PPR-FCN: Weakly Supervised Visual Relation Detection via Parallel Pairwise R-FCN

Hanwang Zhang[†], Zawlin Kyaw[‡], Jinyang Yu[†], Shih-Fu Chang[†]

[†]Columbia University, [‡]National University of Singapore

{hanwangzhang, kzl.zawlin, yjy941124}@gmail.com; shih.fu.chang@columbia.edu

Abstract

We aim to tackle a novel vision task called Weakly Supervised Visual Relation Detection (WSVRD) to detect "subject-predicate-object" relations in an image with object relation groundtruths available only at the image level. This is motivated by the fact that it is extremely expensive to label the combinatorial relations between objects at the instance level. Compared to the extensively studied problem, Weakly Supervised Object Detection (WSOD), WSVRD is more challenging as it needs to examine a large set of regions pairs, which is computationally prohibitive and more likely stuck in a local optimal solution such as those involving wrong spatial context. To this end, we present a Parallel, Pairwise Region-based, Fully Convolutional Network (PPR-FCN) for WSVRD. It uses a parallel FCN architecture that simultaneously performs pair selection and classification of single regions and region pairs for object and relation detection, while sharing almost all computation shared over the entire image. In particular, we propose a novel position-role-sensitive score map with pairwise RoI pooling to efficiently capture the crucial context associated with a pair of objects. We demonstrate the superiority of PPR-FCN over all baselines in solving the WSVRD challenge by using results of extensive experiments over two visual relation benchmarks.

1. Introduction

Visual relation detection (VRD) aims to detect objects and predict their relationships in an image, especially `subject-predicate-object` triplets like `person-hold-ball` (verb), `dog-on-sofa` (spatial), `car-with-wheel` (preposition), and `person1-taller-person2` (comparative) [20]. As an intermediate task between low-level object detection [23] and high-level natural language modeling [42], VRD has received increasing attention recently, in areas of new benchmarks [29, 20], algorithms [24, 48, 7, 49], and visual reasoning [17, 15, 44]. VRD is ex-
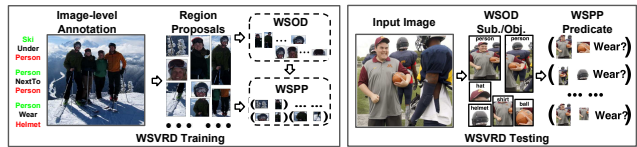

Figure 1. The illustration of WSVRD training and testing stages. It includes weakly supervised object detection (WSOD) and weakly supervised predicate prediction (WSPP). Note that the key difference from WSOD is that WSVRD requires pairwise region modeling for many weakly labeled region pairs in WSPP.

pected to become an important building block for the connection between vision and language.

Like any other visual detection task, VRD is also data-hungry. However, labeling high-quality relation triplets is much more expensive than objects as it requires the tedious inspection of a combinatorial number of object interactions. On the other hand, collecting image-level relation annotation is relatively easier. For example, there are abundant image-caption data [20, 26] and Web image-text pairs [40], where *image-level* relation descriptions can be automatically extracted from the text using state-of-the-art text parsers [37, 1]. Therefore, to make VRD of practical use at a large scale, it is necessary to study the novel and challenging task: *weakly supervised visual relation detection* (WSVRD), with triplet annotation available only at the *image level*.

Figure 1 shows the WSVRD problem studied in this paper. As there are no instance-level object annotations (*e.g.*, bounding boxes), we first exploit region proposal generators [50, 41] for a set of candidate proposals (or RoIs) and then predict their object classes. This step is also known as *Weakly Supervised Object Detection* (WSOD) [3, 6]. Then, as the image-level relation does not specify which pairs of objects are related, it exhaustively enumerates every RoI pairs as candidate `subject-object` pairs for `predicate` prediction (*e.g.*, relationships), which results in that WSVRD is more challenging than WSOD. More specifically, first, as the spatial context annotation of pairwise regions is missing, we should carefully model the spa-

tial constraints in WSVRD otherwise the relationships will be easily confused by incorrect `subject-object` configurations, *e.g.*, one can detect `hat-on-person` correctly but the `hat` is on someone else; second, for N regions, WSVRD has to scan through $\mathcal{O}(N^2)$ region pairs, thus, the weakly supervised learning based on alternating between instance selection and classification (*i.e.*, predicate prediction) in WSVRD is more easily trapped in bad local optimal solution than that in WSOD [21]; third, the $\mathcal{O}(N^2)$ computational cost in WSVRD would become prohibitively expensive if per-RoI fully-connected subnetwork is still adopted [48, 49], since WSVRD usually uses many more object regions (*e.g.*, >100 regions and >10,000 pairs) than supervised VRD (*e.g.*, <20 regions and <400 pairs) to ensure high recall of object instances.

We present a Parallel, Pairwise Region-based, end-to-end Fully Convolutional Network (PPR-FCN) to tackle the above challenges in WSVRD. The architecture is illustrated in Figure 2 and detailed in Section 3. It consists of a WSOD module for weakly supervised object detection and a Weakly Supervised Predicate Prediction (WSPP) module for weakly supervised region pair modeling. PPR-FCN is a two-branch parallel network, inspired by the recent success of using parallel networks to avoid bad local optima in WSOD [3]. We use FCN [23] as our backbone network to exploit its advantages in sharing computation over the entire image, making efficient pairwise score estimation possible [23, 28]. The WSPP module has two novel designs:

1. Position-Sequence-Sensitive Score Map. Inspired by the position-sensitive score map in R-FCN [23], we develop a set of *position-role-sensitive* conv-filters to generate a score map, where every spatial pixel encodes the object class-agnostic spatial context (*e.g.*, `subject` is above `object` for predicate `sit on`) and roles (*e.g.*, the first part of the pixel channels is `subject` and the rest is `object`).

