

## Visual Relationship Detection with Contextual Information

Yugang Li<sup>1,2,\*</sup>, Yongbin Wang<sup>1</sup>, Zhe Chen<sup>2</sup> and Yuting Zhu<sup>3</sup>

**Abstract:** Understanding an image goes beyond recognizing and locating the objects in it, the relationships between objects also very important in image understanding. Most previous methods have focused on recognizing local predictions of the relationships. But real-world image relationships often determined by the surrounding objects and other contextual information. In this work, we employ this insight to propose a novel framework to deal with the problem of visual relationship detection. The core of the framework is a relationship inference network, which is a recurrent structure designed for combining the global contextual information of the object to infer the relationship of the image. Experimental results on Stanford VRD and Visual Genome demonstrate that the proposed method achieves a good performance both in efficiency and accuracy. Finally, we demonstrate the value of visual relationship on two computer vision tasks: image retrieval and scene graph generation.

**Keywords:** Visual relationship, deep learning, gated recurrent units, image retrieval, contextual information.

### 1 Introduction

Majority of real-world images involve multiple objects and the relationships between them contain crucial information for understanding the images. Visual relationships are two localized objects connected by a predicate. Visual relationship is very useful for downstream computer vision applications, e.g., image captioning [Vinyals, Toshev, Bengio et al. (2015)], scene graph generation [Xu, Zhu, Choy et al. (2017)], and visual question answering (VQA) [Antol, Agrawal, Lu et al. (2015)]. Booted by the progresses of deep learning, recent years have witnessed excellent progresses in many fields of computer vision, for example, object detection [Girshick (2015); Redmon, Divvala, Girshick et al. (2016); Ren, He, Girshick et al. (2015)], image classification [Krizhevsky, Sutskever and Hinton (2012); Simonyan and Zisserman (2014)] and quantum image steganography [Liu, Gao, Wang et al. (2019); Qu, Cheng and Wang (2019)]. However,

---

<sup>1</sup> School of Computer and Cyberspace Security, Communication University of China, Beijing, 100024, China.

<sup>2</sup> Academy of Broadcasting Science, Beijing, 100866, China.

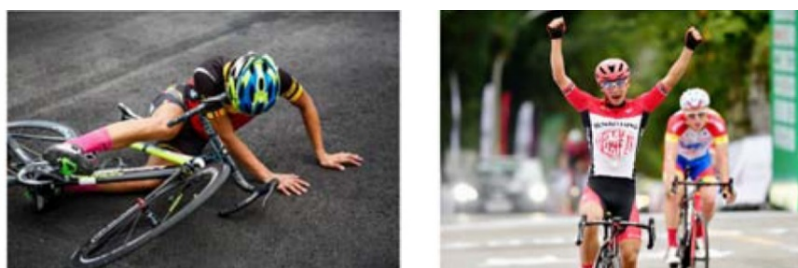
<sup>3</sup> School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, 639798, Singapore.

\* Corresponding Author: Yugang Li. Email: liyugang@abs.ac.cn.

Received: 21 May 2019; Accepted: 01 July 2019.

visual relationship detection still a difficult task.

Visual relationships are the determinant of image holistic understanding, they naturally bridge the semantic gap between language and vision. We represent a relationship as a subject-predicate-object triplet, for example (subject, predicate, object). The predicate can be a verb (e.g., “ride”), spatial (e.g., “on”), comparative (e.g., “taller”), or preposition (e.g., “with”). A common approach that detects visual relationship is to use the statistical patterns of co-occurrence between objects and their spatial layout for inferring. However, most previous methods focus on making local predictions of the relationships [Dai, Zhang and Lin (2017); Lu, Krishna, Bernstein et al. (2016); Zhu, Jiang and Li (2017)], which ignore surrounding contextual information of the objects. However the contextual information may be very useful for relationship prediction, because local prediction is lack of capacity in formation expression. Take a look at Fig. 1 for an illustration, the two images all have person and bike. Traditional visual relationship detector perceives an image by attending to individual objects and their relationship. As a result, the above two images would be labeled as (person, on, bike), which cannot describe the subtle difference between them.



**Figure 1:** Person-fall off-bike (left) and person-ride-bike (right)

A usual visual relationship detection process is as follows. Given a set of images in which the objects have been localized by bounding boxes, then specify the “relationship” among the objects in the images; for example, person and bike maybe related by riding; person and couch maybe related by sitting on. These relations can facilitate the detection of the objects and the relationship between them. A model can improve visual relationship detection if it has the following properties: 1) The model complexity should be able to compensate for the data complexity while still making a good performance gains for the learning problem. 2) It is better if the above feature can generalize to unseen data with little information about unseen observations.

