# Visual Relationship Detection with Deep Structural Ranking

**Kongming Liang**[1,2,3], **Yuhong Guo**[2], **Hong Chang**[1], **Xilin Chen**[1,3]
[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
[2]School of Computer Science, Carleton University, Ottawa, Canada
[3]University of Chinese Academy of Sciences, Beijing 100049, China
kongming.liang@vipl.ict.ac.cn, yuhong.guo@carleton.ca, {changhong, xlchen}@ict.ac.cn

## Abstract

Visual relationship detection aims to describe the interactions between pairs of objects. Different from individual object learning tasks, the number of possible relationships are much larger, which makes it hard to explore only based on the visual appearance of objects. In addition, due to the limited human effort, the annotations for visual relationships are usually incomplete which increases the difficulty of model training and evaluation. In this paper, we propose a novel framework, called Deep Structural Ranking, for visual relationship detection. To complement the representation ability of visual appearance, we integrate multiple cues for predicting the relationships contained in an input image. Moreover, we design a new ranking objective function by enforcing the annotated relationships to have higher relevance scores. Unlike previous works, our proposed method can both facilitate the co-occurrence of relationships and mitigate the incompleteness problem. Experimental results show that our proposed method outperforms the state-of-the-art on the two widely used datasets. We also demonstrate its superiority in detecting zero-shot relationships.

## Introduction

To achieve the goal of holistic image understanding, researchers have made great progress on recognizing, detecting and describing individual objects within an image. Over the past few years, the state-of-the-art of object recognition (Simonyan and Zisserman 2014) and object detection (Ren et al. 2015; Girshick 2015) have been dramatically improved thanks to the advances of deep learning. Based on these basic building blocks for single object understanding, visual relationship detection aims to accurately localize a pair of objects and determine the predicate between them as shown in Fig. 1. As a mid-level learning task, visual relationship detection can benefit many high-level image understanding tasks such as image captioning (Kulkarni et al. 2013; Vinyals et al. 2015; Xu et al. 2015) and visual question answering (Antol et al. 2015).

Different from single object based image understanding, visual relationship detection is used to describe two objects which makes the number of possible relationships much larger. If we represent the relationship as `subject-predicate-object`, the number of possible relationships are $\mathcal{O}(N^2K)$ when there are $N$ object categories and $K$ predicates. Since the object co-occurrence is infrequent, a large number of relationships contain few examples. Therefore previous method (Sadeghi and Farhadi 2011) which directly takes the tuple `subject-predicate-object` as a whole learning task can not scale well on large datasets. One possible solution (Lu et al. 2016) is to learn the objects and predicates separately and fuse the result to predict the relationship. In that way, different relationships (e.g. `truck-on-street, car-on-street`) are merged into the same category if they share the same predicate. Therefore, only $\mathcal{O}(N + K)$ detectors are needed but the samples within the same predicate category are highly diverse as shown in Fig.1(c). This problem also exists in some other relevant tasks. For visual attribute learning, an attribute may be used to describe various objects which

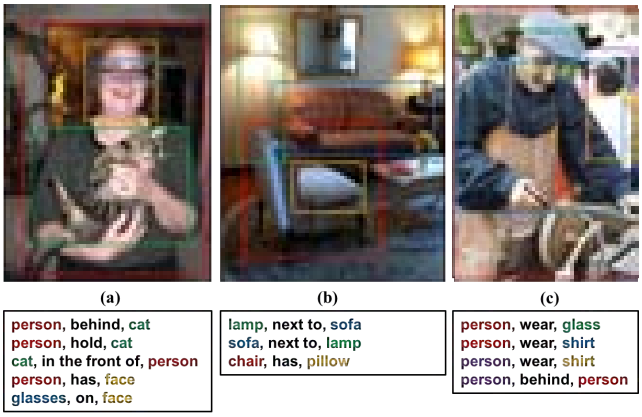


Figure 1: Detecting the visual relationships contained in an image. A visual relationship instance is defined as `subject-predicate-object`. (a) Multiple predicates may co-occur for a paired object, as the person can hold a cat and also be behind the cat. (b) The annotations are usually not complete. For example, `chair-in the front of-lamp` is not included. (c) The diversity within the same predicate is high since person can wear or be behind many kinds of objects.

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

makes its manifestations highly diverse (Farhadi et al. 2009; Liang et al. 2015). For image captioning and visual question answering, a textual word may contain multiple senses (Gella, Lapata, and Keller 2016).

