

# Hierarchical Visual Relationship Detection

Xu Sun<sup>1,3</sup>, Yuan Zi<sup>1</sup>, Tongwei Ren<sup>1,3,\*</sup>, Jinhui Tang<sup>2</sup>, and Gangshan Wu<sup>1</sup>

<sup>1</sup> State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China  
<sup>2</sup> School of Computer Science, Nanjing University of Science and Technology, Nanjing, China

<sup>3</sup> Shenzhen Research Institute of Nanjing University, Shenzhen, China

{sunx, ziyuan}@smail.nju.edu.cn, rentw@nju.edu.cn, jinhuitang@njust.edu.cn, gswu@nju.edu.cn

## ABSTRACT

Acting as a bridge between vision and language, visual relationship detection (VRD) aims to represent objects and their interactions in an image with several relationship triplets. Nevertheless, the conventional VRD task shows little consideration for the penalization of incorrect relationship predictions, which in turn undermines its support for image understanding applications. In this paper, we propose a novel VRD task named *hierarchical visual relationship detection* (HVRD), which encourages predictions with abstract yet compatible relationship triplets when the confidence level of the specific image content is relatively low. Meanwhile, HVRD can handle the inevitable ambiguity of groundtruth annotation in VRD. Based on this, we propose a HVRD method, consisting of hierarchical object detection and hierarchical predicate detection. It can effectively detect the hierarchical visual relationships by exploiting both object concept hierarchy and predicate concept hierarchy with order embedding. We also propose the first datasets for HVRD evaluation, H-VRD and H-VG, by expanding the relationship category spaces of VRD and VG datasets to hierarchical ones respectively. The experimental results show that our method is superior to the state-of-the-art baselines.

## CCS CONCEPTS

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

## KEYWORDS

Visual relationship detection; hierarchical visual relationship; hierarchical predicate detection; order embedding

### ACM Reference Format:


Xu Sun<sup>1,3</sup>, Yuan Zi<sup>1</sup>, Tongwei Ren<sup>1,3,\*</sup>, Jinhui Tang<sup>2</sup>, and Gangshan Wu<sup>1</sup>. 2019. Hierarchical Visual Relationship Detection. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*, October 21–25, 2019, Nice, France. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3343031.3350921>

## 1 INTRODUCTION

As a bridge between vision and language, visual relationship detection (VRD) aims to represent objects and their interactions in images and videos with relationship triplets in the format of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
MM '19, October 21–25, 2019, Nice, France

© 2019 Association for Computing Machinery.  
ACM ISBN 978-1-4503-6889-6/19/10...\$15.00  
<https://doi.org/10.1145/3343031.3350921>



relationship triplets	VRD	HVRD
<person, stand next to, elephant>	correct	1.00
<person, near, elephant>	wrong	0.87
<person, near, animal>	wrong	0.76
<entity, interact, entity>	wrong	0.15
<person, watch, elephant>	wrong	0.00

<person, stand next to, elephant>

Figure 1: An example of HVRD as compared to VRD.

<subject, predicate, object> [20]. It serves as a foundation of many multimedia applications, such as image/video captioning [3, 25, 36] and visual question answering [1].

The application of VRD, is built on the prerequisite that the detected visual relationships should be correct. The reason can be rather straightforward: the incorrect relationship predictions may conflict with the correct ones and prevent the applications from deriving useful information from VRD results. However, the *recall* criterion used in conventional VRD evaluation cannot penalize the incorrect relationships effectively, due to the fact that the recall criterion only evaluates the correct relationships in the VRD results and shows little consideration for the incorrect ones.

A widely-accepted reason for using recall criterion is that there is inevitable ambiguity in the groundtruth annotation of VRD datasets, *i.e.*, the annotators may miss some visual relationships with low saliency and use inconsistent words to represent the same subject/object or predicate. This, however, cannot justify using recall criterion without penalizing the incorrect ones in the VRD results. In other words, we may find a better solution for VRD evaluation if we gain some insights into VRD groundtruth annotation. As a matter of fact, the crucial visual relationships are always annotated, while correct yet unimportant ones may go unnoticed. And it follows that these missed relationship instances should be ranked lower than the annotated ones in terms of importance while prediction. Therefore, it is reasonable to penalize the abovementioned visual relationships of lower importance while evaluation. Moreover, the annotators may use inconsistent words to describe the same subject/object or predicate in annotation, such as “elephant” vs. “animal” and “stand next to” vs. “near”. Those words are all correct descriptions, despite their differences in specificity. It will also be useful to obtain the abstract yet correct relationship instances if the specific ones cannot be detected.

Based on the above observation, we propose a novel VRD task, named *hierarchical visual relationship detection* (HVRD). To tackle the ambiguity of groundtruth annotation in VRD, we establish the connections among relationship triplets by constructing hierarchical concept structures on both subject/object and predicate categories. The relationship concept hierarchy consists of object concept

hierarchy and predicate concept hierarchy, both of which are single root trees and express semantic generalization/specialization relation among concepts at different semantic levels. The closer a concept is to the root, the more abstract the concept is and vice versa. Once a relationship instance is detected, the compatibility between the predicted triplet and annotated ones in groundtruth will be evaluated. For instance, a relationship triplet  $A$  is compatible with a relationship triplet  $B$  if each element in  $A$ , *i.e.*, subject, predicate or object, is the same or ancestor node in the concept hierarchy to the one in  $B$ , and the bounding boxes of both their subjects and objects have high IoU. If the detected relationship triplet is compatible with an annotated one in groundtruth, its score is calculated according to their semantic similarity. Figure 1 gives an example of HVRD as compared to the conventional VRD. The correct relationships, such as  $\langle \text{person, near, animal} \rangle$ , are considered useful in HVRD even they are not specific enough; and the incorrect relationships, such as  $\langle \text{person, watch, elephant} \rangle$ , are penalized to avoid the misunderstanding of image content. Furthermore, to avoid resulting in too many incorrect relationship triplets, we propose a new adaptive evaluation criteria,  $Recall@N(k = \alpha)$ , which only considers top- $\alpha$  detected relationship triplets between an object pair instead of arbitrary number of that, within  $N$  relationship detection instances per image. Here,  $\alpha$  is an adaptive value which is the number of annotated relationship triplets of each object pair.

