances thus offer more comprehensive understandings of the image content.

In this paper, we propose a new task, Visual Relation Of Interest Detection (VROID), to detect VROIs from a given image. VROID can be seen as a further extension of the traditional VRD task. Similar to other existing VRD methods [5, 19, 44], VROID aims to represent the VROIs with relationship triplets and localize the subject and object for each relationship triplet by bounding boxes. Compared to VRD and IOID, VROID needs to tackle more technical challenges. It requires to retain the essential visual relations only, meaning it demands additional essentiality measurement of visual relations compared to VRD, which is much more complex than that of instances as in IOID. The visual relation of two significant instances may be described in diverse manners. For example, in an image illustrating a girl is working on a computer, possible predicates between "person" and "computer" include "beside", "behind", "touch" and "work on", but only "work on" denotes the main content. Hence the essentiality measurement in VROID should be discriminative enough to distinguish among these predicate candidates. Besides, VROID requires to detect the visual relations rather than the objects prior to the relation essentiality measurement, leading to greater difficulty for this new task than IOID due to the lower precision and recall of VRD compared with object detection. VROID can support various applications such as image retrieval [29], tagging [30], captioning [33] and visual question answering [2].

We devise a novel Interest Propagation neural Network (IPNet) for effectively detecting VROIs. It contains three modules: a Panoptic Object Detection (POD) module, a Pair Interest Prediction (PaIP) module and a Predicate Interest Prediction (PrIP) module. Firstly, the POD module extracts all instances in a given input image, which are represented as bounding boxes and corresponding categories. Also, the features of these instances and unions of any two instances are extracted and used in the next two modules. The PaIP module generates the interest probability for each pair of two instances (subject and object in the triplet). Specifically, we first generate the interest feature for every instance with its category, bounding box and feature, and also similarly for each instance pair. Then, the interest features of two instances are propagated to their pair by a modified graph convolutional network (GCN) inspired by [37], which concatenates the interest feature of a pair with the sum of interest features of corresponding instances after linear transformation, and the interest probability of a pair is generated upon the composite interest feature. Meanwhile, we use the PrIP module to predict the interest probability of all candidate predicates for each pair, *i.e.*, whether a predicate is interesting, given that its corresponding subject and object are both interesting. The interest features of pairs are propagated to triplets during the inference by multiplying the interest probability of a pair with that of a predicate under the condition of the pair being interesting.

Since VROID is a new VRD task, there is no existing dataset for VROID, to the best of our knowledge. Hence, we construct the first VROID dataset *ViROI* based on the IOID dataset [39] and the MSCOCO dataset [18]. The ground truth VROIs are manually annotated over images from the IOID dataset using our own annotation tool based on the corresponding captions of the MSCOCO dataset. The whole annotation work has been accomplished by expert annotators, followed by careful data cleaning to ensure annotation accuracy. The ViROI dataset contains 30,120 images and 109,764 annotated VROIs in total, on which we evaluate the proposed IPNet method and proves its better performance over other the baselines.

To sum up, our contributions are threefold: (i) We present a new VROID task aiming to detect the VROIs in a given image that are important for describing the image main content. (ii) We propose a novel IPNet method to address VROID that is proved to be very effective. (iii) We construct the first VROID dataset ViROI, which consists of 30,120 images with 109,764 manually labelled VROIs.

2 RELATED WORKS

Visual Relation Detection. Visual relation detection targets at detecting the relationships between two instances in images. One of its down-stream task, scene graph generation, aims to represent the detected visual relations with graphs. Both visual relations and scene graphs can support many down-stream tasks such as image captioning and image retrieval. Early VRD works [19, 38, 43] focus on relation prediction after object detection is independently performed. Later, the majority of methods like [42] incorporate an object detector into their pipeline, and take advantage of the object features from it for predicting visual relationships. Many efforts have been denoted to lifting performance by applying various techniques. Some methods [6, 19, 45] introduce language priors to facilitate the prediction of predicates in relation triplets. Starting from [35], context information is popularly considered to boost

the prediction accuracy. Some methods [32, 37, 40] propagate the context features among different object pairs through recurrent neural network (RNN) or tree structures, while some others [21, 46] score all the candidates with an attention mechanism based on global context. Starting from [40], several methods [10, 25] use data bias to raise recall of prediction. Following research works construct a balanced dataset [15] or try to make unbiased predictions [31].

In this work we seek to attain the visual relations that are especially important for conveying the main content of a given image. Our new VROID task focuses on the semantic importance of the relations compared with VRD that seeks to find all the visual relations. VROID can be seen as an extension of VRD.

Instance of Interest Detection. Different from traditional object detection that targets at recognizing all the objects, instance of interest detection aims to only detect objects that are beneficial to representing the image content while excluding the unimportant ones for describing a given image, which can benefit tasks such as image captioning and image retrieval, *etc.* In recent years it has been researched. For example, Yu *et al.* [39] proposed to leverage both visual saliency and semantic context to detect objects of interest through a Cross-influential Network. Our new VROID task differs from this task in that VROID aims at more important visual relations while this task pursues more important instances in the images.