Another challenge of visual relationship detection is that the relationships are not annotated completely. This problem mainly comes from three aspects. Firstly, a single image usually contains multiple objects and only a subset of them have been manually localized in the dataset. Secondly, given individual objects, some pairs of the objects are not annotated with any predicates even they do have a relationship. Thirdly, in most cases, only one predicate is defined for an annotated object pair even though the co-occurrence of the predicates are very common. One direct way to deal with incomplete predicate annotation is to take the unlabeled relationships as negative samples and solve the relationship detection problem using a multi-label classification framework based on cross-entropy loss. In addition, since many object pairs only contain a single predicate, some previous works (Dai, Zhang, and Lin 2017; Li et al. 2017; Zhang et al. 2017) choose to formulate relationship detection as a multi-class classification problem based on softmax loss function which totally ignores the co-occurrence of predicates.

In this paper, we propose a deep neural network framework with structural ranking loss to tackle the visual relationship detection problem. Our proposed method separately models object and predicate and fuses their results to predicate the relationship. To handle the diversity of each predicate, we propose to integrate multiple cues as the input feature. Specifically, the proposed network takes visual appearance cue, spatial location cue and semantic embedding cue as the input and further fuse them into a joint feature vector. In addition, we explore different ways to model each input cue and conduct experiments to validate their effectiveness. Based on the fused feature, we propose a structural ranking loss to learn the related predicate for a pair of localized objects. It enforces the annotated relationships to have higher relevance score than the rest of possible relationships. Even though some unannotated relationships may also be related with the given image, they are usually not as salient as the annotated ones. Moreover, we integrate the structural loss with the probability of the predicates conditioned on its subject and object to further reduce the impact of incomplete annotations. Experimental results show that our proposed method outperforms the state-of-the-art on the two widely used datasets Visual Relationship Dataset (Lu et al. 2016) and Visual Genome (Krishna et al. 2016).

The rest of this paper is organized as follows. We first introduce previous works in the following section. In the third section, we present our proposed deep structural ranking framework for visual relationship detection. Then we report the experimental results and conclude the paper in the last section.

## Related Work

Many previous works have tackled the visual relationship prediction problem. (Galleguillos, Rabinovich, and Belongie 2008) attempted to learn relationships that describe spatial information such as "above", "below", "inside", and "around". (Gould et al. 2008) further shown that this kind of relationship can be used to improve multi-class segmentation. Human-object interaction (Yao and Fei-Fei 2010; Chao et al. 2015a) aims to learn the relationships with the `subject` constrained to be human. (Farhadi et al. 2010; Sadeghi and Farhadi 2011; Ramanathan et al. 2015; Atzmon et al. 2016) considered each relationship as a distinct learning task. For example, (Sadeghi and Farhadi 2011) proposed to train a detector for `person-ride-horse` to improve the localization of `person` and `horse`. However, this kind of methods can only be used for a small set of relationships due to the long-tail distribution of visual relationships.

To facilitate large scale relationship understanding, a promising way is to learn the visual relationship separately. (Lu et al. 2016) proposed to use visual appearance features to predict the relationship of a pair of the objects. The visual modules for objects and predicates are trained separately and combined to predict the final relationships. They further leverage language priors to improve the proposed visual modules. One drawback of their method is that the network for extracting visual appearance features is trained separately from the final task. To tackle the above problem, the works of (Dai, Zhang, and Lin 2017; Li et al. 2017; Zhang et al. 2017; Liang, Lee, and Xing 2017) proposed to integrate the feature learning procedure with the target relationship detection task in an end-to-end training manner. More specifically, (Zhang et al. 2017) proposed to model visual relations by mapping subjects and objects into a common low-dimensional relation space. Therefore, the predicate is considered as a translation vector between the subject and object. (Liang, Lee, and Xing 2017) proposed a deep Variation-structured Reinforcement Learning framework to sequentially detect relationship and attribute instances by exploiting global context cues. (Li, Ouyang, and Wang 2017) proposed a Visual Phrase guided Convolutional Neural Network to learn three inter-connected recognition problems (subject- predicate-object) simultaneously through message passing. (Dai, Zhang, and Lin 2017) proposed a novel Deep Relational Network to exploit both spatial configurations and statistical dependencies among relationship predicates, subjects, and objects. However, most of the above choose to model relationship detection as a multi-class classification problem which ignore the relationship co-occurrence.