2. Pairwise RoI Pooling. To shepherd the training of the position-role-sensitive conv-filters, we append a *pairwise RoI pooling* layer on top of the score map for fast score estimation. Our pooling design preserves the spatial context and subject/object roles for relations.

To the best of our knowledge, PPR-FCN is the first detection network for the WSVRD task. We believe that PPR-FCN will serve as a critical foundation in this novel and challenging vision task.

2. Related Work

Fully Convolutional Networks. A recent trend in deep networks is to use convolutions instead of fully-connected (fc) layers such as ResNets [13] and GoogLeNet [39]. Different from fc layers where the input and output are fixed size, FCN can output dense predictions from arbitrary-sized inputs. Therefore, FCN is widely used in segmenta-

tion [28, 27], image restoration [9], and dense object detection windows [34]. In particular, our PPR-FCN is inspired by another benefit of FCN utilized in R-FCN [23]: per-RoI computation can be shared by convolutions. This is appealing because the expensive computation of pairwise RoIs is replaced by almost cost-free pooling.

Weakly Supervised Object Detection. As there are no instance-level bounding boxes for training, the key challenge of WSOD is to localize and classify candidate RoIs simultaneously [6, 43, 38, 16]. The parallel architecture in PPR-FCN is inspired by the two-branch network of Bilen and Vedaldi [3], where the final detection score is a product of the scores from the parallel localization and classification branches. Similar structures can be also found in *et al.* [18, 35]. Such parallel design is different from MIL [30] in a fundamental way as regions are selected by a localization branch, which is independent of the classification branch. In this manner, it helps to avoid one of the pitfalls of MIL, namely the tendency of the method to get stuck in local optima.

Visual Relation Detection Modeling the interactions between objects such as verbs [11, 4], actions [12, 33, 45], and visual phrases [46, 2, 36, 8] are not new in literature. However, we are particularly interested in the VRD that simultaneously detects generic `subject-predicate-object` triplets in an image, which is an active research topic [29, 24, 48, 49, 7, 25] and serves as a building block for connecting vision and language [20, 17, 15, 44]. But, a key limitation is that it is very expensive to label relation triplets as the complexity is combinatorial. Perhaps the most related work to ours is done by Prest *et al.* [32] on weakly-supervised learning human and object interactions. However, their spatial configurations and definitions of relations are limited to one person and one object while our relations include generic objects and diverse predicates. There are recent works on referring expression groundings, *e.g.*, localizing an object by its relationship to another object [14, 47, 31]. However, they require stronger supervision, *i.e.*, at least one of the objects is labeled with bounding box. We also notice that we are not the only work towards the efficiency of VRD. Li *et al.* [24] and Zhang *et al.* [49] proposed to use groundtruth pairwise bounding boxes to learn triplet proposals to reduce the number of region pairs; however, these methods are fully supervised.

3. PPR-FCN

As illustrated in Figure 2, PPR-FCN consists of two modules: 1) WSOD module for object detection and 2) WSPP module for predicate prediction. At test time, PPR-FCN first detects a set of objects and then predicts the predicate for every object pairs. In this section, we will detail each module.

Figure 2. The overview of the proposed PPR-FCN architecture for WSVRD. It has two modules: WSOD for object detection (Section 3.1) and WSPP for predicate prediction (Section 3.2), each module is composed by a pair selection branch and a classification branch.

3.1. WSOD Module

The goal of WSOD module is to predict the object class score $S_c(P)$ for a RoI P of any class $c \in \{0, 1, ..., C\}$. Then, NMS is performed to generate the final detection result: subject and object RoIs (*i.e.*, bounding boxes), and their classes. It is worth noting that any state-of-the-art WSOD method can be used as the WSOD module in PPR-FCN. In this paper, we adopt a parallel design similar in WSDDN [3] as it achieves the state-of-the-art results on benchmarks (cf. Section 4.3) and it is easy to replace the backbone network with R-FCN [23], which is also compatible with the subsequent WSPP module. During training and test, we first use EdgeBox [50] to generate N object RoIs, where N is initially 4,000 and then reduced to 1,000 by NMS with IoU>0.4 and discard those with objectness score<0.2; the objectness score for a region is the sum over all class scores from a 3-epoch pre-trained WSOD with the initial 4,000 RoIs. Then, given the 1,000 RoIs, for each class, we perform NMS with IoU>0.4 and score threshold>0.7 to select 15~30 RoIs, resulting ~100 detected objects, where this number is significantly larger than that in supervised detection (*e.g.*, ~20), since we need to ensure enough recall for true objects. Please see Section 3.3 for the training loss of this module.