We propose a HVRD method, consisting of hierarchical object detection module and hierarchical predicate detection module. It can effectively detect the hierarchical visual relationships by exploiting both object and predicate concept hierarchies with order embedding [31]. Moreover, it is worth noting that to date, there is no available dataset for HVRD, though VRD has several datasets including Visual Relationship Dataset [20] and Visual Genome [15]. Therefore, by expanding the category spaces of VRD and VG datasets to hierarchical ones, we generate the first datasets for HVRD evaluation, including Hierarchical Visual Relationship Dataset (H-VRD) and Hierarchical Visual Genome (H-VG).

The main contributions of this paper include:

- (1) The proposal of a novel HVRD task that aims to improve the correctness of the detected relationship triplets.
- (2) The generation of two HVRD datasets for evaluation, including H-VRD and H-VG.
- (3) The proposal of an order embedding based HVRD method which successfully encodes the knowledge contained in concept hierarchies.

## 2 RELATED WORK

### 2.1 Visual Relationship Detection

Visual relationship detection has been widely used by various visual understanding applications, including image retrieval [2, 13], image captioning [36], scene graph generation [16, 33, 34, 39] and visual question answering [1]. Within the recent few years, a number of visual relationship detection methods have emerged [4, 14, 17, 18, 20, 35, 37, 40–42, 44–46].

Lu *et al.* formally proposed VRD task on static image with the first VRD dataset [20]. They also developed the first VRD method combining deep convolutional neural network and language prior. Yu *et al.* applied knowledge distillation in VRD for the first time [38].

The method distills external linguistic knowledge extracted from large scale textual data. Zhou *et al.* proposed an attention based model, which successfully fuses language and spatial information with CNN feature and achieves great performance [44]. Recently, Yin *et al.* proposed an effective framework, Zoom-Net, with a well designed spatial-context-appearance module [37]. They also exploited the structural knowledge contained in semantic concept hierarchies to improve the visual feature. Nevertheless, Zoom-Net still focuses on conventional VRD.

To apply visual relationship detection in real-world scenario, a large scale dataset, Visual Genome [15] is constructed, which contains more than 100K images and covers a raft of object and predicate categories. It poses a great challenge to the existing methods. Zhang *et al.* developed a large-scale visual relationship detection method to tackle the overlarge category space and extremely imbalanced data, by embedding visual and semantic features into a shared space [42]. The experiment results shows the effectiveness of their method.

It is also worth mentioning that Shang *et al.* formalized video visual relationship detection task (VidVRD) and constructed a dataset for evaluation [29]. Moreover, they proposed the first VidVRD framework with the capability of predicting dynamic visual relationships in video. It adopts a bottom-up strategy and utilizes iDT feature [32] for dynamic relationship recognition.

### 2.2 Hierarchical Object Detection

The exploration of hierarchical visual recognition has existed for a long time. Deng *et al.* constructed an essential large scale hierarchical image dataset, Imagenet [8], the object categories in which are associated with the semantic hierarchy of WordNet [23]. With the structural knowledge within WordNet, large scale and open-ended visual recognition can be promoted.

Encouraged by Imagenet, researchers has proposed numerous hierarchical object detection methods [6, 7, 9–11, 24, 26]. DARTS [9] is a well designed framework for hierarchical object recognition, which is optimized by trade-off between specificity and accuracy. Ordonez *et al.* [24] proposed an entity level object recognition method, which involves “naturalness” of expression, mined from enormous amount of textual data on the Internet. Recently, to tackle object detection over 9,000 categories, Redmon *et al.* [26] constructed a new framework, YOLO9000, which is able to jointly training on classification and detection datasets combined with the concept hierarchy of WordNet. In this way, the model may solve the zero-shot classification problem.

## 3 METHOD

HVRD uses two-dimensional subject/object and predicate concept hierarchies to tackle the inconsistency and diversity of relationship triplets. To make the best of structural knowledge contained in concept hierarchies, we propose an order embedding based HVRD method with multi-modal feature. Our framework is inspired by a fact that the generalization/specialization relation among hierarchical concepts is a typical partial order relation, which possesses reflexivity, transitivity and antisymmetry. Therefore, we use two high dimensional order embedding vector spaces to model the relations among the hierarchical subject/object and

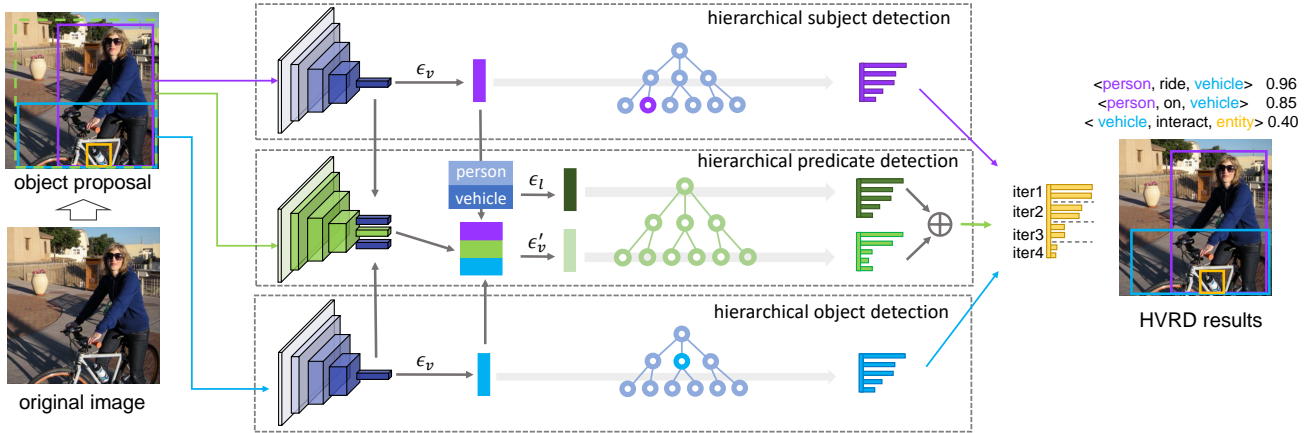


Figure 2: An overview of the proposed HVRD method.  $\epsilon_v$ ,  $\epsilon'_v$  and  $\epsilon_l$  demote visual and language embedding functions.

predicate concepts respectively. We obtain two sets of embedding vectors to represent the two kinds of concepts, which effectively preserve the hierarchical structural knowledge. Within the concept embedding spaces, we construct a HVRD framework consisting of hierarchical object detection module and hierarchical predicate detection module. Figure 2 shows the overview of the proposed method.