As mentioned, although the research of visual relationship detection has gained plenty progresses, but still need to improve in terms of speed and accuracy. In this work, we propose a novel model to cope with the problem of visual relationship detection. The core of the model is a relationship inference network, which is a recurrent structure. Instead of inferring the visual relationship by using local region features, the relationship inference network can refine the feature of object by fusing contextual information extracted from surrounding regions.

In summary, our major contribution is that instead of inferring visual relationship in isolation, we propose a model which can incorporate contextual messages into the object

feature and then infer the visual relationship iteratively by using Gated Recurrent Unit (GRU) which is a generic recurrent neural networks (RNN) unit. We evaluate our model on two datasets: Stanford VRD and Visual Genome [Krishna, Zhu, Groth et al. (2017)] which contains more than 108 K images where every image has an average of 35 objects, 21 pairwise relationships and 26 attributes. The experimental results show that the proposed model outperforms the benchmark methods in accuracy.

## **2 Related work**

In this section, we are going to review some researches relate to the visual relationship detection. In the past few years, intermediate level computer vision tasks have witnessed a resurgence leading to various datasets and many effective algorithms. The recent success of quantum [Liu, Gao, Yu et al. (2018); Qu, Li, Xu et al. (2019)] used in machine learning has also thrust the development of computer vision. Thanks to these progresses, higher-level computer vision tasks, e.g., scene understanding and relationship inference, which depend on the lower-level modules have been studied and made some great progresses. In particular, the task of detecting the visual relationship between objects becomes our next goal-going from low level detection to high level semantic relations detection among objects.

The value of exploiting high-level knowledge from images has been demonstrated in many computer vision tasks. For instance, answer questions related to a given image [Antol, Agrawal, Lu et al. (2015)] has shown good results, and image captioning [Vinyals, Toshev, Bengio et al. (2015); Xu, Ba, Kiros et al. (2015)] can generate high-level knowledge from images. In this work, we put attention on visual relationships detection, which has been demonstrated that high-quality groundings can provide more comprehensive scene understanding.

Visual relationship detection goes beyond just locating the objects in an input image, it also requires understanding the relationship between the objects. The value of visual relationship has been shown in many visual tasks, for example, semantic image retrieval [Johnson, Krishna, Stark et al. (2015)], visual question answering [Teney, Liu and Heng (2017)], and complex query retrieval [Johnson, Krishna, Stark et al. (2015)]. An intuitive method to this task is to treat it as a problem of classification. Early models usually treat different objects combinations and the relationship between them as different classes [Sadeghi and Farhadi (2011)]. This method may work in the situation where the number of objects is small, but it would encounter a difficulty in general—plenty of imbalanced classes. For example, if an image has  $N$  objects and  $K$  predicates, it needs to train  $O(N^2K)$  detectors separately.

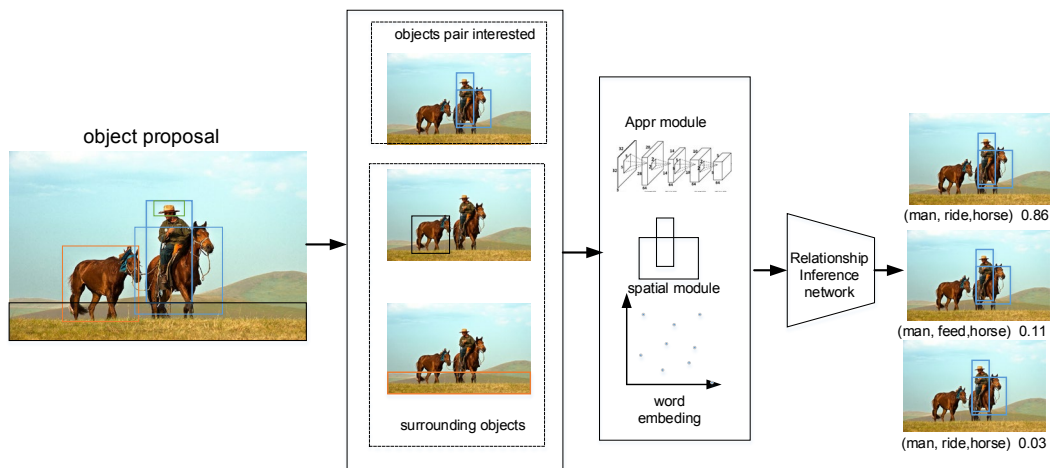
An alternative method is to consider every type of relationship predicate as a class. Some previous work has focus on some special types of relationships [Galleguillos, Rabinovich and Belongie (2008); Gould, Rodgers, Cohen et al. (2008)], such as spatial relationships between objects to improve segmentation. There also some efforts in human-object interaction [Maji, Bourdev and Malik (2011); Yao and Li (2010)] and action recognition which learn discriminative a model which can distinguish between relationship where the subject is a human. However, visual relationship detection is more general because the subject is not must be a human and the predicate does not must be a verb. Most previous work has focus on making local predictions of object relationships, which is not sufficient

for a precise detection. We propose a model that can aggregate surrounding contextual information and generate the visual relationship.