Image Captioning. Image captioning is a task across the visual and language modalities. Inspired by machine translation, the RNN is usually used after a convolution neural network to generate image captions [33]. The attention mechanism is also adopted to automatically generate captions by attending to the salient parts of an image [9, 36]. The output of this task is a sentence describing the given image, while our VROID outputs a relationship triplet like <subject, predicate, object> plus bounding boxes of corresponding instances, which are much more precise in describing the content of a given image.

3 DATASET

Annotation Design. We construct the first dataset for VROID, named ViROI, based on the IOID dataset [39] and the MSCOCO 2017 dataset [18]. The IOID dataset contains 45,000 images in total, providing corresponding instances to each noun in captions of the images. The MSCOCO dataset consists of 123,287 images, each with instances segmentation, stuff segmentation, panoptic segmentation, person keypoints and captions provided. It is designed for training and evaluation of object detection, semantic segmentation, panoptic segmentation, keypoint detection, image captioning, *etc.*

We annotate VROIs in 45,000 images from the IOID dataset to build our new dataset by labelling relationship triplets mentioned in captions of the MSCOCO dataset and selecting the referred instances of each triplet, with the following two steps: 1) Label relationship triplets. We use Stanford CoreNLP Dependency Parser [22] to automatically extract possible relationship triplets as well as subject, predicate and object candidates from captions of MSCOCO. Our annotators check the automatically generated triplets and rectify some wrong ones using these candidates. They are also requested to manually compose relationship triplets based on the image caption if they consider these triplets are "missed" by the system. 2) Select instances of subject and object in each triplet. The

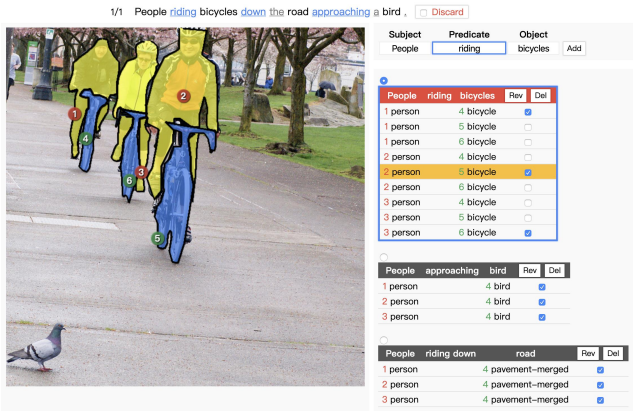

Figure 2: An example of the interface of our VROI annotation tool. The image caption is given at the topmost, the image to be annotated is given in the left, and the right shows tables for annotators to view and work.

IOID dataset provides instance segmentation for each image, and annotators check the correspondence between subject or object in a triplet and the detected instance in the image. However, when both subject and object have multiple corresponding instances, incorrect VROIs would be generated. In this case, our annotators manually pick out the correct subject-object pairs from all the possible pairs automatically generated by the system.

Application of Annotation Tool. We develop an annotation tool with an interface shown in Figure 2. The interface contains image captions, image to be annotated and labelled relations. In the given caption, candidate words are underlined to be clicked to fill into the subject, predicate, or object input box. Once annotators complete the three input boxes and press the "Add" button, a table will show in the right area with relationship triplet in the head and all possible pairs in the body. Meanwhile, the left image displays instance segmentation of subject in yellow and object in blue. If both subject and object have multiple corresponding instances, annotators need to click the segmentation area of subject and object in the left image to check correct pairs. In addition, we request the annotators to discard the images with relation triplet(s) in which subject and object both have too many corresponding instances.

Data Cleaning. To ensure annotation accuracy, we perform data cleaning with following six steps: 1) Lemmatize predicates using Stanford CoreNLP lemmatizer [22], but keep them in passive voice. 2) Filter out images with irregular predicates. We define 4 kinds of canonical predicates according to the VRD dataset: *verb* (*e.g.*, "hold", "ride", "look"), *preposition* (*e.g.*, "on", "with", "in"), *spatial* (*e.g.*, "next to", "in front of", "outside of") and *preposition phrase* (*e.g.*, "stand beside", "sit at", "walk through"). 3) Merge synonymous predicates based on WordNet [23]. 4) Reverse passive relationships if reversible, such as changing passive voice "held by" into active one "hold" by swapping subject and object. For those irreversible, such as "connected to", we keep the relationships unchanged. 5) Filter out relationships with rare predicates that appear less than 6 times in whole data, in order to limit the number of predicates. 6) Filter out duplicate relationships and images without any relationship.

Dataset Statistics. Our ViROI dataset consists of 30,120 images with 109,764 annotated VROIs, with instances and corresponding

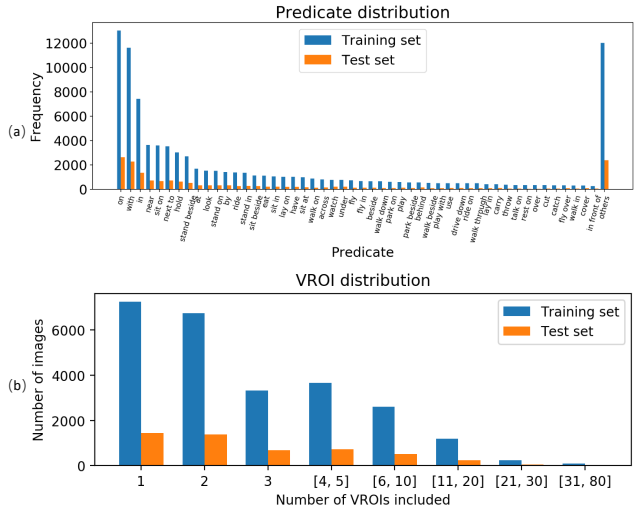

Figure 3: Dataset analysis. (a) Predicate distribution. (b) VROI distribution.