### 3.1 Hierarchical Concept Embedding

According to previous research on VRD [20, 40, 42], knowledge transfer is critical to visual relationship detection in dealing with the large scale category space and extremely sparse data distribution. To this end, word embedding [21, 22] becomes a widely used technique in VRD as knowledge transfer medium, which encodes words with a set of high-dimensional vectors and expresses semantic relation in Euclidean space. However, an obvious defect of the existing word embedding lies in its inability of encoding information with complex spatial structure such as taxonomy.

The concept hierarchies provided by HVRD datasets are represented with tree structure, which is beyond the capability of word embedding. So we utilize order embedding [31] to model the unidirectional generalization/specialization relation among the concepts in different abstraction levels. Here, we define specialization relation,  $\mu_1 \geq \mu_2$ , where  $\mu_2$  is a specialized concept of  $\mu_1$  or the same concept as  $\mu_1$ . Then we have  $\mu_1 \geq \mu_1$  (i.e. reflexivity);  $\mu_1 \geq \mu_2 \Rightarrow \mu_2 \not\geq \mu_1$  (i.e. antisymmetry);  $\mu_1 \geq \mu_2$ ,  $\mu_2 \geq \mu_3 \Rightarrow \mu_1 \geq \mu_3$  (i.e. transitivity). Furthermore, we use  $T_\mu = \{\mu' : \mu' \geq \mu\}$  to indicate order transitive closure of concept  $\mu$ , which contains  $\mu$  and all the generalized concepts of  $\mu$ .

We construct two high-dimensional positive embedding spaces  $\mathbb{R}_+^N$  and  $\mathbb{R}_+^M$  for subject/object and predicate concept hierarchies in the same fashion, where  $N$  is 600 and  $M$  is 300 in our experiments. We define order distance  $\mathcal{D}$  between different concept embedding vectors as follows [31]:

$$\mathcal{D}(\delta(\mu_1), \delta(\mu_2)) = \|\max(0, \delta(\mu_1) - \delta(\mu_2))\|_2, \quad (1)$$

where  $\delta(\cdot)$  is order embedding function;  $\|\cdot\|_2$  is L2 norm. If  $\mu_1 \geq \mu_2$ ,  $\mathcal{D}(\delta(\mu_1), \delta(\mu_2))$  is expected to be zero (minimum). Otherwise, it is expected to be a positive value larger than a threshold  $\beta$ . In optimizing phase, we use a lookup table to embed the concepts

with pairwise max-margin loss as follows [31]:

$$\mathcal{L}_c((\mu, \sigma), (\tilde{\mu}, \tilde{\sigma})) = \mathcal{D}(\delta(\mu), \delta(\sigma)) + \max\{0, \beta - \mathcal{D}(\delta(\tilde{\mu}), \delta(\tilde{\sigma}))\}, \quad (2)$$

where  $(\mu, \sigma)$  and  $(\tilde{\mu}, \tilde{\sigma})$  are different concept pairs in the same concept hierarchy and  $\mu \geq \sigma$ ,  $\tilde{\mu} \not\geq \tilde{\sigma}$ ;  $\beta$  is margin threshold, which equals 1 in our experiments.

### 3.2 Hierarchical Object Detection

Object detection is an essential part of visual relationship detection, whose outputs are used to generate <subject, object> pairs as relationship candidates. And it follows that object detection performance determines the ceiling of relationship detection performance. Existing VRD methods may predict multiple categories for the same object proposal, i.e. category agnostic bounding box, to decrease false negative rate and increase the diversity of relationship candidates, regardless of the potential incompatible predictions. The aforementioned problem could be found in predicate detection as well, which compromises the applicability of VRD. To tackle this problem, we propose a novel hierarchical object detection model, Hier-RCNN, using Fast-RCNN [27] as backbone, and making full use of object concept hierarchies in HVRD datasets. Inspired by [9], we trade off specificity for accuracy, which is perceived as more important by VRD applications. Hier-RCNN is an order embedding based object detection model which encodes visual information with embedding vector, which implicitly represents internal connections among hierarchical object concepts.

**Visual embedding.** We replace the classifier of Fast-RCNN [12] with an embedding network, which is a two-layer fully-connected neural network with ReLU activation, to project CNN feature into object concept embedding space. In training phase, the visual object instances are treated as specialization of their annotated categories. Order distance  $\mathcal{D}$  between visual embedding vector and annotated concept embedding vector is minimized. We adopt a variant of softmax loss function for optimization as follows [43]:

$$\mathcal{L}_e(\mu_g, \mathbf{x}) = -\log \frac{e^{-\mathcal{D}(\delta(\mu_g), \epsilon_v(\mathbf{x}))}}{e^{-\mathcal{D}(\delta(\mu_g), \epsilon_v(\mathbf{x}))} + \sum_{\mu'_g \notin T_{\mu_g}} e^{-\mathcal{D}(\delta(\mu'_g), \epsilon_v(\mathbf{x}))}}, \quad (3)$$

where  $\mathbf{x}$  is the output of fc7 layer in VGG16 [30];  $\mu_g$  is the manual annotated object concept;  $\epsilon_v(\cdot)$  is the visual embedding network;  $\mu'_g$  is incompatible concept to  $\mu_g$ . It is worth noting that only the most specific concepts are used as labels in training data while the abstract concepts are learned without extra overhead. Due to the transitivity of partial order relation, one training sample with specific concept annotation is equivalent to a set of samples with generalized concepts, which saves a considerable amount of computation cost.