### 3 Visual relationship detection

#### 3.1 Overview

Visual relationship plays a very important role in image understanding. Visual relationship detection is the task of recognizing the different interactions between a pair of objects. These interactions can be spatial (e.g., *under*), verbs (e.g., *wear*), prepositions (e.g., *with*), action (e.g., *kick*), comparative (e.g., *taller than*) or a preposition phrase (e.g., *fall off*). The object of our proposed model is to detect the visual relationship in an image efficiently and accurately. Our detection pipeline is illustrated in Fig. 2. The motivation in our model is that the predictions of object and relationship can benefit from their contextual information. For example, when there is “a horse is standing on grass” can increase the probability of predicting the relationship of “person riding horse”. By using this observation, we propose a model which can aggregate contextual information of the objects to infer the visual relationship.



**Figure 2:** An overview of our proposed visual relationship detection framework

Our work is inspired by the recent progresses in machine translation [Sutskever, Vinyals and Le (2014)] which employs Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber (1997)] to aggregate a word with the contextual information to encode a long sentence. After inputting an image, the proposed model first uses an object detection method to locate the objects in the image. For each pair of objects, contextual information will be extracted then by fusing with appearance feature and spatial configurations to generate enhanced features of the located objects. These features will be fed to the relationship inference network. Finally, the relationship triplet ( $s$ ,  $r$ ,  $o$ ) will be generated by choosing the categories which have the highest scores. In this work, we use GRU instead of LSTM, for the purpose of achieving better flexibility in a principled training framework.

### 3.2 Feature extraction

In our work, we extract three types information to improve the predication accuracy. We add a layer to extract the statistical relationship between pairs of objects spatial configurations and appearance of objects:

(1) **Contextual information.** To generate the visual relationship we have to get some initial object bounding boxes, which can be either from algorithmically generated or ground-truth manual annotation. In our work, we use the Faster r-cnn [Ren, He, Girshick et al. (2015)] to automatically generate the objects bounding box set  $BI$  from an image  $I$ , then fed the set  $BI$  to the model.

For every object, we need to get two variables: (1) the class label of an object, and (2) four bounding box offsets of the proposal box coordinates. Given the relationship type set  $R$  and classes set  $C$ , we denote all variables set to be  $x = \{x_i^{cls}, x_i^{bbox}, x_{i-j} \mid i=1 \dots n, j=1 \dots n, i \neq j\}$ , where  $n$  denotes the number of proposal bounding boxes,  $x_i^{bbox} \in \mathfrak{R}^4$  is the bounding box offsets of the  $i$ -th proposal box coordinates,  $x_i^{cls} \in C$  is the class label of  $i$ -th proposal box, and  $x_{i-j} \in R$  is the predicate between  $i$ -th and  $j$ -th proposal bounding boxes.

Formally, we formulate the visual relationship problem as:

$$P(x_{i \rightarrow j} \mid x_k^{cls}, x_k^{bbox}, x_s^{cls}, x_s^{bbox} \dots, k \neq s \neq i \neq j) \tag{1}$$

Here,  $k$  and  $s$  denote the  $k$ -th and  $s$ -th object. By learning the distribution over the next input  $p(x_{t+1} \mid x_t, \dots, x_1)$ , a GRU can be used to learn a distribution of a sequence. We tackle this problem by using GRU [Cho, Van Merriënboer, Bahdanau et al. (2014)] due to its simplicity and effectiveness. In that case, the output at time setp  $t$  is the conditional distribution  $p(x_t \mid x_{t-1}, \dots, x_1)$ . For instance, a multinomial distribution (1-of- $K$  coding) can be outputted by using a softmax function:

$$p(x_{t,j} = 1 \mid x_{t-1}, \dots, x_1) = \frac{\exp(w_j h_{(t)})}{\sum_{j=1}^K \exp(w_j h_{(t)})} \tag{2}$$

For all symbols  $j=1, \dots, K$ , where  $w_j$  denote the rows of a weight matrix  $W$ . At the time step  $t$ , the hidden state  $h_{(t)}$  of the RNN is computed by:

$$h_{(t)} = f(h_{t-1}, x_t) \tag{3}$$

Here,  $f$  is a non-linear function. The hidden state of the GRU updates according to Eq. (3).