VROIs available for each image. We represent each instance with its category, bounding box and segmentation, and represent each VROI with its subject instance, predicate and object instance. There are 133 categories of instances, including 80 thing categories (*e.g.*, person, bicycle, car) and 53 stuff categories (*e.g.*, banner, blanket, bridge). There are 249 categories of predicates, including 77 verbs, 17 prepositions, 5 spatial and 150 preposition phrases. Our dataset contains 12,713 unique VROIs in total and an average of 6.676 things, 4.020 stuff and 3.644 VROIs per image. It is divided into a training set and a test set with similar distribution of predicates and VROIs, containing 25,091 images with 91,496 VROIs and 5,029 images with 18,268 VROIs, respectively. The training set contains an average of 6.684 things, 4.018 stuff and 3.647 VROIs per image. The test set contains an average of 6.634 things, 4.029 stuff and 3.632 VROIs per image. Details about data distribution are shown in Figure 3.

We would like to discuss about the data bias in our dataset. Bias naturally exists in our ViROI dataset because predicates and VROIs are extracted from image captions and the distribution is related to the occurrence frequency in natural language. We compare our ViROI dataset with the pre-processed VG dataset [35]. If the relation triplets are sorted according to the number of occurrence in a descendent order, in our dataset top-50% relation triplets account for 93.17% in training set and 88.87% in test set, while in the other dataset the statistics are 96.10% in training set and 94.88% in test set. We also design a simple baseline using frequency of relation triplets to further discuss data bias in the Section 5.3.

4 METHOD

We propose a novel Interest Propagation Network to solve our new VROID task, which outputs VROIs in triplets plus bounding boxes of corresponding instances for each given input image. The whole framework is shown in Figure 4. IPNet contains a Panoptic Object Detection module, a Pair Interest Prediction module and a Predicate Interest Prediction module. The POD module extracts instances from the input image that are represented as bounding boxes with corresponding categories, and also generates corresponding instance features and union features ("union" here refers to a rectangular area tightly containing each two objects). The PaIP module then predicts the interest score of each instance pair while the PrIP module predicts that of each predicate for each instance pair. Then we combine the interest scores of instance pairs with those of the corresponding predicates as the final interest scores. All VROI candidates are sorted by final interest scores and the highest ones are final results.

4.1 Panoptic Object Detection

Before detecting visual relations, we first extract the instances in the given image, which are divided into thing and stuff [12]. Object detection is typically performed in foreground, but we argue the background can provide scene information that is also important for conveying image content. Hence, we use a panoptic segmentation model in the Detectron2 framework [34] for panoptic object detection. In particular, we adopt the feature pyramid network [16] as backbone to extract five-layer image features. The thing predictor starts with a region proposal network, which takes as input the five-layer image features to generate thing-instance proposals. The image features are cropped and pooled according to the proposals, and the bounding boxes and classes of the proposals are predicted. Then, thing-instance candidates are filtered through non-maximum suppression. The stuff predictor combines the five-layer image features and predicts the category of each pixel. The bounding boxes of stuff-instances are extracted from the corresponding segmentations. The thing-instances and stuff-instances are finally merged.

Besides the above bounding boxes and classes of each instance, we also extract instance features which contain rich information benefiting following modules. Features of thing-instances can be directly extracted with the thing predictor while for stuff-instances this does not work, so we use an instance encoder to generate stuff-instance features with semantic consistency between thing-instances and stuff-instances. The raw features are extracted via ROI pooling with five-layer image features and instance bounding boxes. The instance encoder uses linear transformation to generate target instance features. To make instance features represent consistent semantic information, we predict the corresponding classes according to generated instance features during training.

We also combine each two instances as a pair and compute the union bounding box of subject and object in the pair, and the feature of the union is generated by the instance encoder with the five-layer image features and the union bounding box of the pair.

4.2 Pair Interest Prediction

We predict the interest of each pair with interest propagation from corresponding instances to the pair in this module.

We first predict the interest of instances with features of semantics, location and vision. Word embedding features pretrained by global vectors for the word representation (GloVe) model [24] are used as our semantics features. Location features represent the position comparative to the whole image:

$$Loc_i = \frac{x_i^{min}}{w} \oplus \frac{y_i^{min}}{h} \oplus \frac{x_i^{max} - w}{w} \oplus \frac{y_i^{max} - h}{h}, \quad (1)$$