**Greedy inference.** With the visual embedding network, we project visual features into concept embedding space and measure the order distances between the visual embedding vector and all concept embedding vectors. Then the classification problem can be formed as concept retrieval [43]. Ideally, the distances to positive concepts are smaller than those to negative concepts. Top- $k$  prediction is an intuitive and widely used strategy in retrieval. However, a small number of high-accuracy object detections are required to limit the number of relationship candidates. To this end, we adopt a top-down greedy inferring procedure on the concept hierarchy based on the distances, to predict single concept by trade-off between specificity and accuracy. Specifically, we calculate conditional probabilities  $P(\mu|\mu^p)$  for each non-root node  $\mu$  in concept hierarchy :

$$P(\mu|\mu^p) = \frac{e^{-\mathcal{D}(\delta(\mu), \epsilon_v(\mathbf{x}))}}{\sum_{\mu^c \in C_{\mu^p}} e^{-\mathcal{D}(\delta(\mu^c), \epsilon_v(\mathbf{x}))}}, \quad (4)$$

where  $\mu^p$  is the directly generalized concept of  $\mu$ ;  $C_{\mu^p}$  is a set of directly specialized concepts of  $\mu^p$ , which are siblings. We determine each move along the top-down inferring path on concept hierarchy by iteratively selecting concept nodes with the largest conditional probabilities comparing to their siblings. Then we obtain a set of prediction candidates  $\Phi$  with different abstraction like {"entity", "covering", ..., "pants", "jeans"}. To determine single concept prediction from  $\Phi$ , we calculate semantic specificity reward  $Q(\mu)$  and conditional probability based information entropy  $\mathcal{E}(\mu^p)$  for all  $\mu$  in  $\Delta$  respectively:

$$Q(\mu) = \frac{|T_\mu|}{\max_{\tilde{\mu} \in D_\mu} (|T_{\tilde{\mu}}|)}, \quad (5)$$

$$\mathcal{E}(\mu^p) = - \sum_{\mu \in C_{\mu^p}} P(\mu|\mu^p) \log(P(\mu|\mu^p)), \quad (6)$$

where  $D_\mu$  is the concept set containing  $\mu$  and all descendant concepts of  $\mu$ ;  $Q(\mu)$  is positively associated with semantic specificity of  $\mu$ ;  $\mathcal{E}(\mu^p)$  indicates risk of choosing  $\mu$  at split  $\mu^p$ . We choose concept  $\tilde{\mu}$  as the single concept prediction result from  $\Phi$  by balancing the risk and reward as follows:

$$\tilde{\mu} = \arg \max_{\mu} (Q(\mu) \cdot (1 - \mathcal{E}(\mu^p))), \forall \mu \in \Phi. \quad (7)$$

The confidence of concept prediction  $\tilde{\mu}$  is regularized as:

$$P(\tilde{\mu}|\mathbf{x}) = \frac{1}{1 + \mathcal{D}(\delta(\tilde{\mu}), \epsilon_v(\mathbf{x}))}, \quad (8)$$

which ranges in (0, 1] and used to calculate the confidence of relationship triplet predictions in Equation (11). The proposed top-down greedy inferring procedure is of great importance in its attempt to take both accuracy and specificity into consideration.

### 3.3 Hierarchical Predicate Detection

Hierarchical predicate detection module predicts predicates between the detected objects. The main differences between predicate detection and object detection include: (1) predicates between objects provide richer semantic information than single objects, requiring more comprehensive features; (2) multi-predicate between the same object pair is common. Taking the differences into consideration, we propose a hierarchical predicate detection method adopting similar paradigm to Hier-RCNN, exploiting the compatibility among hierarchical predicate concepts. The proposed method consists of a visual embedding stream and a language embedding stream, and combines implicit visual cue and explicit language prior. Furthermore, with an accuracy oriented sorting strategy, *i.e.* ordered instance sorting, accuracy is effectively improved while recall of predictions is guaranteed.

**Visual embedding.** The basic visual feature we utilize is ROI-pooled CNN feature [12] extracted from relation phrase [28], *i.e.* union region of subject and object regions in image, which is widely used by VRD methods. However, this feature is unable to effectively capture the context information. To provide richer visual cues, we adopt a two-stage feature fusion, extended from [42]. In the first stage, we use two independent CNNs to extract visual features ( $\mathbf{x}_s$ ,  $\mathbf{x}_o$ ,  $\mathbf{x}_p$ ) for subject region, object region, and relation phrase respectively instead of a shared network. This design delves into the fact that appearance of individual object is vastly different from that of relation phrase. We generate a (4096 × 3)-dimensional feature vector  $\mathbf{f}$  by concatenating the raw features ( $\mathbf{x}_s$ ,  $\mathbf{x}_o$ ,  $\mathbf{x}_p$ ). In this way, local and global visual information complement each other. Then we feed  $\mathbf{f}$  into two fully-connected layers with ReLU activation to generate low level hidden feature  $\mathbf{h}$  which is a 600-dimensional vector. In the second stage, we further concatenate subject and object visual embedding vectors  $\epsilon_v(\mathbf{x}_s)$  and  $\epsilon_v(\mathbf{x}_o)$  with  $\mathbf{h}$  to generate a implicit semantic and visual combined feature  $\tilde{\mathbf{x}}$ . At last, we project  $\tilde{\mathbf{x}}$  into predicate concept embedding space with an embedding network  $\epsilon'_v(\cdot)$  whose structure is same as  $\epsilon_v(\cdot)$ . The visual embedding vector encodes various comprehensive yet implicit features. Explicit information, on the other hand, is generated via a language embedding stream as a complement.

**Language embedding.** Various language cues have been utilized by existing VRD methods. The most intuitive language prior is conditional probability based on statistics of training data, which only captures the most frequently recurring relationship triplets without any deduction capability. A more comprehensive idea is to infer unseen relationship triplets by knowledge transfer from the ones already appeared in training data with the help of word embedding [20]. Widely used word embeddings are trained with large scale textual data on the Internet, in which only unstructured and sparse knowledge is contained. However, HVRD datasets provide structural text data, *i.e.* object and predicate concept hierarchies, which can serve VRD better. The problem mentioned above may be eliminated by extracting and leveraging the structural knowledge from the concept hierarchies of HVRD datasets.

Our model encodes the structural knowledge provided by concept hierarchies with order embedding vectors, which significantly improve the performance of our method in HVRD tasks and is validated by the component analysis in Section 4.4. Specifically,

We concatenate subject and object concept embedding vectors as language context feature and project it into the predicate concept embedding space with a network  $\epsilon_l(\cdot)$ , which is similar to  $\epsilon_v(\cdot)$  and trained with the same loss function as in Section 3.2. We measure the distances between language embedding vector and all predicate concept embedding vectors. Finally, we combine the distances measured by visual and language streams in a linear manner and obtain the confidence scores for each predicate concepts:

$$S(\mu_p, \tilde{\mathbf{x}}, \mathbf{y}) = \gamma \cdot \mathcal{D}(\delta(\mu_p), \epsilon'_v(\tilde{\mathbf{x}})) + (1-\gamma) \cdot \mathcal{D}(\delta(\mu_p), \epsilon_l(\mathbf{y})), \quad (9)$$

$$P(\mu_p | \tilde{\mathbf{x}}, \mathbf{y}) = \frac{1}{1 + S(\mu_p, \tilde{\mathbf{x}}, \mathbf{y})}, \quad (10)$$

where  $\mu_p$  is predicate concept;  $\tilde{\mathbf{x}}$  is the visual feature mentioned in Section 3.3;  $\mathbf{y}$  is language feature;  $\gamma$  is a parameter, which is 0.3 in our experiments. Triplet score is calculated as:

$$P(\mu_s, \mu_p, \mu_o | \mathbf{x}_s, \mathbf{x}_o, \tilde{\mathbf{x}}, \mathbf{y}) = P(\mu_s | \mathbf{x}_s) \cdot P(\mu_p | \tilde{\mathbf{x}}, \mathbf{y}) \cdot P(\mu_o | \mathbf{x}_o). \quad (11)$$