3.2. WSPP Module

WSPP module predicts the predicate score $S_r(P_i, P_j)$ of any predicate $r \in \{1, ..., R\}$ for two RoIs detected by the previous WSOD module. As shown in Figure 2, it is a two-branch network with independent parameters for pair selection (*i.e.*, which pair of regions are related) and classification. In particular, the input feature map for WSPP is the same as WSOD, which is the base CNN feature map followed by a trainable conv-layer as in R-FCN [23]. The predicate score, *i.e.*, the likelihood of subject-object pair being associated with predicate r, is defined as:

$$S_r(P_i, P_j) = S_r^{sel}(P_i, P_j) \cdot S_r^{cls}(P_i, P_j), \quad (1)$$

where we split the challenging estimation of the predicate score using only image-level annotation into two simpler

problems: one is responsible for pair selection and the other is for predicate classification. In particular, S_r^{sel} (or S_r^{cls}) is the predicate score from the selection (or classification) branch. S_r^{sel} is softmax normalized over all possible region pairs with respect to a predicate class, *i.e.*, $S_r^{sel}(P_i, P_j) \leftarrow \text{softmax}_{i,j} S_r^{sel}(P_i, P_j)$; while S_r^{cls} is softmax normalized over possible predicate classes for a region pair, *i.e.*, $S_r^{cls}(P_i, P_j) \leftarrow \text{softmax}_r S_r^{cls}(P_i, P_j)$. Note that such normalizations assign different objectives to two branches and hence they are unlikely to learn redundant models [3]. Essentially, the normalized selection score can be considered as a soft-attention mechanism used in weakly supervised vision tasks [35, 5] to determine the likely RoIs. Next, we will introduce how to calculate the scores before normalization. Without loss of generality, we use S_r^{cls} as the example and discard the superscript.

3.2.1 Position-Sequence-Sensitive Score Map

First, predicate score should be position-sensitive as the spatial context of two objects is informative for the relationship. Second, as the predicate score is usually dependent on the role-sequence of two RoIs, the score should be also role-sensitive to ensure asymmetric scores of $S_r(P_i, P_j)$ and $S_r(P_j, P_i)$. For example, for ride score, person-ride-bike is more likely than bike-ride-person; person-on-bike is different from bike-on-person, as the former usually indicates "person riding a bike" while the latter suggests "person carrying a bike". Inspired by the *position-sensitive* score map design in R-FCN [23], we propose to use two sets of trainable size 1×1 and stride 1 conv-filters to generate $2 \cdot k^2 R$-channel *position-role-sensitive* score maps from the input feature map. As illustrated in Figure 3, the first $k^2 R$-channel score map encodes R predicate scores at k^2 spatial positions for subject and the second $k^2 R$-channel map encodes scores for object. By using these filters, the computation of predciate prediction is amortized over the entire image. Note that the score maps are class-agnostic, *i.e.*, they are only aware of whether a spatial location is subject or object but not aware of whether

it is "dog" or "car". This is scalable to relation detection for many classes and predicates as the complexity is only $\mathcal{O}(C + R)$ but not $\mathcal{O}(C^2 R)$.

3.2.2 Pairwise RoI Pooling

To sheperd the training of the above position-role-sensitive filters, we design a *pairwise RoI pooling* strategy to obtain the predicate score $S_r(P_i, P_j)$ for a RoI pair. It includes three pooling steps: 1) subject pooling, 2) object pooling, and 3) joint pooling. Thus, the final $S_r(P_i, P_j)$ is the sum of these steps:

$$S_r(P_i, P_j) = S_r^{sub}(P_i) + S_r^{obj}(P_j) + S_r^{joint}(P_i, P_j). \quad (2)$$

Next, we will detail the three pooling steps as illustrated in Figure 3.

Subject/Object Pooling. This pooling aims to score whether an RoI is `subject` or `object` in a relation. Without loss of generality, we use subject pooling as the walk-through example. We first divide the RoI P into $k \times k$ spatial grids. Suppose $(x, y) \in g(i, j)$ is the set of pixels within the grid $g(i, j) \in P$, where $1 \leq i, j \leq k$, and $X_{x,y,g(i,j),r}$ is the score of the r-th predicate at the position (x, y) inside grid $g(i, j)$ in the subject score map X, then the subject pooling for P is defined as:

$$S_r^{sub}(P) = \operatorname*{vote}_{g(i,j) \in P} \left(\operatorname*{pool}_{(x,y) \in g(i,j)} \left(X_{x,y,g(i,j),c} \right) \right), \quad (3)$$

where $k = 3$, pool$(\cdot)$ is mean pooling, and vote$(\cdot)$ is average voting (*e.g.*, average pooling for the scores of the grids). $S_r^{sub}(P)$ is position-sensitive because Eq. (3) aggregates responses for a spatial grid of RoI subject to the corresponding one from the k^2 maps (*e.g.*, in Figure 3 left, the dark red value pooled from the top-left grid of the RoI) and then votes for all the spatial grids. Therefore, the training will shepherd the $k^2 R$ subject filters to capture subject position in an image.

Joint Pooling. The above subject/object pooling does not capture the relative spatial context of a predicate. Therefore, we use joint pooling to capture how two RoIs interacts with respect to a predicate. As shown in Figure 3 right, different from the single RoI pooling where the k^2 spatial grids are over the entire RoI, the pairwise RoI pooling is based on the grids over the joint region and the pooling result for the `subject` P_i or `object` P_j is from the intersected grids between P_i (or P_j) and $P_i \cup P_j$, where the latter joint RoI is divided into $k \times k$ spatial grids. Denote $(x, y) \in g(i', j')$ as the pixel coordinates within the grid $g(i', j') \in P_i \cup P_j$, where $1 \leq i', j' \leq k$, and $X_{x,y,g(i',j'),r}^s$ (or $X_{x,y,g(i',j'),r}^o$) is the score of the r-th predicate at the position (x, y) within $g(i', j')$ from the
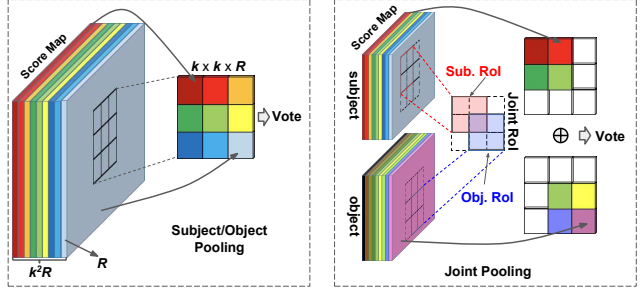
Figure 3. Illustrations of pairwise RoI pooling with $k^2 (k = 3)$ spatial grids and R predicates. **Left**: subject/object pooling for a single RoI. **Right**: joint pooling. For each score map, we use k^2 colors to represent different position channels. Each color channel has R predicate channels. For the joint pooling, uncolored pooling results indicate zero and the back-propagation is disabled through these grids. Note that the score maps in subject/object pooling and joint pooling are different, *i.e.*, there are $2 \cdot 2 \cdot k^2 R$ conv-filters.