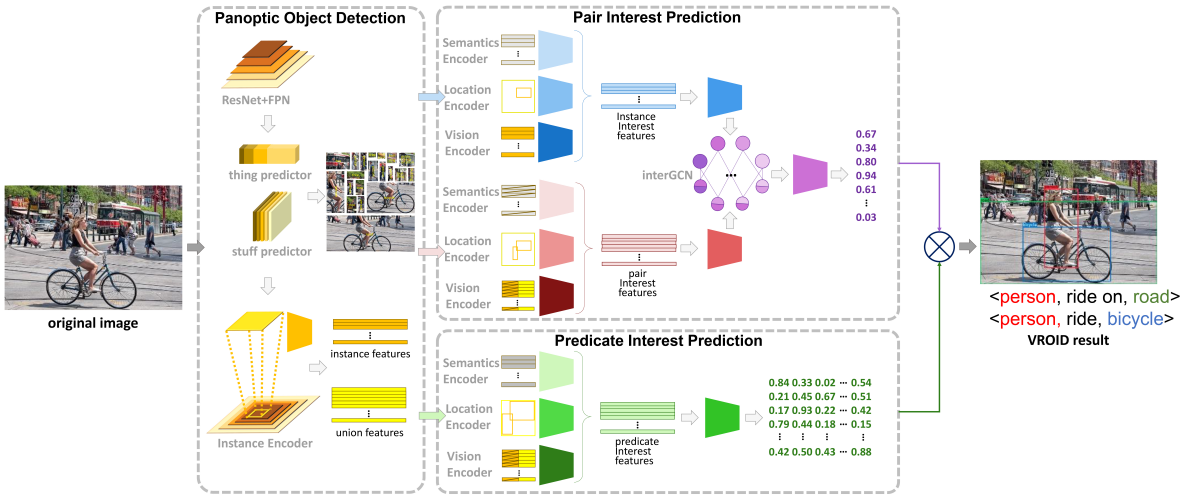

Figure 4: An overview of our IPNet. It consists of a Panoptic Object Detection module, a Pair Interest Prediction module and a Predicate Interest Prediction module.

where Loc_i is the location feature of the instance i, $\oplus$ is the concatenation operation, coordinates of the instance top-left corner and bottom-right corner are x_i^{min}, x_i^{max}, y_i^{min}, y_i^{max}, and height, width of the image are h, w. The vision features are the instance features generated as detailed in Section 4.1. Then three types of features are transformed to the same dimension and concatenated.

Before propagating the interest of instances to corresponding pairs, we also extract the pair features of semantics, location and vision. The semantics features are the difference in word embedding features of subject and object, which distinguishes a pair from its counterpart composed of reversed subject and object. The location features in this module represent the position of the subject and object comparative to the whole image:

$$Loc_p = \biguplus_{i \in \{s_p, o_p\}} \left(\frac{x_i^{min}}{w} \oplus \frac{y_i^{min}}{h} \oplus \frac{x_i^{max} - w}{w} \oplus \frac{y_i^{max} - h}{h}\right), \quad (2)$$

where Loc_p is the location feature of the pair p, $\oplus$ is the concatenation operation, $\biguplus$ represents concatenation on the instance level, s_p and o_p represent the subject instance and the object instance of the pair p, respectively. The same as Equation (1), the coordinates of an instance's top-left corner and bottom-right corner are x_i^{min}, x_i^{max}, y_i^{min}, y_i^{max} and the height, width of the image are h and w. The vision features are composed by those related to the pair's union box and the difference of subject features and object features:

$$F_p = (F_i^s - F_i^o) \oplus F_i^u, \quad (3)$$

where F_p is the feature of pair p, F_i^s and F_i^o represent the feature of the subject and object instance of pair p, respectively, F_i^u is the feature generated by the union box of subject and object of pair p and $\oplus$ is the concatenation operation. Vision features contain subject and object information and context information.

We modify the GCN in [37] to combine the interest features of a pair with its related instances, which is named interGCN:

$$G_p' = G_p \oplus \frac{\sum_{i \in I}(\epsilon_p^i \bar{G}_i)}{\sum_{i \in I} \epsilon_p^i}, \quad (4)$$

where G_p' denotes the updated interest feature of pair p after interGCN, G_p is the interest feature of pair p, $\bar{G}_i$ is the transformed interest feature of instance i passed from the instance interest module, I is the set containing all instances, ϵ_p^i represents the weight factor of instance i to pair p and $\oplus$ is the concatenation operation. We pose a constraint that only instances serving as subject or object of a pair can make influence:

$$\epsilon_p^i = \begin{cases} 1 & i \in \{s_p, o_p\} \\ 0 & i \notin \{s_p, o_p\} \end{cases}, \quad (5)$$

where s_p and o_p represent the subject instance and the object instance of the pair p, respectively.

4.3 Predicate Interest Prediction

We then generate possible predicates of interest for each subject-object pair. Similarly as the PaIP module, we also combine features of semantics, location and vision in this module. Inspired by [42], we transform word embedding features of the subject and object categories through a two-step long short term recurrent neural network [7]. The location features represent the position of the subject and object comparative to the union box:

$$Loc_p' = \biguplus_{i \in \{s_p, o_p\}} \left(\frac{x_i^{min}}{w'} \oplus \frac{y_i^{min}}{h'} \oplus \frac{x_i^{max} - w'}{w'} \oplus \frac{y_i^{max} - h'}{h'}\right), \quad (6)$$

where Loc_p' is the location feature of the pair p, $\oplus$, $\biguplus$, s_p, o_p, x_i^{min}, x_i^{max}, y_i^{min} and y_i^{max} are the same with those in Equation (2), and the h' and w' represent the height and width of the union box of pair p, respectively. The visual features are the same as those used in Section 4.2.