**Ordered instance sorting.** With confidence scores corresponding to all predicate concepts, we apply the greedy inference procedure introduced in Section 3.2 to iteratively generate  $n$  predicate predictions for each object pair.  $n$  is the maximum of predicate predictions, which is equal to  $|C_{\mu_r}|$ , where  $\mu_r$  is the root concept. For each iteration, we collect all the chosen concepts at each split along the inferring path, *i.e.* the concept candidate set  $\Phi$ , and set the status of concepts in  $\Phi$  as visited. In a new inference iteration, the visited concepts are eliminated. In the relationship instance sorting phase, we split the relationship instances into  $n$  batches according to the index of iteration from which their predicates are chosen. The batch corresponding to early iteration is ranked near the top. Within each batch, the relationship instances are ranked according to their confidence scores. The design logic is that we place accuracy first and iteratively supplement predicates for each object pair to improve diversity. In this way, false positive rate is effectively controlled.

## 4 EXPERIMENTS

### 4.1 Tasks

As defined in Section 1, the input of HVRD is a given image, and its output is a list of relationship instances. A detected relationship instance is judged as correct if the detected subject and object boxes spatially hit a groundtruth relationship boxes respectively, *i.e.* the IoUs exceed a threshold, and predicted relationship triplet is compatible to the groundtruth relationship triplet associated with the hit subject and object boxes. The threshold for IoU is set as 0.5 in our experiments, the same as VRD.

Considering object detection’s heated status, especially in open-ended category space, we conduct our experiments in another task called *hierarchical predicate detection* (HPD) to emphasize the effectiveness in object interaction recognition. Predicate detection has been widely used in the existing works of VRD [4, 14, 17, 20, 35, 37, 41, 45]. The input of predicate detection is a pair of localized objects with their categories, and its output is the predicate(s) between the objects.

### 4.2 Evaluation Criteria

The common evaluation criterion for VRD task is  $Recall@N(k=m)$ . Here,  $N$  denotes the maximal number of relationship instances allowed to return on the whole image;  $m$  is a fixed value, which is set to 1, *i.e.*, only one relationship instance for a pair of objects allowed to return, or is set to the number of all the possible predicates, *e.g.*, on VRD dataset [20],  $k=70$ . However, in VRD task,  $m$  as a fixed value is not a suitable option, because the number of the annotated relationship instances on a pair of objects can be arbitrary. If  $m$  is defaulted to 1, it is too rigid to evaluate the performance of VRD methods to explore multiple relationship instances on a pair of objects; if  $m$  is set to present the number of all the possible predicates, it may be so flexible that incorrect relationship instances are encouraged to return to increase recall. Hence, we use  $k=\alpha$  in our experiments, here  $\alpha$  is an adaptive number for each pair of objects which is equal to the number of the annotated relationship instances on the object pair. With the adaptive  $\alpha$ , the evaluated VRD methods are encouraged to generate more accurate relationship instances and to guarantee high recall at the same time.

SGGen+ is an improved evaluation criterion proposed by Yang *et al* [34], based on an observation that minor mistakes in object recognition will lead to severe punishment in conventional VRD evaluation. However, this problem can be solved naturally with the assistance of semantic hierarchy in HVRD task settings. We define a new evaluation criterion named  $H-Recall@N(k=\alpha)$ , which adopts a soft judgment strategy. To each annotated relationship instance  $g_i$  in groundtruth, we calculate its *hit score*  $s(g_i)$  as follows:

$$s(g_i) = \max_{r_k \in R_{\alpha_i}} \varphi(g_i, r_k), \quad (12)$$

where  $R_{\alpha_i}$  denotes the set of detected relationship instances, whose bounding box IoUs on subject and object to the ones of  $g_i$  both exceed the predefined threshold, with the top  $\alpha_i$  confidences;  $\alpha_i$  denotes the number of groundtruth relationship instances which have the same subject and object to  $g_i$ ;  $\varphi(\cdot)$  denotes the semantic similarity between two relationship instances in HVRD task, which is calculated as the mean value of the similarities between their subjects, predicates and objects:

$$\varphi(g, r) = \begin{cases} 0, & \varphi^S(g, r) \cdot \varphi^P(g, r) \cdot \varphi^O(g, r) = 0, \\ \frac{1}{3}(\varphi^S(g, r) + \varphi^P(g, r) + \varphi^O(g, r)), & \text{otherwise,} \end{cases}, \quad (13)$$

where  $\varphi^S(g, r)$ ,  $\varphi^P(g, r)$  and  $\varphi^O(g, r)$  denote the similarities between the subjects, predicates and objects of two relationship instances  $g$  and  $r$  in object concept hierarchy and predicate concept hierarchy, respectively. The calculation of  $\varphi^S(g, r)$ ,  $\varphi^P(g, r)$  and  $\varphi^O(g, r)$  are similar, *e.g.*, the calculation of  $\varphi^S(g, r)$  is as follows:

$$\varphi^S(g, r) = \begin{cases} \frac{d_{g^S, r^S}}{d_{g^S}}, & r^S \in T_{g^S}, \\ 0, & \text{otherwise,} \end{cases}, \quad (14)$$

where  $g^S$  and  $r^S$  denote the subjects of  $g$  and  $r$ , respectively;  $T_{g^S}$  denotes the transitive closure of  $g^S$  in object concept hierarchy;  $d_{g^S}$  and  $d_{r^S}$  denote the distances from the concept “entity” to  $g^S$  and  $r^S$  in object concept hierarchy, respectively. Note that we only use  $\varphi^P(g, r)$  as the evaluation criterion in HPD task, *i.e.*,  $\varphi(g, r) = \varphi^P(g, r)$  in Equation (13). Based on Equation (12)-(14), we



Figure 3: Qualitative results on H-VRD dataset. The green boxes show the groundtruth relationship triplets and the red boxes show the hierarchical relationship triplets predicted by the proposed method with evaluation scores.

can calculate  $H\text{-Recall}@N(k=\alpha)$  as follows:

$$H\text{-Recall}@N(k=\alpha) = \frac{1}{|G|} \sum_{s \in S_G^N} s, \quad (15)$$

where  $G$  denotes the set of relationship instances in groundtruth;  $S_G^N$  denotes the set of hit scores of all the groundtruth relationship instances in  $G$  according to the top- $N$  predictions with the highest scores, which are calculated by Equation (12).