## Deep Structural Ranking

For visual relationship detection, we need to detect a set of objects and output the predicates between each pair of them. We use $\mathcal{P}$ to denote the set of all the annotated object pairs. For each element $(s, o) \in \mathcal{P}$, $s$ and $o$ represent the subject and object involved in a relationship. Then we use $\mathcal{P}_{s,o}$ to denote all the predicates which are annotated for the pair $(s, o)$. We use $\mathcal{R} = \{(s, p, o)|(s, o) \in \mathcal{P} \wedge p \in \mathcal{P}_{s,o}\}$ to denote all the visual relationship contained in an input image. In the training stage, the bounding boxes and labels for all the subjects and objects are observed. For testing, we first conduct object detection to acquire the location and label information for all the objects. Then we predict the relevan-

t predicates for each pair of them. As the predicates contain abundant semantic meaning and may generalize across different types of object pair, we propose a deep convolutional network which combines multiple cues to sufficiently learn the representation for an input instance. In this paper, we fuse the visual appearance cue, spatial location cue and semantic embedding cue for an relationship instance as its joint feature. Based on that, we propose to train the multiple cue network in an end-to-end manner with a novel structural ranking loss which considers both the natural co-occurrence and incompleteness of predicate annotations for visual relationship detection.

## Network Architecture

In this section, we introduce the proposed deep neural network for visual relationship detection. It integrates multiple cues to learn the representations of each input relationship instances. The proposed framework is shown in Fig. 2.

**Visual Appearance Cue.** Based on the visual appearance of two localized objects, human can easily acquire the appropriate predicates to describe their relationship. More specifically, human can determine both the object category information of two localized objects and the holistic interpretation from the context around them. Therefore, visual appearance cue is a key factor for the final learning task. For a relationship instance $(s, p, o)$, we use $\mathbf{b}_s = (x_s, y_s, w_s, h_s), \mathbf{b}_o = (x_o, y_o, w_o, h_o)$ to denote the bounding box coordinates, width and height of its corresponding `subject` and `object`. In addition, we use $\mathbf{b}_p$ to denote the area of the predicate which is usually represented as the union of $\mathbf{b}_s$ and $\mathbf{b}_o$. To capture the surrounding context, $\mathbf{b}_p$ can also encompass the subject and object with a small margin referring to (Dai, Zhang, and Lin 2017). Then we adopt a convolutional neural network as the backbone and extract the RoI Pooling features of $\mathbf{b}_s, \mathbf{b}_o$ and $\mathbf{b}_p$ from the last convolutional layer. In this paper, we choose VGG16 (Simonyan and Zisserman 2014) as the base network and the RoI features are further fed into two fully connected layers to be used as the visual appearance features. In our experiments, we first use the union area as the visual appearance and further combine it with the feature extracted on the separate subject and object area.

**Spatial Location Cue.** Spatial Location is complementary to visual appearance, as many of the `predicates` (e.g. `above, under, on the left of`) reflect the spatial or preposition relation between two objects which is not easy to learn only from visual feature. Here we explore two ways to leverage spatial information. Firstly, we try to use the relative location feature which is scale-invariant and also specifies the relative height/width between the two counterpart `subject` and `object`. We denote the location feature as a 4-d vector $(l_x, l_y, l_w, l_h)$. For `subject`, its spatial location cue is represented as:

$$l_x = \frac{x_s - x_o}{x_o}, l_y = \frac{y_s - y_o}{y_o}, l_w = log\frac{w_s}{w_o}, l_h = log\frac{h_s}{h_o}, \quad (1)$$

where $(l_x, l_y)$ specifies the translation and $(l_w, l_h)$ specifies the log-space height/width shift relative to its counterpart. Secondly, we use the spatial masks of `subject` and `object` to explore the spatial information for a relationship instance. A spatial mask is defined to be a binary matrix where only the pixels within the bounding box area are nonzero. The spatial mask is first generated as the same size of the input image and further down-sampled to the size $32 \times 32$. Then the spatial masks of both `subject` and `object` are concatenated as the input of a spatial neural network which compresses the spatial masks into a low-dimensional vector via three convolutional layers. In our experiments, we compare the above two ways to validate their effectiveness on leveraging spatial location cue.

**Semantic Embedding Cue.** Since the same predicate may be used to describe different types of object pairs (e.g. `car-near-street`, `person-near-building`), it is hard to recognize the predicate only based on the visual appearance. To exploit the visual manifestation across different object categories, we introduce a semantic embedding layer to integrate the category information of `subject` and `object`. Specifically, the proposed semantic embedding layer maps the object category into a feature embedding vector. Then we concatenate the embedding vectors of `subject` and `object` to learn the joint representations of the object pair through a fully connected layer. The learned joint representations are further fused with both visual and spatial features. Instead of learning the embeddings of object categories, we can also initialize the parameters of semantic embedding layer with the pre-trained word representations. In this paper, we introduce word2vec (Mikolov et al. 2013) as the off-the-shelf language model to acquire the word representations. In this way, the semantic relatedness between objects is well integrated into the learning procedure which especially benefits zero-shot visual relationship detection as shown in (Lu et al. 2016).