`subject` (or `object`) score map. Therefore, the joint RoI pooling is defined as:

$$\begin{aligned} S_r^{joint}(P_i, P_j) = &\operatorname*{vote}_{g(i',j') \in P_i \cup P_j} \left(\operatorname*{pool}_{(x,y) \in g(i',j') \cap P_i} \left(X_{x,y,g(i',j'),r}^s \right) \right. \\ &+ \left. \operatorname*{pool}_{(x,y) \in g(i',j') \cap P_j} \left(X_{x,y,g(i',j'),r}^o \right) \right), \end{aligned}$$
$$(4)$$

where $g(i', j') \cap P_i$ denotes the intersected pixels between $g(i', j')$ and P_i; in particular, if $g(i', j') \cap P_i = \phi$, pool$(\cdot)$ is zero and the gradient is not back-propagated. We set $k = 3$, pool$(\cdot)$ to average pooling, and vote$(\cdot)$ to average voting. For example, for relation `person-ride-bike`, the pooling result of `person` RoI is usually zero at the lower grids of the joint RoIs while that of `bike` RoI is usually zero at the upper grids.

3.3. Loss Functions

We follow the conventional image-centric training strategy [34], *i.e.*, a training mini-batch arises from the set of region proposals in a single image. We resized images to the longer side of 720 pixels. Multiple scales at $\{0.6, 0.8, 1.0, 1.2, 1.4\}$ and random horizontal flips are applied to the images during training.

WSOD Loss. Suppose $\mathcal{C}$ is the set of image-level object class groundtruth, $S_c = \sum_i S_c(P_i)$ is the image-level class score[2], the loss is defined as:

$$\mathcal{L}_{img}^{obj} = -\sum_{c=1}^{C} \left(\mathbb{1}_{[c \in \mathcal{C}]} \log S_c + \mathbb{1}_{[c \notin \mathcal{C}]} \log(1 - S_c) \right), \quad (5)$$

where $\mathbb{1}_{[x]}$ is 1 if x is true and 0 otherwise. However, the above image-level loss does not guarantee the spatial

---

[2] Note that S_c is also a sum of element-wise product of softmax normalized scores, *i.e.*, $S_c(P_i) = S_c^{loc}(P_i) \cdot S_c^{cls}(P_i)$ and thus it is < 1.

smoothness of detection scores. Inspired by the positive and negative bounding box sampling in supervised object detection [10], we regularize the smoothness as: 1) for each foreground class $c \in \{1, ..., C\}$, the top high-scored regions (*e.g.*, top 5) and their neighborhood with IoU ≥ 0.5 should both have high scores; we consider them as the pseudo positive regions; and 2) the neighborhood of the pseudo positive regions with $0.1 \leq$ IoU ≤ 0.5 should be pseudo background regions ($c = 0$). In this way, our spatial smoothness regularization loss is:

$$\mathcal{L}_{reg}^{obj} = -\sum_{c \in \mathcal{C}} \sum_{i \in \mathcal{C}_c} \log S_c(P_i) - \sum_{i \in \mathcal{B}} \log S_0(P_i), \quad (6)$$

where $\mathcal{C}_c$ is the set of pseudo positive regions for class $c \neq 0$, and $\mathcal{B}$ is the set of pseudo background regions. We follow a similar sampling strategy as in [34]: 256 regions are sampled, where at least 75% are the pseudo background regions.

WSPP Loss. Suppose $\mathcal{R}$ is the set of the image-level relation groundtruth triplets, specifically, $(s, r, o) \in \mathcal{R}$, where $s, o \in \{1, ..., C\}$ are the labels of the `subject` and `object`. Suppose $\mathcal{C}_s$ and $\mathcal{C}_o$ are the region sets of `subject` s and `object` o, respectively. Denote $S_r = \sum_{i \in \mathcal{C}_s, j \in \mathcal{C}_o} S_r(P_i, P_j)$ as the image-level predicate score, the image-level predicate prediction loss is defined as:

$$\mathcal{L}_{img}^{pred} = -\sum_{r=1}^{R} \left(\mathbb{1}_{[(s,r,o) \in \mathcal{R}]} \log S_r + \mathbb{1}_{[(s,r,o) \notin \mathcal{R}]} \log(1 - S_r) \right). \quad (7)$$

Overall Loss. The overall loss of PPR-FCN is a multi-task loss that consists of the above WSOD and WSVRD losses:

$$\mathcal{L}_{PPR-FCN} = \mathcal{L}_{img}^{obj} + \mathcal{L}_{img}^{pred} + \alpha \mathcal{L}_{reg}^{obj}, \quad (8)$$

where α is empirically set to 0.2. We train the PPR-FCN model by SGD with momentum [19].