To be compatible with the existing evaluation criterion in hierarchical object detection [9], we also use another criterion  $B\text{-Recall}@N(k=\alpha)$  in our experiments.  $B\text{-Recall}@N(k=\alpha)$  extends the definition of binary recall criterion in hierarchical object detection evaluation to HVRD evaluation, which is calculated by binarizing the hit score in Equation (12), *i.e.*, if positive,  $s(g_i)$  is set to 1, otherwise 0.

### 4.3 Datasets

Visual Relationship Detection (VRD) dataset [20] and Visual Genome (VG) dataset [15] are two widely used datasets in visual relationship detection. Specifically, VRD dataset contains 5,000 images, 100 object categories, 70 predicate categories and 37,993 relationship instances among 6,672 unique triplets. VG dataset contains 99,658 images with abundant noisy relationship annotations described with natural language. Based on VRD dataset and VG dataset, we construct two datasets for HVRD, named H-VRD and H-VG, by expanding their flat relationship category spaces to hierarchical ones, respectively. To the best of our knowledge, H-VRD and H-VG are the first HVRD datasets.

The hierarchical relationship category spaces in both H-VRD dataset and H-VG dataset consist of two parts, *i.e.*, object concept hierarchy and predicate concept hierarchy. First, we construct the object concept hierarchy by associating raw object categories to WordNet synsets. We use WordNet package of NLTK [19] for automatic association and obtain a directed concept graph initially. The directed edge indicates generalization/specialization relation between two concepts on each side of it. As for the concept nodes with multiple parents caused by polysemous phenomenon, we manually select the most appropriate ones and obtain a single-root tree structure. The most abstract concept in object concept hierarchy is “entity”. Next, we construct the predicate concept hierarchy with an intuitive strategy, similar to the construction of object concept hierarchy. The most abstract concept in the

predicate concept hierarchy is named “interaction”, which is further divided into four high level concepts, namely “action”, “spatial”, “possession” and “comparative”. All transitive verbs belong to the category of “action”. Simple spatial relations are associated with the category of “spatial” (*e.g.* “under” and “above”). “Possession” comprises near-synonymy predicates, such as “has” and “with”. The last category of “comparative” consists of comparative phrases, such as “taller than”. It is worth noting that the phrases with the combination of prepositions and verbs are also taken into consideration, *e.g.*, “stand next to” and “sit on” are associated with “next to” and “on” respectively. The intransitive verbs are left out of the aforementioned categories, as they are more likely to characterize the action states of subjects rather than the interaction with objects.

The constructed H-VRD dataset contains 5,000 images with 155 hierarchical object concepts and 80 hierarchical predicate concepts. It contains 37,993 relationship instances among 6,672 unique relationship triplets. The distance from the concept “entity” to the leaf object concepts in object concept hierarchy ranges from 3 to 10 at an average of 7.3, and the distance from the concept “interaction” to the leaf predicate concepts in predicate concept hierarchy varies between 3 and 6, at an average of 3.9. The H-VRD dataset is randomly divided into training and test sets with 4,000 and 1,000 images respectively. Considering the extremely large scale of VG dataset, we construct H-VG dataset on the pruned version provided by [40]. The constructed H-VG dataset contains 99,658 images with 301 hierarchical object concepts and 118 hierarchical predicate concepts. It contains 1,174,692 relationship instances among 19,237 unique relationship triplets. The distance from the concept “entity” to the leaf object concepts in object concept hierarchy ranges from 3 to 12, at an average of 7.7, and the distance from the concept “interaction” to the leaf predicate concepts in predicate concept hierarchy varies between 3 and 5, at an average of 4.0. We split H-VG dataset into training and test sets with 73,801 and 25,857 images respectively.

### 4.4 Component Analysis

There are three key components in the proposed method: feature representation, hierarchical relationship triplet recognition, and relationship instance sorting. We validate their influences on the performance of our method in HPD and HVRD tasks on H-VRD dataset.

**Table 1: Evaluation of our method with different components in HPD and HVRD tasks on H-VRD dataset.  $HR@N$  and  $BR@N$  are the abbreviations of  $H\text{-Recall}@N(k=\alpha)$  and  $B\text{-Recall}@N(k=\alpha)$ , respectively.**

Method	HPD				HVRD			
	HR@50	HR@100	BR@50	BR@100	HR@50	HR@100	BR@50	BR@100
PD-V	56.96	56.96	62.04	62.04	15.37	17.70	16.49	18.96
PD-L	58.93	58.93	64.75	64.75	14.56	17.75	15.59	19.02
Ours\HR	57.82	57.82	60.93	60.93	11.99	14.46	12.23	14.75
Ours-HOR	57.82	57.82	60.93	60.93	13.59	17.26	14.34	18.25
Ours-HPR	60.28	60.28	66.20	66.20	14.77	16.76	15.23	17.27
Ours-mix	<b>60.31</b>	<b>60.32</b>	<b>66.28</b>	<b>66.29</b>	15.56	18.24	16.64	19.51
Ours	60.28	60.28	66.20	66.20	<b>15.94</b>	<b>18.66</b>	<b>17.03</b>	<b>19.94</b>

**Feature representation.** The proposed method extracts two types of features for HVRD: visual feature and language feature. The former is used for both subject/object detection and predicate detection, and the latter is only used for predicate detection. As visual feature is indispensable for subject/object detection, we construct two varieties of our method by using different features for predicate detection: visual feature only (PD-V) and language feature only (PD-L).

The top two rows in Table 1 show the performance of these two baselines, and the last row in Table 1 shows the performance of our method. We can see that visual feature and language feature complement each other, and our method obtains best performance by fusing the two kinds of features. It validates the effectiveness of both the visual feature and the language feature.