We predict the interest possibility of each predicate for all instance pairs in the image, given that both their subjects and objects are interesting. Considering our ViROI dataset only provides relations of interest annotation, inspired by [14], we apply semi-supervised learning as

$$L_{rela} = l_{rela}(r_{pred}^l, r_{gt}^l) + \beta l_{rela}(r_{pred}^u, r_{gt}^u), \quad (7)$$

where L_{rela} represents the loss in PrIP module, l_{rela} is the loss function, r^l_{pred} and r^u_{pred} denote prediction of labelled data and unlabelled data, respectively, r^l_{gt} and r^u_{gt} are ground truth of labelled data and unlabelled data, respectively, and β is the weight of unlabelled data's loss. The predictions with probability larger than a threshold directly work as the gt^u. The threshold is the γ-highest prediction probability of the labelled data, and γ is the number of positive samples in these labelled data. The value of β gradually rises to 1 along with the increase of iterations because the prediction is not accurate at the beginning.

4.4 Loss Function

Uninteresting relations are far more than relations of interest. Hence, normal training easily leads a model to predicting all visual relations as not interesting. Focal loss is proposed [17] to address the category imbalance that always challenges object detection. Considering that interesting relations only make a tiny fraction of the total possible relations, we use a modified focal loss function:

$$L^{pos} = -(1 - p^{pos})^2 \log(\mathrm{p^{pos}}), \tag{8}$$

$$L^{neg} = -p^{neg} \log(1 - \mathrm{p^{neg}}), \tag{9}$$

where L^{pos}, L^{neg} respectively denote the loss of positive and negative samples, and p^{pos}, p^{neg} respectively denote the probability score of positive and negative samples. This loss function will severely penalizes the model if it wrongly divides a positive sample into the negative category.

Possible instances of interest, pairs of interest and predicates of interest only account for a small proportion among all instances, pairs and predicate categories, respectively. The modified focal loss is used in all modules except for the instance encoder in the POD module. We use the typical cross entropy loss for the class prediction in the instance encoder. Total loss is the summarization of losses of all the modules:

$$L_{total} = L_{class} + L^{pos}_{ins} + L^{neg}_{ins} + L^{pos}_{pair} + L^{neg}_{pair} + L^{pos}_{rela} + L^{neg}_{rela}, \tag{10}$$

where L_{total} is the total loss of the whole end-to-end network, L_{class} is the loss of instance's class prediction in the instance encoder, L^{pos}_{ins} and L^{neg}_{ins} are positive and negative loss of instance interest prediction, L^{pos}_{pair} and L^{neg}_{pair} are positive and negative loss of pair interest prediction, and finally L^{pos}_{rela} and L^{neg}_{rela} are positive and negative loss of predicate interest prediction.

4.5 Relation Interest Inference

To generate the final triplet interest, the pair interest prediction and the predicate interest prediction are combined:

$$I_{spo} = E_{so} \cdot I_{so} \cdot P_{spo}, \tag{11}$$

where I_{spo} represents the interest of the triplet made up of subject s, predicate p and object o, I_{so} represents the interest probability of the pair composed by subject s and object o and P_{spo} is the interest probability of predicate p, given that the pair composed by the subject s and the object o is interesting. The E_{so} is a parameter, which is 0 when subject s and object o are the same and 1 otherwise.

5 EXPERIMENTS

5.1 Experimental Settings

All the experiments are conducted with i7-8086K 4.00GHz 12 cores CPU, 64GB memory and one TITAN V GPU, on our newly built ViROI dataset. We use typical metrics *Recall@k* and *Precision@k*, and also a new metric $\Psi@k$ for performance evaluation. The metric *Recall@k* is computed by

$$Recall@k = \frac{TP_k}{TP_k + FN_k}, \tag{12}$$

where TP_k and FN_k denote the number of correct relations predicted and unpredicted in the top k confident relation predictions, respectively. A correct relation is predicted if there is a relation prediction whose subject and object are within the same category as the correct relation's and with bounding box having $IOU > 0.5$, and the predicate is also within the same category as in ground truth. All predictions are sorted based on their confidence and then matched with ground truth in a descending order, and each prediction can only be matched once. Slightly different from the traditional definition of *Recall@k* [19], we allow each subject-object pair to have multiple predicates to adapt to our ViROI dataset, which is used and named as *No Graph Constraints Recall@k* in [40].

Recall@k is most widely used in VRD because it is almost impossible to exhaustively annotate all possible relations in the images [19]. But VROIs account for a small ratio to all visual relations and the prediction needs to be precise. Predicting VROIs as many as possible leads to increase of *Recall@k*, but decrease of precision. So we also use *Precison@k* for evaluation:

$$Precision@k = \frac{TP_k}{TP_k + FP_k}, \tag{13}$$

where TP_k is the same as in *Recall@k*, FP_k is the number of wrong relations predicted in the top k confident relation predictions.

Usually the value of k is set to 50 and 100 [19]. However in our test set, images with less than 10 and 20 VROIs account for 94.2% and 98.8% (shown in Figure 3 (b)), respectively, and a single image contains up to 76 VROIs. So we set the value of k to 10, 20, 50, 100 and θ. We set a variable θ for k considering the number of VROIs per image varies greatly and a fixed value of k may not accurately reflect performance. We use θ to compute *Recall@*θ [28], where θ is the number of correct relations in the image, *i.e.*, the value of k varying with images.