4. Experiments

4.1. Datasets

We used two recently released datasets with a wide range of relation annotation. Every image from the above two datasets is annotated with a set of `subject-predicate-object` triplets, where every instance pair of `subject` and `object` is labeled with bounding boxes. At training time, we discarded the object bounding boxes to conform with the weakly supervised setting.

VRD: the Visual Relationships Dataset collected by Lu *et al.* [29]. It contains 5,000 images with 100 object classes and 70 predicates, resulting in 37,993 relation annotations with 6,672 unique relations and 24.25 predicates per object class. We followed the official 4,000/1,000 train/test split.

VG: the latest Visual Genome Version 1.2 relation dataset constructed by Krishna *et al.* [20]. VG is annotated by crowd workers and thus the relations labeling are noisy, *e.g.*, free-language and typos. Therefore, we used the pruned version provided by Zhang *et al.* [48]. As a result, VG contains 99,658 images with 200 object categories and 100 predicates, 1,174,692 relation annotations with 19,237 unique relations and 57 predicates per object category. We followed the same 73,801/25,857 train/test split.

4.2. Evaluation Protocols and Metrics

Since the proposed PPR-FCN has two modules: WSOD and WSPP, we first evaluated them separately and then overall. Thus, we have the following protocols and metrics that are used in evaluating one object detection task [3, 6, 18] and three relation-related tasks [29, 48]:

1) Object Detection. We used the WSOD module trained with image-level object annotations to detect objects in VRD and VG. We followed the Pascal VOC conventions that a correct detection is at least 0.5 IoU with the groundtruth.

2) Predicate Prediction. Given the groundtruth objects with bounding boxes, we predict the predicate class between every pair of regions. This protocol allows us to study how well the proposed position-role-sensitive score map and pairwise RoI pooling perform without the limitations of object detection.

3) Phrase Detection. We predict a relation triplet with a bounding box that contains both `subject` and `object`. The prediction is correct if the predicted triplet is correct and the predicted bounding box is overlapped with the groundtruth by IoU>0.5.

4) Relation Detection. We predict a relation triplet with the `subject` and `object` bounding boxes. The prediction is correct if the predicted triplet is correct and both of the predicted `subject` and `object` bounding boxes are overlapped with the groundtruth by IoU>0.5.

Note that both the objects and relations in VRD and VG are not completely annotated. Therefore, the popular Average Precision is not a proper metric as the incomplete annotation will penalize the detection if we do not have that particular groundtruth[3]. To this end, following [29, 48], we used Recall@50 (**R@50**) and Recall@100 (**R@100**) for evaluation. R@K computes the fraction of times a groundtruth is in the top K confident predictions in an image.

4.3. Evaluations of Object Detection

Comparing Methods. We compared the proposed WSOD module named **WSOD** with three state-of-the-

---

[3]For example, even though R-FCN is arguably better than than Faster R-CNN, R-FCN only achieves 6.47% mAP while Faster-RCNN achieves 13.32% mAP on VRD.

Table 1. Weakly supervised object detection performances (R@K%) of various methods on VRD and VG. The last row is supervised object detection performances by R-FCN.

Dataset	VRD		VG	
Metric	R@50	R@100	R@50	R@100
WSDDN [3]	15.08	16.37	6.22	6.89
ContextLocNet [18]	9.74	11.27	4.74	4.91
WSL [22]	13.10	13.59	5.43	6.28
WSOD	**25.34**	**26.54**	**8.85**	**9.12**
R-FCN [23]	57.64	58.31	18.37	18.90

Table 2. Predicate prediction performances (R@K%) of various methods on VRD and VG. The last two rows are fc-based methods.

Dataset	VRD		VG	
Metric	R@50	R@100	R@50	R@100
Pos+Pairwise	24.30	24.30	43.30	43.64
Pos+JointBox	29.57	29.57	45.69	45.78
PosSeq+Pairwise	42.74	42.74	61.57	61.71
PosSeq+Pairwise+fc	**47.43**	**47.43**	**64.17**	**64.86**
VTransE [48]	44.76	44.76	62.63	62.87

art weakly supervised object detection methods: 1) **WSDDN** [3], the weakly-supervised deep detection network. It has a two-branch localization and classification structure with spatial regularization; 2) **ContextLocNet** [18], the context-aware localization network. Besides it is also a two-branch structure, the localization branch is further sub-branched to three context-aware RoI pooling and scoring subnetworks; 3) **WSL** [22], the weakly supervised object localization model with domain adaption. It is a two-stage model. First, it filters out the noisy object proposal collection to mine confident candidates as pseudo object instances. Second, it learns a standard Faster-RCNN [34] using the pseudo instances. We used their official source codes as implementations on the VRD and VG datasets in this paper. For fair comparison, we used the same ResNet-50 [13] as the base network. We also provided the fully-supervised detection model **R-FCN** [23] as the object detection upper bound.

Results. From Table 1, we can see that our WSOD is considerably better than the state-of-the-art methods. This is largely contributed by the parallel FCN architecture. It is worth noting that the quality of the top 1,000 proposal RoIs is significant to WSOD; if we directly used the original scores of EdgeBox, the performance will drop significantly by about 5 points. Note that we are still far behind the fully supervised method such as R-FCN, which shows that there is still a large space to improve WSVRD by boosting WSOD. As illustrated in Figure 5, WSOD usually detects the discriminative parts of objects, which is a common failure in state-of-the-art models. We also compared WSOD with other methods on the completely annotated Pascal VOC 2007, where we also achieved the best 39.8% mAP, surpassing WSDDN (39.3%), ContextLocNet (36.3%), and WSL (39.5%).