(2) **Spatial configurations.** Spatial configurations also contain important information for visual relationship inference, for example, the relative positions and relative sizes of objects. In this work, we use a 4-d vector  $(t_x, t_y, t_w, t_h)$ , which denotes the bounding box parameterization to represent the spatial configurations. Where  $(t_x, t_y)$  is a scale-invariant translation and  $(t_w, t_h)$  is the log-space height/width shift relative to its counterpart object or subject. Here, we take object for an example:

$$t_x = \frac{x - x'}{w'}, t_y = \frac{y - y'}{h'}, t_w = \log \frac{w}{w'}, t_h = \log \frac{h}{h'} \tag{4}$$

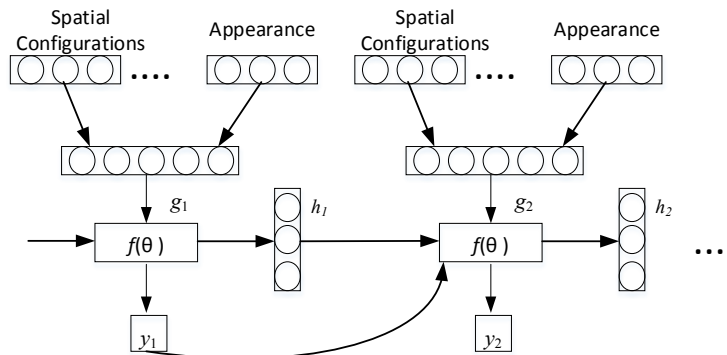
Here  $(x', y', w', h')$  and  $(x, y, w, h)$  are the box coordinates of object and subject respectively.

(3) **Appearance.** Intuitively, the type of the relationship can be reflected by the appearance of the located objects and their pairwise combination. In this work, we extract the appearance feature by using the 16 layers VGG network [Simonyan and Zisserman (2014)] to a bounding box which encloses the objects with a very small margin. It is a  $D$ -d vector transformed from a convolutional feature of shape  $X \times Y \times C$ .

The above three contextual features can be fused by three weights, which are learnable because the feature contribution varies among different relations. The fused feature is then fed to the relationship inference network, which will be introduced in the next section.

### 3.3 Joint detection

After we get the object features and its contextual information, the next consideration is how to fuse them together. A common way [Dai, Zhang and Lin (2017); Li, Ouyang, Wang et al. (2017)] is to build three independent CNN branches, each branch extracts useful information from other branches to refine the feature. But these methods ignored the information of surrounding objects or other global information. In this work, we present an inference network to fuse these features for a more accurate relationship detection. The relationship inference network is a recurrent structure, which is widely used in many deep learning tasks, e.g., visual attention [Mnih, Heess and Graves (2014)], machine translation [Sutskever, Vinyals and Le (2014)], multi-label image classification [Wang, Yang, Mao et al. (2016)] and transcribe speech utterances to characters [Chan, Jaitly, Le et al. (2016)]. In this work, we show that the straightforward use of GRU architecture can solve the problem of visual relationship generation by iteratively message passing. GRU takes the concatenation of the features as input and then combine with the representation at previous time  $h_{t-1}$  to produce the new state of the model  $h_t$ . The relationship inference network is shown in Fig. 3, the spatial configurations and appearance features are mapped into a space by independent linear layers and then combined using another linear layer to produce the feature representation  $g$ . Function  $f(\theta)$  takes the feature representation  $g$  and combines with the hidden representation at previous time  $h_{t-1}$  to produce the new hidden state  $h_t$ .



**Figure 3:** Relationship inference network

The relationship inference network processes input sequentially, and incrementally incorporates contextual information to enhance the object features for visual relationship

detection. At the end of the model, we add a softmax layer to produce the scores for the relationship predicate.

### 3.4 Architecture details

The combined contextual information, spatial configurations and appearance feature will be input to the relationship inference network for joint inference. The model produces the prediction of visual relationship by choosing the classes which have the highest scores.

At the time of training, all modules of our framework, that is object detection, joint detection and pair filtering are trained respectively. We sample a set which contains 256 region proposal bounding boxes produced by the region proposal network (RPN). If the bounding box has an intersection over union (IoU) larger than 0.7 with other ground truth regions, we annotate it with a positive label, and if the  $\text{IoU} < 0.3$  we annotate it with a negative label. The positive proposals are inputted to the classification layer. After that we use non-maximum suppression (NMS) for all classes with the  $\text{IoU} > 0.4$  and on average produce 15.6 detected objects.