The value of k may be much larger than the number of correct relations in an image, and the precision would be very low even in the condition of best prediction. We thus apply a new metric $\Psi@k$ for more reasonable evaluation:

$$\Psi@k = \frac{TP_k}{TP^{max}_k}, \tag{14}$$

where TP_k is the same as in *Recall@k*, TP^{max}_k is the maximum number of correct relations predicted in k relationship predictions, which is equal to the smaller value between k and θ. $\Psi@k$ aims to evaluate the performance of a method compared with the best in theory. Note that *Recall@*θ, *Precision@*θ and $\Psi@\theta$ are actually the same since the number of top predictions is equal to that of VROIs. So we only keep *Recall@*θ in the following analysis.

Table 1: Ablation results of our method *vs.* different variants. R@θ, R@k, P@k and Ψ@k are abbreviations of *Recall@*θ, *Recall@*k, *Precision@*k and Ψ@k, respectively.

Method	R@θ	R@10	P@10	Ψ@10	R@20	P@20	Ψ@20	R@50	P@50	Ψ@50	R@100	P@100	Ψ@100
triplet as output	15.20	23.53	8.55	26.88	31.20	5.67	32.51	42.42	3.08	42.59	51.05	1.85	51.05
output with triplet	20.01	30.18	10.96	34.49	38.44	6.98	40.05	48.93	3.55	49.13	57.05	2.07	57.05
output without pair	0.18	1.62	0.59	1.85	3.47	0.63	3.61	7.89	0.58	8.01	13.38	0.49	13.38
only raw predicate	13.03	22.21	8.07	25.38	30.61	5.56	31.90	41.86	3.04	42.03	50.21	1.82	50.21
no instance	20.14	29.76	10.81	34.01	37.71	6.85	39.30	48.35	3.51	48.55	56.23	2.04	56.23
output with instance	18.37	27.53	10.00	31.46	35.48	6.44	36.97	46.11	3.35	46.30	54.29	1.97	54.29
no semantics features	19.48	29.12	10.58	33.27	37.23	6.76	38.79	47.63	3.46	47.83	55.55	2.02	55.55
no locations features	20.20	29.95	10.88	34.23	38.20	6.94	39.80	48.75	3.54	48.95	57.04	2.07	57.04
bce loss	13.58	20.95	7.61	23.94	27.21	4.94	28.35	36.39	2.64	36.54	43.42	1.58	43.42
Ours	**20.93**	**30.75**	**11.17**	**35.13**	**38.79**	**7.05**	**40.43**	**49.60**	**3.60**	**49.80**	**57.50**	**2.09**	**57.50**

5.2 Component Analysis

We evaluate the effect of five components in our method: interest propagation from pairs to triplets, interest propagation from instances to pairs, semantics features, location features and our loss function. The results are shown in Table 1.

We first test how interest propagation from pairs to triplets influences the performance.We modify the IPNet and make four variants. 1) The first variant concatenates the pair interest features and the corresponding predicate interest features to predict the final interest score of each triplet, instead of multiplying the interest scores of instance pairs and corresponding predicates as in the original method. 2) The second variant multiplies the triplet interest scores produced as in the first variant with the pair interest scores and the predicate interest scores. 3) The third variant takes the interest scores produced by PrIP as the final interest scores, without applying multiplication. 4) The fourth variant removes PaIP module and takes output of PrIP module as the final interest score of each relation triplet. Results of the four variants are shown in the first four lines of Table 1. It can be seen that their performance *w.r.t.* these metrics are always lower than those of our original method (denoted as Ours), demonstrating that the mechanism of propagating interest estimate from pairs to triplets is beneficial to the detection performance and that it is better to be implemented during inference than fusing features during training. Furthermore, comparing the results of the first and the forth variants, we find that directly estimating interest for relation triplets hurts the performance. This is because the tiny proportion of VROIs compared to all visual relations makes the network tend to predict all the visual relations as uninteresting.

We also evaluate the effect of interest propagation from instances to pairs. The PaIP module generates instance interest features and pair interest features and combines them with an interGCN, where interest propagates from instances to pairs. We design two variants for this test, and show their results in the fifth and sixth lines of Table 1. 1) We remove the component which generates instance interest features and the interGCN which combines instance interest features and pair interest features. For this variant, Ψ drops by more than 1.0% when interest propagation from instances to pairs is omitted. The proportion of interesting instances among all instances is much higher than that of interesting relations among all relations, meaning interesting instances are more easily predicted. Therefore, combining pair interest features with instance interest features can strengthen the prediction power of PaIP. But the feature of a pair is transformed from the features of subject and object instances, during which the input features for instances and pairs contain duplicate information. That may be why the improvement is not significant. 2) We propagate interest from instances to pairs during the inference period in the same way as that from pairs to triplets. We can see the performance is even worse than that when interest propagation from instances to pairs is omitted. This may be because an instance is likely to be contained in lots of triplets. Simply multiplying the interest scores of the subject and object instances cannot distinguish interesting relations, and combining instance interest features during training works better.

We use semantics features in both PaIP and PrIP module. To evaluate their effects, we remove them and show results of this variant in the seventh line. Its performance regarding Ψ is almost 2.0% lower than that of our original method. With the instance encoder in the POD module, the vision features contain the information about instance categories, but using the word embeddings of the instance categories still improves the performance. This proves that semantics information is important for detecting VROIs.