## Structural Ranking Loss

In this section, we propose to cascade the multi-cue based convolutional neural network with a structural ranking loss function. Compared with multi-class based method, the proposed method is no longer under the assumption of only single predicate existing in each object pair which can better facilitate the predicate co-occurrence. At the same time, we do not use the hard constraint to separate the relationship instances for being positive or negative which makes our method more robust and flexible than using the cross-entropy based loss for multi-label learning.

For an input image $x$, we first extract the feature representations of visual appearance cue, spatial location cue and semantic embedding cue for each relationship instance tuple $r = (s, p, o) \in \mathcal{R}$. The learned features combined with multiple cues are further concatenated and fused into a joint feature vector through one fully connected layer. We use $f(\cdot)$ to denote the above procedure. Therefore, the fused features for relationship instance $r = (s, p, o)$ can be denoted as $f(x, s, o)$. Then we formulate a compatibility function between the input image $x$ and the relationship instance $r$ as following:

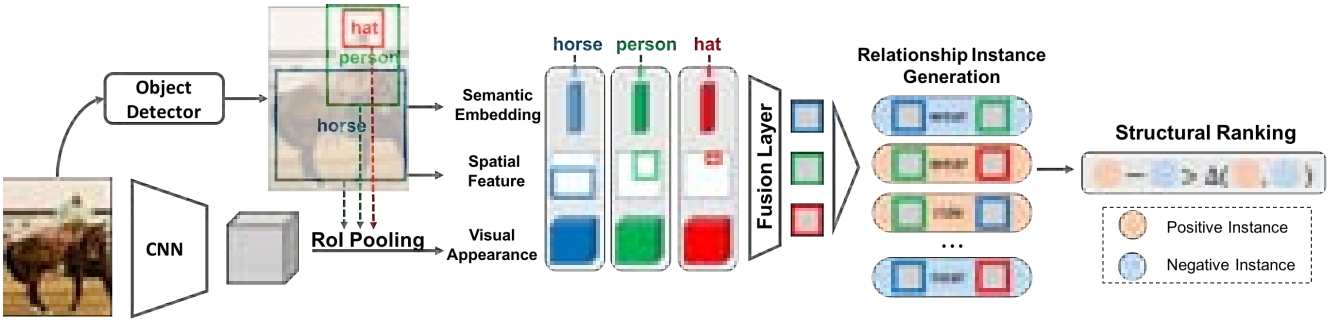


Figure 2: The proposed deep structural ranking framework for visual relationship detection. It contains a deep convolutional network which combines multiple cues to learn the representation for an input instance. A structural ranking loss is further cascaded with the network to enforce the annotated relationships with higher relevance score.

$$\Phi(x, r) = \Phi(x, \{s, p, o\}) = \mathbf{w}_p \top f(x, s, o), \quad (2)$$

where $\mathbf{w}_p$ denotes the parameters to be learned for $p^{th}$ `predicate`. For visual relationship detection, not all the related predicates between subject and object are annotated. However, the annotated predicates are usually more salient than the unannotated predicates. To address the incompleteness of annotations, we propose to minimize a structural ranking criterion for the input image x and its associated relationship instances $\mathcal{R}$:

$$\mathcal{L}(x) = \sum_{r \in \mathcal{R}} \sum_{r' \in \mathcal{R}'} [\triangle(r, r') + \Phi(x, r') - \Phi(x, r)]_+, \quad (3)$$

where $\mathcal{R}' = \{(s', p', o') | (s', o') \in \mathcal{P} \wedge p' \notin \mathcal{P}_{s',o'}\}$ is the set of all the relationship instances that are not manually annotated, $[\cdot]_+ = max(0, \cdot)$ denotes to only remain the positive part of input and $\triangle(\cdot, \cdot)$ is a margin that depend on the difference between the two input relationship instances.

We design $\triangle(\cdot, \cdot)$ to take the incompleteness problem of visual relationship detection into account. We formulate this margin between relationship instances via the prior probabilistic distribution of relationships conditioned on the subject and object pair. The assumption is if the prior probability is higher, the annotation of relationship instance is more likely to be missing from the annotations. Therefore, we define the following adaptive margin function:

$$\begin{aligned} \triangle(r, r') &= \triangle(\{s, p, o\}, \{s', p', o'\}) \\ &= 1 + P(p|c_s, c_o) - P(p'|c_{s'}, c_{o'}), \end{aligned} \quad (4)$$

where we use $c_s, c_o$ to denote the category label for subject and object respectively. With this adaptive margin, the incomplete annotation with a high prior probability to appear is less penalized.