**Hierarchical relationship triplet recognition.** Hierarchical relationship triplet recognition is the core module of our method, which consists of two components: hierarchical object recognition (HOR) and hierarchical predicate recognition (HPR). The former recognizes the subject and the object in a relationship triplet with the assistance of object concept hierarchy, and the latter recognizes the predicate in a relationship triplet with the assistance of predicate concept hierarchy. To validate their effectiveness, we generate three baselines: using neither of HOR nor HPR (Ours\HR), HOR only (Ours-HOR), and HPR only (Ours-HPR).

The third to fifth rows in Table 1 show the performance of these three baselines. There is little room for doubt that both HOR and HPR can provide significant improvement in HVRD task by comparing the performance of Ours-HOR and Ours-HPR to that of Ours\HR, and our method obtains the best performance by using HOR and HPR together. However, it should be noted that there is no improvement by using HOR only (compare Ours\HR vs. Ours-HOR) or together with HPR (compare Ours-HPR vs. Ours) in HPD task, because the categories of subjects and objects are given in HPD task. It shows that the two components in hierarchical relationship triplet recognition are both effective in HVRD task.

**Relationship instance sorting.** Our method uses an ordered sorting strategy to enhance the diversity of detected relationship instances. To validate its effectiveness, we generate a baseline by mixing all the detected relationship instances in sorting (Ours-mix).

The sixth row in Table 1 shows the performance of Ours-mix, which apparently degrades on HVRD and slightly increases on HPD as compared to that of Ours. Both sorting strategies can improve the performance of our model. The results show that mixed sorting strategy only takes effect when handling small number

of predictions, however impractical in application scenario. The proposed ordered sorting strategy shows enhanced robustness.

#### 4.5 Comparison with State-of-the-Arts

We compare the performance of the proposed method with four state-of-the-art methods, namely Lu’s [20], VTS [40], DR-net [5] and DSR [17], in HPD and HVRD tasks on both H-VRD dataset and H-VG dataset. It should be noted that we only compare with VTS and DSR on H-VG dataset because they can be evaluated on the cleaned version of VG provided by [40] with relatively less adaptation. Since all these methods aimed at the conventional VRD tasks, we adapt them to satisfy the requirements of HVRD tasks so that the compared methods can also generate hierarchical predicates and hierarchical visual relationship instances with the assistance of object concept hierarchy and predicate concept hierarchy. Besides, to eliminate the influence of object proposal, which is still an open problem, we provide the same object proposals for all methods in our experiments to make pair comparison.

Figure 3 shows some qualitative results of the proposed method on H-VRD dataset. Table 2 and Table 3 show the comparison results on H-VRD dataset and H-VG dataset, respectively. From the comparison results, we conclude that:

(1) Our method is superior to the state-of-the-art baselines on all the evaluation criteria in both HPD and HVRD tasks. For instance, in HPD task, our method improves the performance by 6.05% and 11.41% on HR@100 and BR@100 respectively as compared to the top-performing baseline (DSR) on H-VRD dataset; and in HVRD task, our method improves the performance by 1.76% and 2.44% on HR@100 and BR@100 respectively as compared to the top-performing baseline (DR-net) on H-VRD dataset.

(2) All the B-Recall values are larger than the corresponding H-Recall values by reasonable margins according to the evaluation results of our method. For instance, Table 2 shows that B-Recall@100 is larger than H-Recall@100 by 5.92% and 1.28% in HPD task and HVRD task on H-VRD dataset respectively. It shows that our method explores and utilizes both object and predicate concept hierarchies effectively. Since the correct generalized relationship triplets are treated equally to the specific ones in B-Recall criteria, on the one hand, higher B-Recall criteria means that our method can predict correct generalized predicates and relationships. On the other hand, the margins between the H-Recall values and the B-Recall values are reasonable means our method achieves higher correctness while the semantic specificity of the predictions are guaranteed. An intuitive trick, in which only the most generalized

**Table 2: Evaluation of different methods in HPD and HVRD tasks on H-VRD dataset.**

Method	HPD				HVRD			
	HR@50	HR@100	BR@50	BR@100	HR@50	HR@100	BR@50	BR@100
Lu’s	50.32	50.32	50.75	50.75	13.81	14.92	13.84	15.26
VTS	50.08	50.08	50.59	50.59	11.84	13.95	12.04	15.15
DR-net	53.62	53.62	54.02	54.02	14.80	16.90	14.84	17.50
DSR	54.19	54.23	54.71	54.79	14.64	16.82	14.68	17.46
Ours	<b>60.28</b>	<b>60.28</b>	<b>66.20</b>	<b>66.20</b>	<b>15.94</b>	<b>18.66</b>	<b>17.03</b>	<b>19.94</b>

**Table 3: Evaluation of different methods in HPD and HVRD tasks on H-VG dataset.**

Method	HPD				HVRD			
	HR@50	HR@100	BR@50	BR@100	HR@50	HR@100	BR@50	BR@100
VTS	64.44	64.66	65.24	65.47	6.19	8.17	6.21	8.63
DSR	64.27	68.56	65.12	69.47	0.31	0.57	0.32	0.57
Ours	<b>73.89</b>	<b>73.99</b>	<b>76.11</b>	<b>76.25</b>	<b>9.40</b>	<b>11.29</b>	<b>9.77</b>	<b>11.74</b>

**Table 4: Evaluation of different methods in PD and VRD tasks.  $R@N$  is the abbreviation of  $Recall@N(k=\alpha)$ .**

Method	PD				VRD			
	VRD dataset		VG dataset		VRD dataset		VG dataset	
	R@50	R@100	R@50	R@100	R@50	R@100	R@50	R@100
VTS	48.30	48.30	61.26	61.47	10.64	11.29	5.98	7.32
DSR	51.52	51.54	60.99	65.08	<b>14.10</b>	<b>15.55</b>	0.30	0.52
Ours\HR	<b>53.75</b>	<b>53.75</b>	<b>71.01</b>	<b>71.13</b>	13.29	14.69	<b>8.76</b>	<b>10.22</b>
Ours	45.78	45.78	64.00	64.16	3.65	4.38	3.44	4.30

relationship triplet, *i.e.*  $\langle \text{entity}, \text{interact}, \text{entity} \rangle$ , is predicted, can obtain very high B-Recall values. However, the corresponding H-Recall values will drop dramatically because of the extremely low specificity of the triplet predictions.