In our work, we use Conditional Random Field (CRF) to aggregate statistical relations into the discriminative task. For visual relationships detection task, the formulation of CRF is:

$$p(r, s, o | x_r, x_s, x_o) = \frac{1}{Z} \exp(\Phi(r, s, o | x_r, x_s, x_o; W)) \quad (5)$$

where  $x_r$  denotes predicate feature which combines both the spatial configurations and the appearance of the enclosing box;  $x_s$  and  $x_o$  are the fused features of the subject and the object respectively;  $Z$  is the normalizing constant; and  $W$  is the model parameters. The joint potential  $\Phi$  can be computed by adding the individual potentials as:

$$\begin{aligned} \Phi = & \psi_a(s | x_s; W_a) + \psi_a(o | x_o; W_a) + \psi_r(r | x_r; W_r) \\ & \varphi_{rs}(r, s | W_{rs}) + \varphi_{ro}(r, o | W_{ro}) + \varphi_{so}(s, o | W_{so}) \end{aligned} \quad (6)$$

where  $\psi_a$  is a unary potential, it connects objects with their features;  $\psi_r$  connects the feature  $x_r$  with the relationship predicate;  $\varphi_{rs}$ ,  $\varphi_{ro}$  and  $\varphi_{so}$  are binary potentials, they compute the statistical relations between the relationship predicate  $r$ , the subject  $s$  and the object  $o$ .

### 3.5 Implementation details

The object of the proposed model is to exploit the visual relationship from images. At training time, a set of images are input to our model. The input images are annotated with the relationship of the objects, and each subject or object is located by a bounding box and labelled as (subject, predicate, object). At test time, the input is an image without annotations and outputs a pair of objects located by bounding boxes and the relation prediction score of it.

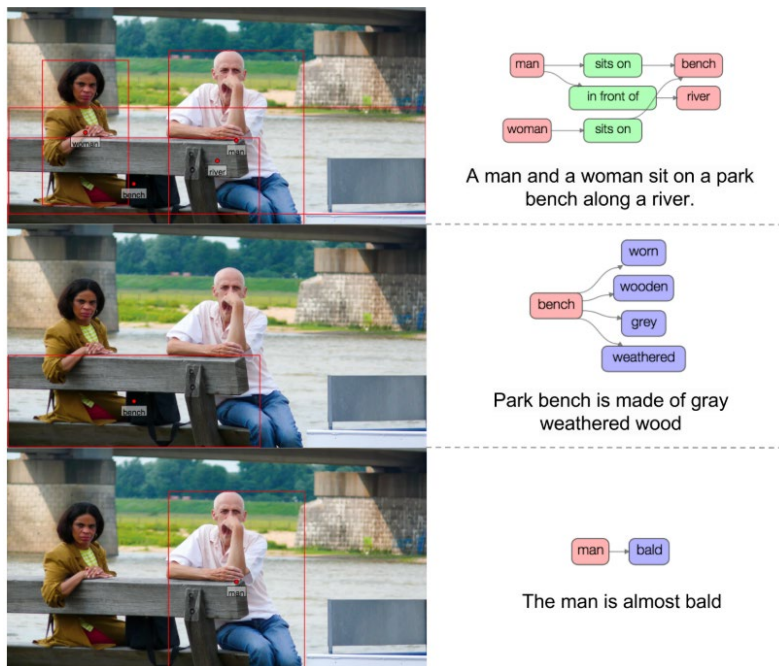
For the purpose of avoiding the gradient exploding/vanishing problem, we train the model end-to-end by using stochastic gradient descent (SGD) and set the momentum rate 0.8. We employ the cross entropy loss for the relationship predicate and object classes. For model initializations, we extract visual features by using a VGG-16 network which is pre-trained by MS-COCO. We fix the parameters of convolution layers, and only optimize other layers.

We randomly initialize the relationship inference network with Gaussian weights.

#### 4 Experiments

The datasets for visual relationship detection are different with the datasets for other computer vision tasks. A relationship detection dataset should have more than just objects localized by bounding box in images, the images should also have a set of relationships. In this work, we show the effectiveness of our model on two datasets: Visual Genome and Stanford VRD. VG is proposed for cognitive tasks, e.g., question answering and image description. The dataset contains more than 108K images, every image contains an average of 21 pairwise relationships, 35 objects and 26 attributes. Different with previous dataset, Visual Genome treats relationships as first-class citizens. Fig. 4 is an example of VG, from which we can see that the dataset composed by many elements. Compare with Visual Phrases and MS-COCO, Visual Genome is more commonly used. We use 3000 images to train our model and then perform visual relationship detection on the other 1000 images.

We train our model using TensorFlow [Abadi, Barham, Chen et al. (2016)]-a widely used deep learning architecture, which use dataflow graphs to represent computation, shared state and supports a variety of applications. We use the ImageNet to pre-train the appearance module, while relationship inference network and the spatial module are initialized randomly. After initialization, we use SGD to optimize the entire network.