## Inference

We use the object detector RCNN (Girshick 2015) to localize all the object proposals within an test image $x$ and acquire the category label and confidence score for each object proposal. Then, we perform non-maximum suppression (NMS) for every object category. We denote the set of object proposals as $\mathcal{B}$ and generate all the candidate object pairs as $\mathcal{P}^* = \{(s, o) | s \in \mathcal{B} \wedge o \in \mathcal{B} \wedge s \neq o\}$ which are the combination of all the detected objects.

Finally, we can acquire the compatibility score between the object pair $(s, o)$ and predicate $p$:

$$score_{\{s,p,o\}} = \Phi(x, \{s, p, o\}), \forall p, (s, o) \in \mathcal{P}^*. \quad (5)$$

By utilizing RoI pooling operation, all the object proposals can share the representations below the last convolutional layer which highly reduces the amount of computation. As the representation of each object proposal is used repeatedly across different object pairs, we can inference the problem even the number of candidate object pairs is fairly large. In our experiments, we found that integrating the above equation with the confident score from the object detector can give further performance gain. Therefore, we propose to model the conditional probabilistic prior by including the object confidence from the trained detector:

$$P(c_s, p, c_o | s, o) = P(p|c_s, c_o) P(c_s|s) P(c_o|o), \quad (6)$$

where $P(c_s|s)$ and $P(c_o|o)$ are the confidence of the subject and object acquired from the detection network. Then we combine the score from the proposed deep neural network and the prior confidence defined in Eqn. (5) and Eqn. (6) respectively as the final score of a relationship instance. According to the compatibility score, we can sort all the relationship instances $r = \{s, p, o\}$ and output the top instances as the predictions for the test image $x$.

## Experiments

We evaluate the proposed methods on two recently released datasets. Experimental results show that our proposed method can surpass the previous state-of-the-art on visual relationship detection. In addition, we investigate the effect of using different components for the proposed method. Finally, experiments on zero-shot relationship detection show that our proposed method can generalize well on the relationships which do not exist in the training data.

## Datasets and Experimental Settings

**VRD (Lu et al. 2016).** VRD (Visual Relationship Dataset) contains 5000 images with 100 object categories and 70 predicates. Totally, VRD contains 37,993 relation instances among 6,672 unique `subject-predicate-object` triplets. We use the default dataset split which contains 4,000 training images and 1,000 test images. There are 1,877 relationship triplets only appearing in the test set which are further used for zero-shot evaluations.

**VG (Krishna et al. 2016).** The annotations of original VG (Visual Genome) dataset are very noisy. Therefore, we use the cleaned up version (Zhang et al. 2017) by using official pruning of objects and relations. For example, young woman and lady are merged to woman. In summary, VG contains 99,658 images with 200 object categories and 100 predicates. There are totally 1,174,692 relation instances among 19,237 unique triplets. The default split contains 73,801 images for training and 25,857 images for testing.

In our experiments, we involve two relevant tasks to evaluate the proposed method: **Predicate detection** and **Relationship detection**. For Predicate detection, the task is to predict the correlated predicates given a pair of localized objects. For Relationship detection, the algorithm needs to firstly localize the objects appearing in the input image and then predict the relevant predicates between each pair of them. It evaluates both object detection and predicate detection. Following the original paper (Lu et al. 2016), we use Recall@50 (R@50) and Recall@100 (R@100) as the evaluation metrics. R@K compute the fraction of true positive relationships over the total relevant relationships among the top K confident predictions. Because of the incompleteness of annotations, mean average precision (mAP) is usually pessimistic as some of the relevant predictions may not have that particular ground truth. Therefore we use recall based evaluation metrics for comparison.

## Implementation Details

Our proposed deep structural ranking model use VGG16 as the base network. We use Adam optimizer to train the whole network and the learning rate is set to be 0.00001. During training, the first five convolutional layers of the base network are fixed without tuning. For the newly added layers, the learning rate is multiplied by 10 to accelerate the learning process. We train the proposed model for 5 epochs and divide the learning rate by a factor of 10 after the third epoch. Our implementations are based on the Pytorch deep learning framework on a single GeForce GTX TITAN X. For VRD dataset, the training time for one epoch (4000 training images) is around 10 minutes.