#### 4.6 Discussion

An interesting question about our method is whether it is effective in conventional predicate detection (PD) and VRD tasks, though it is proposed for HVRD task. To validate its performance in these tasks, we remove the hierarchical relationship triplet recognition module from our method, *i.e.*, using the Ours\HR in Table 1. We also define a criterion  $Recall@N(k=\alpha)$ , which follows the definition of recall criterion in VRD evaluation by only replacing  $k=m$  with  $k=\alpha$ . It is calculated by replacing the similarity measurement in Equation (13) with the requirement of the same relationship triplets in the detected relationship instance and the groundtruth relationship instance, *i.e.*,  $\varphi(g, r)$  is equal to 1 if the relationship triplets of  $g$  and  $r$  are the same, and 0 otherwise. We compare our method with two state-of-the-art methods, VTS [40] and DSR [17], on VRD and VG datasets. Table 4 shows the comparison results. We can see that our method obtains comparable results in most cases and even outperforms the state-of-the-arts, *e.g.* R@50 and R@100 in predicate detection task on VRD and VG datasets. The results lead to the thinking that whether HVRD is more meaningful compared with VRD, or in other words, whether HVRD helps to explore more relationship instances with correct description of image content. From Table 2 to 4, we can see that all the methods have shown evident improvements in performance on the criteria of HR@N as compared with the corresponding ones of R@N. There is some

doubt that the improvement on HR@N is caused by considering the generalized relationship instances, which are ignored in the calculation of R@N criteria even these relationship instances are also detected. Such a situation is hardly possible to occur because the generalized concepts of both object and predicate are not considered in VRD, which assumes exclusive object and predicate categories. Hence, the improvements on HR@N criteria indicate the meaning of HVRD in exploring more relationship instances with correct description.

## 5 CONCLUSION

In this paper, we proposed a novel VRD task, namely HVRD, which aims to tackle the ambiguity of manual annotation by exploiting the compatibility of relationship triplets with concept hierarchies. We also generated the first datasets, H-VG and H-VRD, for HVRD evaluation. Moreover, we proposed the first HVRD method consisting of hierarchical object detection module and hierarchical predicate detection module based on order embedding. The experiment results show that our method is superior to the state-of-the-art baselines.

## 6 ACKNOWLEDGEMENTS

This work is supported by National Science Foundation of China (61202320, 61732007), Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20180307151516166), and Collaborative Innovation Center of Novel Software Technology and Industrialization.



## REFERENCES

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiaseen 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.
- [2] Yusuf Aytar, O Bilal Orhan, and Mubarak Shah. 2007. Improving semantic concept detection and retrieval using contextual estimates. In *IEEE International Conference on Multimedia and Expo*. 536–539.
- [3] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Iklizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research* 55 (2016), 409–442.
- [4] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. 2018. Context-dependent diffusion network for visual relationship detection. In *ACM International Conference on Multimedia*. 1475–1482.
- [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] Jia Deng, Alexander C Berg, Kai Li, and Li Fei-Fei. 2010. What does classifying more than 10,000 image categories tell us?. In *European Conference on Computer Vision*. 71–84.
- [7] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. 2014. Large-scale object classification using label relation graphs. In *European conference on computer vision*. 48–64.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
- [9] Jia Deng, Jonathan Krause, Alexander C Berg, and Li Fei-Fei. 2012. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3450–3457.
- [10] Jia Deng, Sanjeev Sathesh, Alexander C Berg, and Fei Li. 2011. Fast and balanced: Efficient label tree learning for large scale object recognition. In *Advances in Neural Information Processing Systems*. 567–575.
- [11] Nan Ding, Jia Deng, Kevin P Murphy, and Hartmut Neven. 2015. Probabilistic label relation graphs with ising models. In *IEEE International Conference on Computer Vision*. 1161–1169.
- [12] Ross Girshick. 2015. Fast r-cnn. In *IEEE International Conference on Computer Vision*. 1440–1448.
- [13] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*. 241–257.
- [14] 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.
- [15] 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 (2017), 32–73.
- [16] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *IEEE International Conference on Computer Vision*. 1261–1270.
- [17] Kongming Liang, Yuhong Guo, Hong Chang, and Xilin Chen. 2018. Visual relationship detection with deep structural ranking. In *AAAI Conference on Artificial Intelligence*. 7098–7105.
- [18] Xiaodan Liang, Lisa Lee, and Eric P Xing. 2017. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 848–857.
- [19] Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *arXiv preprint cs/0205028* (2002).
- [20] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *European Conference on Computer Vision*. 852–869.
- [21] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).
- [22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.
- [23] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38 (1995), 39–41.
- [24] Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. From large scale image categorization to entry-level categories. In *IEEE International Conference on Computer Vision*. 2768–2775.
- [25] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*. 6504–6512.
- [26] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7263–7271.
- [27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. 91–99.
- [28] Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*. 1745–1752.
- [29] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. 2017. Video visual relation detection. In *ACM International Conference on Multimedia*. 1300–1308.
- [30] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [31] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361* (2015).
- [32] Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *IEEE International Conference on Computer Vision*. 3551–3558.
- [33] 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.
- [34] Jianwei Yang, Jiaseen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *European Conference on Computer Vision*. 670–685.
- [35] 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.
- [36] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *European Conference on Computer Vision*. 684–699.
- [37] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. 2018. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *European Conference on Computer Vision*. 322–338.
- [38] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. 2017. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision*. 1974–1982.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] Ji Zhang, Yannis Kalantidis, Marcus Rohrbach, Manohar Paluri, Ahmed Elgammal, and Mohamed Elhoseiny. 2018. Large-scale visual relationship understanding. *arXiv preprint arXiv:1804.10660* (2018).
- [43] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. 2017. Open vocabulary scene parsing. In *IEEE International Conference on Computer Vision*. 2002–2010.
- [44] Hao Zhou, Chuanping Hu, Chongyang Zhang, and Shengyang Shen. 2019. Visual Relationship Recognition via Language and Position Guided Attention. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2097–2101.
- [45] Yaohui Zhu and Shuqiang Jiang. 2018. Deep structured learning for visual relationship detection. In *AAAI Conference on Artificial Intelligence*. 7623–7630.
- [46] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. 2017. Towards context-aware interaction recognition for visual relationship detection. In *IEEE International Conference on Computer Vision*. 589–598.