**Figure 4:** An example from Visual Genome

##### 4.1 Evaluation metric

The evaluation metric which we use is recall@100 and recall@50 [Alexe, Deselaers and Ferrari (2012)]. The recall@ $k$  metric indicates the proportion of ground-truth relationship



triplets are contained in the top  $k$  most confident triplet predictions of an image. We choose this metric because the sparsity of the relationship annotations in VG-metrics like mean average precision (mAP) would be unable to penalize positive predictions on unlabeled relationships.

#### 4.2 Comparative results

We compared our final model with two baseline models: (1) Visual Phrase (VP) [Sadeghi and Farhadi (2011)]: a representative approach which treats different triplet as a different class; and (2) Visual Relationship (VR) [Lu, Krishna, Bernstein et al. (2016)]: this model contains two components—a language module which can capture the language priors between objects, and a vision module which makes detections from images.

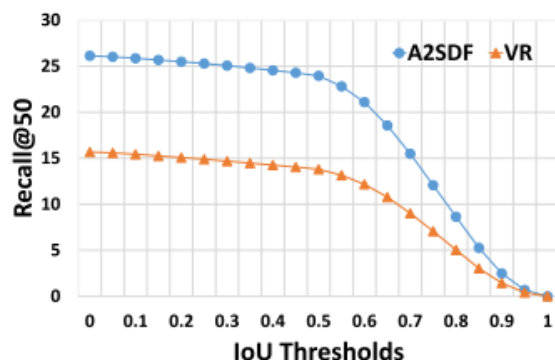
Tab. 1 demonstrates the performances of our model. On the dataset, we can see that: (1) VP performs very poorly, because it is unable to tackle such a huge and imbalanced class space. (2) VR performs substantially better, but remains unsatisfactory. (3) Our proposed model outperforms the above two methods. The results demonstrate that aggregate the contextual information from other hidden states can make the network yields superior performances.

**Table 1:** Comparison with VP and VR, measured by recall@50 and recall @100

Dataset	Model	Recall@50	Recall@100
VRD	VP	0.84	1.32
	VR	42.34	44.57
	Our model	76.66	82.43
VG	VP	0.97	1.91
	VR	47.87	47.87
	Our model	81.68	87.63

In order to investigate the extent each feature can influence on the task of visual relationship detection, we ablate our model into four methods according to the different features they use: (1) contextual information (2) spatial configurations (3) appearance and (4) all that uses contextual information, spatial configurations and appearance and fuse the above features with a scaling layer, respectively. We compare our method under two task settings: (1) **Relationship detection**. In this task, given a set of images, we need to output a set of (subject, predicate, object) triplets and localize both subject and object having at least 0.5 IoU with their ground-truth; (2) **Phrase detection**. In this task, the relationship triplet is treated as a whole. We need to output a set of (subject, predicate, object) triplets and the entire relationship having at least 0.5 IoU with the ground-truth.

Tab. 2 shows that: 1) fusing all the information can get the best performance in all type experiments; 2) for spatial relationship detection, location features perform better. This is because the errors which is made by relationship detection back-propagated to the object detection module. Despite the great gain compared with others, the recalls on phrase detection remains weak, because the limitations of the object detector module.



**Figure 5:** The performance of union-box detection

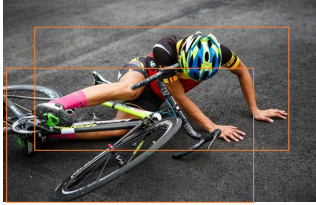
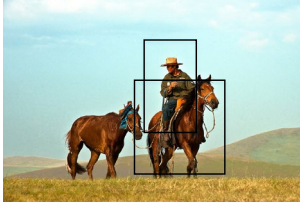
Fig. 5 demonstrates the union-box detection results evaluated by different IoU thresholds. The object detector can just get about 30% of object recall, measured by R@50. In order to get a better result on these tasks, we need to get a better object detector.

**Table 2:** Evaluation of the influence exert by different features on visual phrase detection and visual relationship detection

Dataset	Model	Phrase Det.		Relationship Det.	
		Recall@50	Recall@100	Recall@50	Recall@100
VRD	Contextual	52.12	55.26	63.13	65.14
	Spatial	53.14	54.26	62.14	64.56
	Appearance	55.74	57.28	70.26	75.82
	All	69.34	71.26	76.66	82.43
VG	Contextual	53.13	54.35	64.46	65.86
	Spatial	54.63	56.42	65.78	67.53
	Appearance	56.47	58.36	66.12	67.84
	All	73.23	75.12	81.68	87.63

We compared our model with different variants of the proposed method. Tab. 3 shows the predicted relationship on two random images. For the first image, VR and VP incorrectly predict the relationship. These models perform bad because they always tend to predict the visual relationship which they usually see at training time. In comparison, our proposed model can predict and localize the objects correctly in the image. The results show that our proposed model outperform the other models.