The location features are also used in PaIP and PrIP module, and we evaluate their effects by removing them in the two modules and comparing with the original method, as shown in the eighth line and the last line of Table 1. A performance decline is still observed. We also find adding the location encoder brings only a little improvement, even much less than that brought by adding semantics information. This means whether a relation is a VROI does not have a strong correlation with the location of its subject and object, which has more influences upon salient object detection that mainly depends on visual information.

Finally, we evaluate the effect of our loss function. Different from the original method, we use the typical binary cross entropy loss in this variant and show the results in the ninth line. Obviously, our original method performs much better, which shows penalizing wrongly predicted interesting samples in this case contributes a lot.

5.3 Comparison with Other Methods

We compare our method with baselines for VRD, SGG and image captioning tasks. We also design a baseline simply using frequency.

Table 2: Comparison results of our method *vs.* different baselines. R@θ, R@k, P@k and Ψ@k are abbreviations of *Recall*@θ, *Recall*@k, *Precision*@k and Ψ@k, respectively.

Method	R@θ	R@10	P@10	Ψ@10	R@20	P@20	Ψ@20	R@50	P@50	Ψ@50	R@100	P@100	Ψ@100
STA [38]	4.52	7.71	2.81	8.81	12.08	2.20	12.59	20.02	1.46	20.10	27.03	0.98	27.03
MFURLN [42]	5.73	9.32	3.39	10.65	13.24	2.41	13.79	19.84	1.44	19.93	25.28	0.92	25.28
IMP [35]	3.99	6.32	2.38	7.22	8.87	1.67	9.25	12.46	0.94	12.51	15.56	0.59	15.56
Graph R-CNN [37]	11.34	16.92	6.15	19.33	22.19	4.03	23.12	28.86	2.10	28.98	33.03	1.20	33.03
neural motifs [40]	15.09	21.93	7.97	25.06	27.34	4.97	28.49	33.67	2.45	33.80	37.60	1.37	37.60
VCTree [32]	17.78	25.96	9.43	29.67	32.26	5.86	33.62	40.38	2.93	40.55	46.05	1.67	46.05
VCTree [32]+DSS [8]	17.74	25.93	9.42	29.63	32.23	5.85	33.59	40.38	2.93	40.55	46.05	1.67	46.05
VCTree [32]+NLDF [20]	17.68	25.89	9.41	29.58	32.23	5.85	33.58	40.38	2.93	40.54	46.05	1.67	46.05
ARNet [3]	3.98	-	-	-	-	-	-	-	-	-	-	-	-
MMT [4]	4.94	-	-	-	-	-	-	-	-	-	-	-	-
Frequency	11.25	16.30	5.92	18.62	23.88	4.34	24.89	34.56	2.51	34.71	42.57	1.55	42.57
Ours	**20.93**	**30.75**	**11.17**	**35.13**	**38.79**	**7.05**	**40.43**	**49.60**	**3.60**	**49.80**	**57.50**	**2.09**	**57.50**

All the baseline models are retrained on our ViROI dataset with the code and default settings provided by their authors. The results are shown in Table 2.

We use two VRD baselines, STA [38] and MFURLN [42]. STA does not have its own object detector, and MFURLN's object detector is independent and not trained on MSCOCO for panoptic objects. Thus we use the objects extracted by the panoptic segmentation model in the Detectron2 framework [34] to predict relationships for STA and MFURLN. We can see that the performance of these two methods is rather poor and far worse than ours on the VROID task. This may be because they miss some object features from the object detector, proving that VRD methods can not well handle the VROID task. Also, the tiny proportion of interesting relations among all the relations may also be part of the reason why the VRD methods fail on our new task.

We use four SGG methods as baselines, including IMP [35], Graph R-CNN [37], neural motifs [40] and VCTree [32]. All the methods except IMP have their own internal object detectors but not trained on MSCOCO for panoptic objects. So we retrain their object detection models on our ViROI dataset; for IMP, we use the objects detected by Detectron2. According to the results in Table 2, some SGG methods perform much better than VRD methods, but the best one is still slightly inferior to ours. The possible reason is that they do not take interest estimation into account. Thus the SGG methods are able to solve the VROID task to some extent, but not well enough.

We also build two baselines using VCTree, whose performance is the best among the above baselines, to generate VROI candidates and apply salient object detection (DSS [8] and NLDF [20]) as post-processing to re-score the obtained VROI candidates by multiplying the maximum saliency values in the bounding box areas of the corresponding subject and object. The performance is slightly worse than that of the baseline only using VCTree, which proves that using salient object detection as post-processing is not an effective way to distinguish VROIs.

Since the VROIs in our dataset are annotated from the image captions, we design a baseline method of image captioning, dependency parsing and referring relationships. Similar to the annotating process, this method first generates a caption of the given image, then extracts semantic relationships from the caption based on dependency parsing [27], and finally maps each semantic relationship to the referred visual relationship in the image. We use ARNet [3] and MMT [4] for image captioning, Stanford CoreNLP Dependency Parser [22] for dependency parsing, and DSG [26] for referring relationships. Because the semantic relationships extracted from a single caption are very limited, each of them can be linked to only one visual relationship. So this method will not be effective enough, and the final number of relationships can hardly exceed 10. This is why we only give *Recall*@θ metric for these baselines in Table 2. The results prove that such a method performs even worse than the VRD methods.