## Comparative Results

Firstly, we compared our proposed method with two baseline methods **JointCNN** and **JointBox**. JointCNN aims to predict the three components of a relationship instance jointly. For example, on VRD dataset, it has a 270 (100+100+70) way classification model to predict the subject, object and predicate labels. JointBox (Zhang et al. 2017) trains a softmax classifier that classifies joint bounding boxes of the subject and object into predicates. Beyond that, several state-of-the-art comparison methods are involved. **VR-V** and **VR-LP** are two variants of (Lu et al. 2016) which represent using only visual appearance and full model with language prior respectively. **VR-V** is a two-stage model which use R-CNN (Girshick 2015) to detect the objects and **VR-LP** further combines **VR-V** with word2vec language priors. **VTE** (Zhang et al. 2017) is a novel Visual Translation Embedding network designed for visual relation detection. It is an end-to-end and fully-convolutional architecture which use softmax loss function which only rewards the deterministically accurate predicates.**VRL** (Liang, Lee, and Xing 2017) is a visual relationship model based on variation-structured reinforcement learning which sequentially discover object relationships in the input image. **ViP-CNN** (Li, Ouyang, and Wang 2017) presents a convolutional neural network with Phrase-guided Message Passing Structure (PMPS) to establish the connection among relationship components and help the model consider the relationship learning problems jointly. Finally, we compare with the deep relational network **DR-Net** (Dai, Zhang, and Lin 2017) which develops dual spatial masks to represent the spatial configurations and exploits the statistical dependencies between objects and relationships. For our proposed method, we use the spatial mask as the spatial feature and also apply word2vec to acquire the category embedding.

From the results on both predicate and relationship detection in Table 1, we can see JointCNN performs the worst. This shows that network is hard to train with the supervision from three components simultaneously. Compared with JointCNN, JointBox achieves better performance as the learning task only aims to predict the predicate between the subject and object. From the performance gap between VR-V and VR-LP, we observe that integrating the category information and its language prior knowledge can improve the performance of only using visual appearance. Integrating the feature learning procedure with visual relationship detection learning task into a joint learning task can consistently improve the performance, as VTE, VRL, ViP-CNN and DR-Net achieve much higher Recall than VR-LP. Both VTE and ViP-CNN propose to simultaneously detect the objects and predict the relationship within a single deep network. However, this multi-task learning strategy does not show much performance gain as it increases the difficulty of model training. For example, it is hard to balance the contribution for each object function and the low-level features for object detection and relationship learning may not be shareable. Since both ViP-CNN and DR-Net leverage the dependencies between objects and relationships, DR-Net achieves better performance as its proposed spatial mask can exploit the spatial configurations more effectively.

On both learning tasks, the proposed method outperforms the recent state-of-the-art methods dramatically. For predicate detection, we improve the state-of-the-art by around 11% according to Recall@100 on both VRD and VG. For relationship detection, our proposed method further improve the Recall@100 by 2.4% on VRD dataset compared to the previous best performance. With the same spatial feature as DR-Net, our proposed method can outperform DR-Net ac-

| Methods | Predicate Det. | | Relationship Det. | |
|---|---|---|---|---|
| | R@50 | R@100 | R@50 | R@100 |
| JointCNN | 1.47 | 2.03 | 0.07 | 0.09 |
| JointBox | 25.78 | 25.78 | - | - |
| VR-V | 7.11 | 7.11 | 1.58 | 1.85 |
| VR-LP | 47.87 | 47.87 | 13.86 | 14.70 |
| VTE | 44.76 | 44.76 | 14.07 | 15.20 |
| VRL | - | - | 18.19 | 20.79 |
| ViP-CNN | - | - | 17.32 | 20.01 |
| DR-Net | 80.78 | 81.90 | 17.73 | 20.88 |
| Ours | **86.01** | **93.18** | **19.03** | **23.29** |

Table 1: Performances (%) on VRD dataset. "-" denotes the performance has not been reported in the original paper.

| Methods | Predicate Det. | |
|---|---|---|
| | R@50 | R@100 |
| JointBox | 46.59 | 46.77 |
| VTE | 62.63 | 62.87 |
| Ours | **69.06** | **74.37** |

Table 2: Predicate detection results (%) on VG dataset.

cording to all the evaluation metrics. This demonstrates the co-occurrence and incompleteness of the predicates can not be neglected for detecting the visual relationships. Comparing the predicate detection and relationship detection, we can find that the performance of object detector is a crucial factor for visual relationship detection. In order to correctly predict a relationship instance, the object detector must simultaneously localize the subject and object with 0.5 intersection over union (IoU) according to ground truth location. Even the state-of-the-art object detector can not achieve satisfactory performance.

## Component Analysis

Our proposed network contains multiple cues to learn the holistic representation of a relationship instance. In this section, we discuss on how they influence the final performance. We use $V_1$ to denote only using the union of the subject and object bounding boxes as the input of the network. In addition, $V_2$ combines the visual appearance features from both the union and the separate subject and object bounding boxes. For spatial location cue, we use $S_1$ to denote the scale-invariant relative location feature and the relative height/width feature between the subject and object. $S_2$ represents using the spatial masks as the input spatial location cue. $E$ means using the 300-D word2vec representations to initialize the semantic embedding layer. The experimental results are shown in Table 3.