**Table 3:** This table lists the relationship detection results for a specific object pair. The first row lists images containing with the bounding box pairs, and following rows list the most probable relationship predicted by different methods

		
VP	(people, ride, bicycle)	(man, on, horse)
VR	(people, ride, bicycle)	(man, ride, horse)
Our model	(people, fall off, bicycle)	(man, ride, horse)

### 4.3 Applications

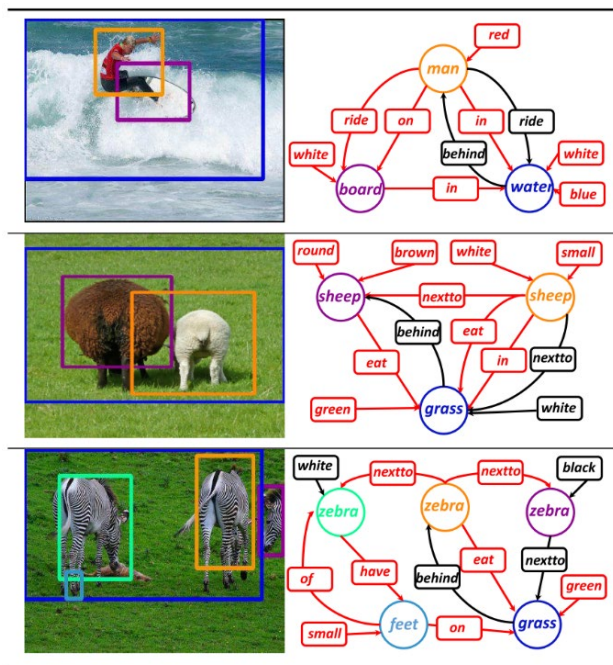
In this section, we will introduce two applications based on the visual relationship: image retrieval and scene graph generation, which are both important computer vision tasks. The task of scene graph generation is to produce a directed graph for a given image. The scene graph contains objects, objects attributes and the relationship between objects. See Fig. 6 for an illustration [Dai, Zhang and Lin (2017)]. The main challenge of scene graph generation is the visual relationship detection. In the experiment, we detect the visual relationship by using the proposed model, and then generate the scene graph. We compute the similarity between the ground truth and the generated scene graph by using average similarity as the metric. In our experiments, the proposed model can obtain better scene graphs than other methods.

Visual relationships can also improve the performance of image retrieval. In the experiments, we sample 1000 images as the test set. We choose 1 of the 1000 images randomly as a query and ranks the rest 999, and use two annotators to rank image results for each of the input queries. We evaluate the results by using median rank, R@1, R@5 and R@10. We compare our model with three most used image descriptors: CNN, GIST and SIFT. We rank the results of an input query by using the  $L_2$  distance. Given a test image, our model generates a set of visual relationships  $\{R_1, \dots, R_n\}$  with a probability of  $\{P_1^q, \dots, P_n^q\}$  respectively. Next, for every image  $I_i$  in the test set, it predicts  $\{R_1, \dots, R_n\}$  with a confidence of  $\{P_1^i, \dots, P_n^i\}$ . We calculate the matching score between the query and an image as  $\sum_{j=1}^n P_j^q * P_j^i$ . We compare our model with another four models by retrieval an image with a relationship (person, ride, horse), the results are shown in Tab. 4.

**Table 4:** Comparisons result of the five model

Model	R@1	R@5	R@10	Median Rank
GIST	0.00	5.60	8.70	68
SIFT	0.70	6.10	10.3	54
CNN	3.15	7.7	11.5	20
Visual Phrases	8.72	18.12	28.04	12
Our Model	11.02	32.12	48.12	4

From Tab. 4 we can see that, GIST and SIFT descriptors perform bad with a median rank of 68 and 54 respectively, the reason is that they only measure the structural similarity of inquire images. The performance of CNN descriptor is better with a median rank of 20 because it captures object-level information. Our method performs best, because it can capture the visual relationships of the query image. The experiments show that visual relationships can improve efficiency of image retrieval.