4.4. Evaluations of Predicate Prediction

Comparing Methods. Note that the task of predicate prediction is in the supervised setting given the groundtruth of subject and object. Thus, we removed the WSOD module and the localization branch from the WSPP module. In this experiment, our goal is to compare our proposed position-role-sensitive score map and pairwise RoI pooling, namely **PosSeq+Pairwise** with other three ablated meth-

ods: 1) **Pos-Pairwise** denoting position-sensitive score map followed by pairwise RoI pooling; 2) **Pos+JointBox** denoting position-sensitive score map followed by joint boxes RoI pooling, where the joint RoI is the tight groundtruth regions that cover both subject and object; and 3) **PosSeq+Pairwise+fc** denoting position-role-sensitive score map followed by pairwise RoI pooling, but the score is obtained by fully-connected subnetworks. Note that this fc-based method is also comparable to 4) **VtransE** [48] using the concatenated RoI features from subject and object as the input to its fc prediction network.

Results. From Table 2, we can see that our PosSeq+Pairwise outperforms the baselines with non-order score maps and non-pairwise pooling significantly. As illustrated in Figure 4, compared to the conventional position-sensitive score maps and pooling, we can observe that PosSeq+Pairwise can capture the contextual configuration better. For example, for the relation bus-on-road, the subject responses are more active at upper positions while the object response are more active at lower positions, and thus the spatial context of on is depicted by adding the pairwise pooling; however, Pos+JointBox seems agnostic to relations but more likely sensitive to objects.

It is worth noting that pooling-based methods are worse than fc-based methods such as VTransE and PosSeq+Pairwise+fc, which contains region-based fully-connected (fc) subnetworks for relation modeling. We noticed that some prepositions such as of and by, and verbs such as play and follow, contain very diverse visual cues and may not be captured by only spatial context. Therefore, fc layers followed by the concatenation of subject and object RoI features might be necessary to model such high-level semantics. Nevertheless, the fact that PosSeq+Pairwise+fc considerably outperforms VTransE demonstrates the effectiveness of exploiting the pairwise spatial context. Note that such unshared region-based subnetworks will lead to inefficient learning in WSPP as there are tens of thousands candidate RoI pairs and millions of fc parameters.

4.5. Evaluations of Phrase & Relation Detection

Comparing Methods. We evaluate the overall performance of PPR-FCN for WSVRD. We compared the following methods: 1) **GroundR** [35], a weakly supervised vi-
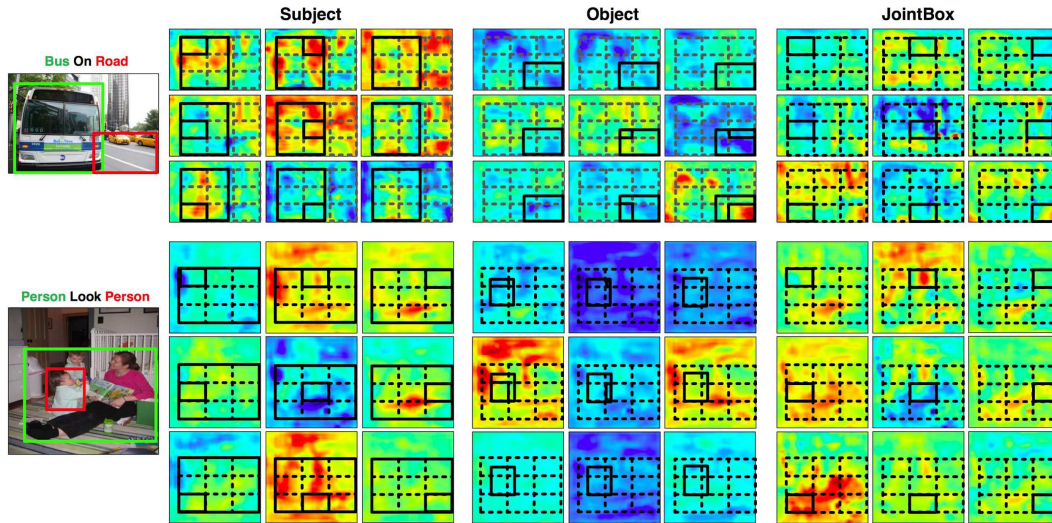
Figure 4. Two illustrative examples (without the base ResNet-50 fine-tuned) of 9 ($k = 3$) position-role-sensitive score maps trained by pairwise RoI pooling (Subject and Object) and the position-sensitive score maps trained by joint boxes pooling (JointBox). Dashed grids are the joint RoI of `subject` and `object`. In Subject and Object score maps, `subject` and `object` RoIs are in solid rectangles. The solid grids denotes pooling at corresponding positions. Note that position-role-sensitive pooling is defined as null if the RoI has no overlap with the joint RoI at a position, *e.g.*, `road` at top left. Please view in color and zoom in.
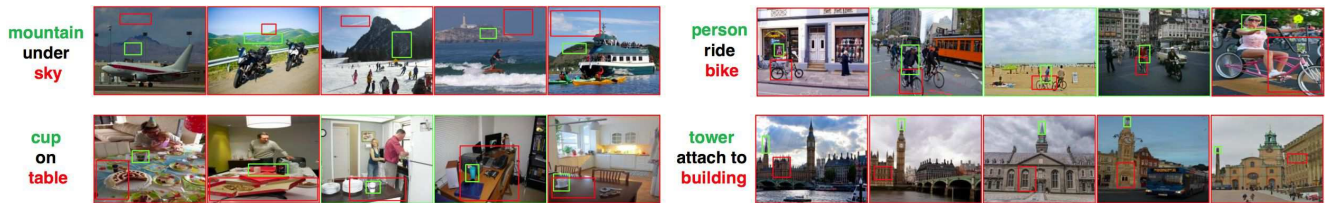
Figure 5. Illustrative top 5 relation detection results from VRD. Red and green borders denote incorrect and correct detections, respectively. Most of the failed cases are due to the wrongly detected objects.