For the comparison on visual appearance cue, $V_2$ can improve the performance of $V_1$ by 4% and 1% according to Recall@50 and Recall@100 respectively. From our perspective, the main reason is using the visual feature from the separate subject and object regions give more specific information since the union bounding box may contain other objects which is usually noisy for learning the visual relationship. By further adding the spatial location cue, the performance can be further improved as many of the predicates are used to describe the spatial information of the subject and object (e.g. sky above street). Both $S_1$ and $S_2$ can effectively model the spatial information to enhance visual relationship detection. On predicate detection, $S_1$ achieves better performance than $S_2$. However, the performance gain is not very obvious on relationship detection as the location of the detected objects are as accurate as the ground-truth annotations. This further shows that spatial masks are more robust than the coordinates based spatial location features. Combining the visual appearance and semantic embedding also improves the performance as $V_2E$ dramatically outperforms $V_2$. Since the visual appearance is limited and varies within the same object category, it is usually hard to infer the category labels for subject and object based on their visual appearance. We further show the proposed method which combines all the three cues can achieve the best performance. However, the performance gain of spatial location cue and semantic embedding cue are not complementary since $V_2SE$ only improve the recall of $V_2E$ by around 2%. This is mainly because the location information is also useful to infer the category label of an object and vice versa. $S_2$ can achieve slightly better performance than $S_1$ when combined with visual appearance and semantic embedding. Finally, we show some qualitative results in Fig. 3. The top5 predications are shown for each image. By using spatial and semantic information, our proposed method can avoid predicting the relationships which are not reasonable (e.g. `person-wear-bag, shelf-on-laptop`).

| | Predicate Det. | | Relationship Det. | |
|---|---|---|---|---|
| | R@50 | R@100 | R@50 | R@100 |
| Comparison on different feature | | | | |
| $V_1$ | 66.14 | 82.53 | 17.71 | 21.91 |
| $V_2$ | 70.58 | 83.58 | 17.66 | 22.03 |
| $V_2S_1$ | 79.53 | 89.16 | 18.72 | 22.83 |
| $V_2S_2$ | 78.14 | 88.27 | 18.86 | 22.49 |
| $V_2E$ | 84.38 | 91.81 | 18.19 | 22.56 |
| $V_2S_1E$ | **86.32** | 93.03 | 18.91 | 23.24 |
| $V_2S_2E$ | 86.01 | **93.18** | **19.03** | **23.29** |
| Comparison on different loss function | | | | |
| $V_2S_2E$+$CE$ | 79.22 | 89.15 | 17.95 | 21.89 |
| $V_2S_2E$+$SM$ | 82.06 | 90.59 | 18.02 | 22.09 |

Table 3: Component Analysis (%) on VRD dataset.

We further conduct experiments on comparing the proposed structural ranking loss with cross-entropy and softmax loss function. Cross-entropy loss is a widely used objective function for multi-label learning which can well facilitate the co-occurrence of predicates. Softmax loss function is adopted as the objective function in (Dai, Zhang, and Lin 2017; Li et al. 2017; Zhang et al. 2017) since most of the object pairs have only one predicate annotation. We use the same input features $V_2S_2E$ which combine multiple cues for fair comparison. As shown in the bottom of Table 3, using either cross-entropy or softmax loss function decreases the

Figure 3: Qualitative Results of the proposed methods. The white box denotes the results of using only visual feature while the grey box denotes the result of using multiple cues. The correct predictions are shown in green while the wrong predictions are in red. The correct predictions without annotating is shown in blue.

| Methods | Predicate Det. | | Relationship Det. | |
|---|---|---|---|---|
| | R@50 | R@100 | R@50 | R@100 |
| VR-V | 3.52 | 3.52 | 0.67 | 0.78 |
| VR-LP | 8.45 | 8.45 | 3.13 | 3.52 |
| VTE | - | - | 1.71 | 2.14 |
| VRL | - | - | **7.94** | 8.52 |
| Ours | **60.90** | **79.81** | 5.25 | **9.20** |

Table 4: Zero-shot Performances (%) on VRD dataset. The methods without reporting the performance on zero-shot setting are excluded from comparison.

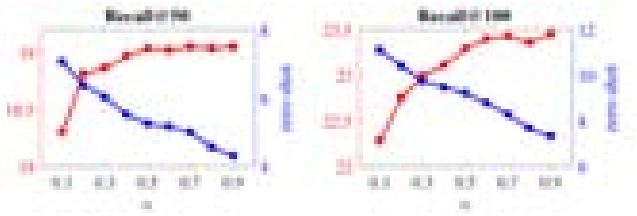

Figure 4: Experiments on the influence of prior knowledge on VRD dataset. Red points indicate the performance of visual relationship detection while the blue points indicate performing using the zero-shot setting.

performance on both tasks.