We also design a simple baseline using frequency of relation triplets in training set as the interest score after panoptic object detection with the Detectron2 framework. This baseline evaluates the impact of data bias, and it demonstrates inferior performance compared to some of the above baselines, proving that a simple baseline using data bias cannot easily achieve good performance.

From all the comparison experiment results, it is proved that our IPNet is effective and advantageous for solving the VROID task.

6 CONCLUSIONS

We proposed a novel task named VROID that aims to detect VROIs for a given image. An IPNet was introduced to solve the VROID task, which consists of panoptic object detection and interest propagation from instances to triplets. Considering absence of any dataset for VROID, we constructed the first ViROI dataset that will be released publicly available soon. The experimental results validated the effectiveness of different components of our method. We also compared it with some baselines, and our method outperforms all of them.

ACKNOWLEDGEMENT

This work is supported by Natural Science Foundation of Jiangsu Province (BK20191248), National Science Foundation of China (61732007), Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20180307151516166), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5561–5570.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE International Conference on Computer Vision*. 2425–2433.

[3] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. 2018. Regularizing rnns for caption generation by reconstructing the past with the present. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7995–8003.

[4] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2019. M^2: Meshed-Memory Transformer for Image Captioning. *arXiv preprint arXiv:1912.08226* (2019).

[5] Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3076–3086.

[6] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. 2019. Scene graph generation with external knowledge and image reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1969–1978.

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[8] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3203–3212.

[9] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *IEEE International Conference on Computer Vision*. 4634–4643.

[10] Seong Jae Hwang, Sathya N Ravi, Zirui Tao, Hyunwoo J Kim, Maxwell D Collins, and Vikas Singh. 2018. Tensorize, factorize and regularize: Robust visual relationship learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1014–1023.

[11] Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162* (2017).

[12] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 9404–9413.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[14] Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*.

[15] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. VrR-VG: Refocusing Visually-Relevant Relationships. In *IEEE International Conference on Computer Vision*. 10403–10412.

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2117–2125.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*. 2980–2988.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. 740–755.

[19] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*. Springer, 852–869.

[20] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin A Eichel, Shaozi Li, and Pierremarc Jodoin. 2017. Non-local Deep Features for Salient Object Detection. IEEE Conference on Computer Vision and Pattern Recognition.

[21] Jianming Lv, Qinzhe Xiao, and Jiajie Zhong. 2020. AVR: Attention based Salient Visual Relationship Detection. *arXiv preprint arXiv:2003.07012* (2020).

[22] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics System Demonstrations*. 55–60.

[23] George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing*. 1532–1543.

[25] Francois Plesse, Alexandru Ginsca, Bertrand Delezoide, and Francoise Preteux. 2020. Focusing Visual Relation Detection on Relevant Relations with Prior Potentials. In *IEEE Winter Conference on Applications of Computer Vision*. 2980–2989.

[26] Moshiko Raboh, Roei Herzig, Jonathan Berant, Gal Chechik, and Amir Globerson. 2020. Differentiable scene graphs. In *The IEEE Winter Conference on Applications of Computer Vision*. 1488–1497.

[27] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language*. 70–80.

[28] Xu Sun, Yuan Zi, Tongwei Ren, Jinhui Tang, and Gangshan Wu. 2019. Hierarchical Visual Relationship Detection. In *Proceedings of the 27th ACM International Conference on Multimedia*. 94–102.

[29] Jinhui Tang, Zechao Li, Meng Wang, and Ruizhen Zhao. 2015. Neighborhood Discriminant Hashing for Large-Scale Image Retrieval. *IEEE Transactions on Image Processing* 24, 9 (2015), 2827–2840.

[30] Jinhui Tang, Xiangbo Shu, Guojun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. 2017. Tri-Clustered Tensor Completion for Social-Aware Image Tag Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 8 (2017), 1662–1674.

[31] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased Scene Graph Generation from Biased Training. *arXiv preprint arXiv:2002.11949* (2020).

[32] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6619–6628.

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[34] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[35] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5410–5419.

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*. 2048–2057.

[37] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *European Conference on Computer Vision*. 670–685.

[38] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2018. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *European Conference on Computer Vision*. 36–52.

[39] Fan Yu, Haonan Wang, Tongwei Ren, Jinhui Tang, and Gangshan Wu. 2019. Instance of Interest Detection. In *ACM International Conference on Multimedia*. 1997–2005.

[40] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5831–5840.

[41] Zhengjun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2019. Context-Aware Visual Policy Network for Fine-Grained Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1.

[42] Yibing Zhan, Jun Yu, Ting Yu, and Dacheng Tao. 2019. On Exploring Undetermined Relationships for Visual Relationship Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5128–5137.

[43] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 5532–5540.

[44] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *IEEE International Conference on Computer Vision*. 4233–4241.

[45] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. 2019. Large-scale visual relationship understanding. In *AAAI Conference on Artificial Intelligence*, Vol. 33. 9185–9194.

[46] Sipeng Zheng, Shizhe Chen, and Qin Jin. 2019. Visual Relation Detection with Multi-Level Attention. In *ACM International Conference on Multimedia*. 121–129.