sual phrase grounding method. We used the image-level triplets as the input short phrases for their language embedding LSTM; 2) **VisualPhrase-WSDNN**, its idea was originally proposed in [36] that considers a whole relation triplet as a class label. As it can be reduced to a weakly supervised object detection task, we used WSDDN [3] pipeline to implement VisualPhrase. Note that our parallel FCN architecture cannot be adopted in VisualPhrase since the number of relation classes is too large to construct the conv-filters. 3) **VTransE-MIL**, we followed the same pipeline of VTransE [48] but using the NoisyOR Multiple Instance Learning (MIL) [30] as the loss function for object and relation detections. 4) **PPR-FCN-single**, we only use the classification branch to implement PPR-FCN. We also compared three fully supervised models as baselines VTransE[48], Lu's-VLK [29], and Supervised-PPR-FCN, which is our proposed PPR-FCN applied in the supervised setting. Note that GroundR and VisualPhrase are based on joint boxes prediction and thus they can only perform phrase detection.

Results Table 3 and 4 reports the phrase and relation detection performances of various methods. We have the fol-

lowing observations:

1) For phrase detection, GroundR and VisualPhrase-WSDDN perform much poorly than VTransE-MIL and PPR-FCN. The reason is two-fold. First, EdgeBox is not designed to generate joint proposals of two interacted objects and thus we limited the number of proposals to 300 to handle 90,000 region pairs, where the top 300 proposals may be low recall of objects. Second, as discovered in [29, 48], once we consider the relation as a whole class label, the training samples for each relation class are very sparse, which will worsen the training of WSPP.

2) PPR-FCN outperforms VTransE-MIL in both phrase and relation detections. The reason is that VTransE does not explicitly model the spatial context in relation modeling while our PPR-FCN does. Note that this is crucial since the context can remove some incorrect `subject-object` configurations, especially when the supervision is only at the image level. For example, Figure 6 shows that the position-role-sensitive score map and pooling design in PPR-FCN can correct misaligned `subject` and `object` when there are multiple instances.

3) Our parallel design of PPR-FCN is significantly better

Table 3. Phrase detection performances (R@K%) of various methods in weakly supervised and supervised settings (bottom three rows) on VRD and VG.

Dataset	VRD		VG	
Metric	R@50	R@100	R@50	R@100
GroundR [35]	0.15	0.18	0.33	0.81
VisualPhrase-WSDDN	0.26	0.37	0.21	0.78
VTransE-MIL	4.09	6.15	1.53	2.02
PPR-FCN-single	3.56	4.31	0.87	0.98
PPR-FCN	**6.93**	**8.22**	**2.41**	**3.23**
Lu's-VLK [29]	16.17	17.03	–	–
VTransE [48]	19.42	22.42	9.46	10.45
Supervised-PPR-FCN	**19.62**	**23.15**	**10.62**	**11.08**

than its counterpart PPR-FCN-single. This demonstrates that for weakly supervised learning with many candidate instances (*i.e.*, region pairs), the parallel design without parameter sharing can prevent from bad solutions.

4) There is a large gap between WSVRD and supervised VRD, *e.g.*, PPR-FCN can only achieve less than a half of the performance of supervised VRD such as Supervised-PPR-FCN and VTransE. We believe that the bottleneck is mainly due to the WSOD module that tends to detect small discriminative part instead of the whole object region. As shown in Figure 5, most of the failed relation detection is due to the failure of object detection. For example, for large and background-like objects such as mountain, sky and building, only small regions are detected; for tower, only the most discriminative "spire" is detected.

5) Even though the fully-connected subnetworks is very helpful in predicate prediction as we discussed in Section 4.4, Supervised-PPR-FCN can still outperform the fc-based VTransE due to the effectiveness of the pairwise RoI pooling, which can correct wrong spatial context (Figure 6) Note that since PPR-FCN is designed for WSVRD, we cannot remove bad RoI pairs using pairwise groundtruth bounding boxes, which may lead to significant improvement in supervised settings [24].

6) Thanks to the FCN architecture introduced in PPR-FCN, it can not only speed up the WSOD, but also efficiently handle tens of thousands region pairs in WSVRD. For example, as reported in Table 5, PPR-FCN is about $2\times$ faster than VTransE-MIL using per-region fc subnetworks. It is worth noting that the number of parameters of PPR-FCN is much smaller that VTransE-MIL (*e.g.*, millions of fc parameters) as we only have $\mathcal{O}(k^2(C+1+R))$ conv-filters. Our current bottleneck is mainly due to the EdgeBox [50] proposal generation time, as we strictly stick to the weak supervision setting that any module should not exploit bounding boxes. However, in practice, we can use generic class-agnostic RPN [34] to generate proposals in 100 ms/img.