**Figure 6:** An illustration of three images and their corresponding scene graphs [Dai, Zhang and Lin (2017)]. In the scene graphs, the red boxes specify correct prediction and the black boxes specify wrong prediction

## 5 Conclusion

In this paper, we proposed a novel model to cope with the problem of visual relationship detection of an image. Our model generates the visual relationship by fusing the object features with the surrounding contextual information. The core of the proposed model is a relationship inference network, which extends its expressive ability of deep neural networks to the task of relational modeling. We evaluate our model on two commonly used datasets, the experimental results demonstrate the efficiency of our model. In this paper, we also demonstrated the effectiveness of visual relationship in image retrieval and scene generation. In the future, we have the following research plans: (1) apply our model in other computer vision tasks, e.g., VQA or image caption generation; (2) solve the problem of zero-shot relation learning.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A. et al.** (2016): Tensorflow: a system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265-283.
- Alexe, B.; Deselaers, T.; Ferrari, V.** (2012): Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189-2202.
- Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D. et al.** (2015): Vqa: visual question answering. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425-2433.
- Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O.** (2016): Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4960-4964.
- Cho, K.; Van Merriënboer, B.; Bahdanau, D.; Bengio, Y.** (2014): On the properties of neural machine translation: encoder-decoder approaches. arXiv:1409.1259.
- Dai, B.; Zhang, Y.; Lin, D.** (2017): Detecting visual relationships with deep relational networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3298-3308.
- Galleguillos, C.; Rabinovich, A.; Belongie, S.** (2008): Object categorization using co-occurrence, location and appearance. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8.
- Girshick, R.** (2015): *Fast R-cnn*. arXiv:1504.08083.
- Gould, S.; Rodgers, J.; Cohen, D.; Elidan, G.; Koller, D.** (2008): Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, vol. 80, no. 3, pp 300-316.
- Hochreiter, S.; Schmidhuber, J.** (1997): Long short-term memory. *Neural Computation*, vol. 9, no. 8, pp. 1735-1780.

- Johnson, J.; Krishna, R.; Stark, M.; Li, L. J.; Shamma, D. et al.** (2015): Image retrieval using scene graphs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3668-3678.
- Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K. et al.** (2017): Visual genome: connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32-73.
- Krizhevsky, A.; Sutskever, I.; Hinton, G. E.** (2012): Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pp. 1097-1105.
- Li, Y.; Ouyang, W.; Wang, X.; Tang, X. O.** (2017): ViP-CNN: visual phrase guided convolutional neural network. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1347-1356.
- Liu, W. J.; Gao, P. P.; Yu, W. B.; Qu, Z. G.; Yang, C. N.** (2018): Quantum relief algorithm. *Quantum Information Processing*, vol. 17, no. 10, pp. 280.
- Liu, W.; Gao, P.; Wang, Y.; Yu, W.; Zhang, M.** (2019): A unitary weights based one-iteration quantum perceptron algorithm for non-ideal training sets. *IEEE Access*, vol. 7, pp. 36854-36865.
- Lu, C.; Krishna, R.; Bernstein, M.; Li, F.** (2016): Visual relationship detection with language priors. *European Conference on Computer Vision*, pp. 852-869.
- Maji, S.; Bourdev, L.; Malik, J.** (2011): Action recognition from a distributed representation of pose and appearance. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3177-3184.
- Mnih, V.; Heess, N.; Graves, A.** (2014): Recurrent models of visual attention. *Advances in Neural Information Processing Systems*, pp. 2204-2212.
- Qu, Z.; Cheng, Z.; Wang, X.** (2019): Matrix coding-based quantum image steganography algorithm. *IEEE Access*, vol. 7, pp. 35684-35698.
- Qu, Z.; Li, Z.; Xu, G.; Wu, S.; Wang, X.** (2019): Quantum image steganography protocol based on quantum image expansion and grover search algorithm. *IEEE Access*, vol. 7, pp 50849-50857.
- Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A.** (2016): You only look once: unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788.
- Ren, S.; He, K.; Girshick, R.; Sun, J.** (2015): Faster r-cnn: towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, pp. 91-99.
- Sadeghi, M. A.; Farhadi, A.** (2011): Recognition using visual phrases. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1745-1752.
- Simonyan, K.; Zisserman, A.** (2014): Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
- Sutskever, I.; Vinyals, O.; Le, Q. V.** (2014): Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, pp. 3104-3112.

**Teney, D.; Liu, L.; van den Hengel, A.** (2017): Graph-structured representations for visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9.

**Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D.** (2015): Show and tell: a neural image caption generator. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156-3164.

**Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C. et al.** (2016): Cnn-rnn: a unified framework for multi-label image classification. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285-2294.

**Xu, D.; Zhu, Y.; Choy, C. B.; Li, F. F.** (2017): Scene graph generation by iterative message passing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410-5419.

**Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. et al.** (2015): Show, attend and tell: neural image caption generation with visual attention. *International Conference on Machine Learning*, pp. 2048-2057.

**Yao, B.; Li, F.** (2010): Modeling mutual context of object and human pose in human-object interaction activities. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 17-24.

**Zhu, Y.; Jiang, S.; Li, X.** (2017): Visual relationship detection with object spatial distribution. *IEEE International Conference on Multimedia and Expo*, pp. 379-384.