### Zero-shot Visual Relationship Detection

Due to the long tail distribution of relationships, it is hard to collect the images for all the possible relationships. So it is crucial for a model to have the generalizability on detecting zero-shot relationships. On VRD dataset, it contains 1,877 relationships that only exist on the test set. Even though `elephant-stand on-street` never occurs on the training set, we can use the correlated relationships (e.g. `dog-stand on-street`) to infer the unseen relationship. The performance on zero-shot predicate and relationship detection is reported in Table 4. Our proposed method can achieve better performance on detecting zero-shot relationships especially on predicate detection.

We further explore the influence of using the inference strategy defined in Eqn. (6) on zero-shot relationship detection. Since the conditional probability of zero-shot relationship is zero, the performance will be decreased when directly using the conditional probability calculated on the training set. Therefore, we proposed to formulate the conditional probability as:

$$P(p|c_s, c_o) = \alpha P_{\mathcal{D}}(p|c_s, c_o) + (1-\alpha)\tilde{P}(p|c_s, c_o), \quad (7)$$

where $P_{\mathcal{D}}(\cdot)$ is calculated based on the training data, $\tilde{P}(\cdot)$ is the prior distribution of the relationships existing in the real world and $\alpha \in [0, 1]$ is used to balance the influence of the above two distributions. Specifically, $P_{\mathcal{D}}(p|c_s, c_o)$ is measured by the conditional probability of predicate p given the category labels of its subject and object. In this paper, we simple set $\tilde{P}(p|c_s, c_o)$ to be a uniform distribution. However, the true prior knowledge can be collected from some external data (Chao et al. 2015b). We leave it as the future work. As shown in Fig. 4, the performance of zero-shot relationship detection trends to decrease when the value of $\alpha$ is increased. Without further tuning, we set $\alpha = 0.5$ for all the experiments .

## Conclusion

In this paper, we propose a deep structural ranking framework for visual relationship detection. Our proposed model is based on the convolutional neural network which combines multiple cues. To facilitate the co-occurrence and incompleteness of visual relationships, we proposed a structure ranking loss which enforces the annotated relationships to have higher relevance score since the annotated relationships are usually more salient. We explore different ways to effectively leverage auxiliary information. Experimental results show that our proposed method outperform the state-of-the-art according to both predicate detection and relationship detection on the widely used datasets.

## Acknowledgments

Research supported by China Scholarship Council (No. 201604910935), Natural Science Foundation of China (No. 61390515), NSERC and the Canada Research Chairs program.

## References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433.

Atzmon, Y.; Berant, J.; Kezami, V.; Globerson, A.; and Chechik, G. 2016. Learning to generalize to new compositions in image understanding. *arXiv*.

Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015a. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 1017–1025.

Chao, Y.-W.; Wang, Z.; Mihalcea, R.; and Deng, J. 2015b. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4259–4267.

Dai, B.; Zhang, Y.; and Lin, D. 2017. Detecting visual relationships with deep relational networks. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1778–1785. IEEE.

Farhadi, A.; Hejrati, M.; Sadeghi, M. A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; and Forsyth, D. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, 15–29.

Galleguillos, C.; Rabinovich, A.; and Belongie, S. 2008. Object categorization using co-occurrence, location and appearance. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8. IEEE.

Gella, S.; Lapata, M.; and Keller, F. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of NAACL-HLT*, 182–192.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Gould, S.; Rodgers, J.; Cohen, D.; Elidan, G.; and Koller, D. 2008. Multi-class segmentation with relative location prior. *International Journal of Computer Vision* 80(3):300–316.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A. C.; and Berg, T. L. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(12):2891–2903.

Li, Y.; Ouyang, W.; Wang, X.; and Tang, X. 2017. Vip-cnn: Visual phrase guided convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Y.; Ouyang, W.; and Wang, X. 2017. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection.

Liang, K.; Chang, H.; Shan, S.; and Chen, X. 2015. A unified multiplicative framework for attribute learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2506–2514.

Liang, X.; Lee, L.; and Xing, E. P. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 852–869. Springer.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Ramanathan, V.; Li, C.; Deng, J.; Han, W.; Li, Z.; Gu, K.; Song, Y.; Bengio, S.; Rosenberg, C.; and Fei-Fei, L. 2015. Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1100–1109.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Sadeghi, M. A., and Farhadi, A. 2011. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1745–1752. IEEE.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of The 32nd International Conference on Machine Learning*, 2048–2057.

Yao, B., and Fei-Fei, L. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 17–24. IEEE.

Zhang, H.; Kyaw, Z.; Chang, S.-F.; and Chua, T.-S. 2017. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.