5. Conclusion

We presented a parallel, pairwise region-based, fully convolutional network: PPR-FCN, for the challenging task
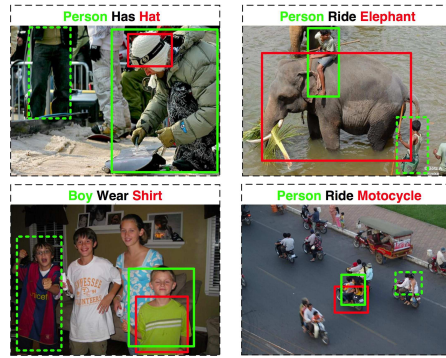

Figure 6. Qualitative examples of relation detection on VG. Compared to the results of PPR-FCN (solid green bounding boxes), VTransE-MIL (dashed green bounding boxes) is more likely to misalign subject to object if there are multiple instances of subject.

Table 4. Relation detection performances (R@K%) of various methods in weakly supervised and supervised settings (bottom three rows) on VRD and VG.

Dataset	VRD		VG	
Metric	R@50	R@100	R@50	R@100
VTransE-MIL	4.28	4.54	0.71	0.90
PPR-FCN-single	3.56	4.15	1.08	1.63
PPR-FCN	**5.68**	**6.29**	**1.52**	**1.90**
Lu's-VLK [29]	13.86	14.70	–	–
VTransE [48]	14.07	15.20	5.52	6.04
Supervised-PPR-FCN	**14.41**	**15.72**	**6.02**	**6.91**

Table 5. Titan X GPU test time (ms/img) of the fc subnetwork based weakly supervised method, VTransE-MIL and PPR-FCN (excluding the proposal generation time cost by EdgeBox, which is 700 ms/img). Both VTransE-MIL and PPR-FCN adopts ResNet-50 as the base CNN and 100 detected object proposals, *i.e.*, 10,000 region pairs for predicate prediction.

VTransE-MIL	PPR-FCN
270	150

of weakly supervised visual relation detection (WSVRD). PPR-FCN has two novel designs towards the optimization and computation difficulties in WSVRD: 1) PPR-FCN is a parallel FCN network for simultaneous classification and selection of objects and their pairwise relations, and 2) the position-role-sensitive conv-filters and pairwise RoI pooling that captures the spatial context of relations. Thanks to the shared computation on the entire image, PPR-FCN can be efficiently trained with a huge amount of pairwise regions. PPR-FCN provides the first baseline for the novel and challenging WSVRD task, which can foster practical visual relation detection methods for connecting computer vision and natural language. We found that the bottleneck of PPR-FCN is the WSOD performance. Therefore, future research direction may focus on jointly modeling WSOD and WSVRD by incorporating relations as the contextual regularization for objects.

References

[1] G. Angeli, M. J. Premkumar, and C. D. Manning. Leveraging linguistic structure for open domain information extraction. In *ACL*, 2015. 1

[2] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik. Learning to generalize to new compositions in image understanding. In *EMNLP*, 2016. 2

[3] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 1, 2, 3, 5, 6, 7

[4] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 2

[5] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017. 3

[6] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *TPAMI*, 2017. 1, 2, 5

[7] B. Dai, Y. Zhang, and D. Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017. 1, 2

[8] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 2

[9] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. In *ICCV*, 2013. 2

[10] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 5

[11] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008. 2

[12] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 2009. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. 2016. 2, 6

[14] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. *arXiv preprint arXiv:1611.09978*, 2016. 2

[15] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017. 1, 2

[16] Z. Jie, Y. Wei, X. Jin, J. Feng, and W. Liu. Deep self-taught learning for weakly supervised object localization. In *CVPR*, 2017. 2

[17] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 1, 2

[18] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, 2016. 2, 5, 6

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016. 1, 2, 5

[21] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010. 2

[22] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, 2016. 6

[23] Y. Li, K. He, J. Sun, et al. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 1, 2, 3, 6

[24] Y. Li, W. Ouyang, and X. Wang. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, 2017. 1, 2, 8

[25] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017. 2

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[27] S. Liu, C. Wang, R. Qian, H. Yu, R. Bao, and Y. Sun. Surveillance video parsing with single frame supervision. In *CVPR*, 2017. 2

[28] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[29] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 1, 2, 5, 7, 8

[30] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *NIPS*, 1998. 2, 7

[31] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2

[32] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *TPAMI*, 2012. 2

[33] V. Ramanathan, C. Li, J. Deng, W. Han, Z. Li, K. Gu, Y. Song, S. Bengio, C. Rossenberg, and L. Fei-Fei. Learning semantic relationships for better action retrieval in images. In *CVPR*, 2015. 2

[34] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 4, 5, 6, 8

[35] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 2, 3, 6, 8

[36] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011. 2, 7

[37] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language*, 2015. 1

[38] H. O. Song, R. B. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, T. Darrell, et al. On learning to localize objects with minimal supervision. In *ICML*, pages 1611–1619, 2014. 2

[39] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[40] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 1

[41] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 2013. 1

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1

[43] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014. 2

[44] Q. Wu, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, 2016. 1, 2

[45] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

[46] M. Yatskar, L. Zettlemoyer, and A. Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 2016. 2

[47] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *ECCV*, 2016. 2

[48] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua. Visual translation embedding network for visual relation detection. In *CVPR*, 2017. 1, 2, 5, 6, 7, 8

[49] J. Zhang, M. Elhoseiny, S. Cohen, W. Chang, and A. Elgammal. Relationship proposal networks. In *CVPR*, 2017. 1, 2

[50] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 1, 3, 8