

# Tell Me What They're Holding: Weakly-Supervised Object Detection with Transferable Knowledge from Human-Object Interaction

Daesik Kim,<sup>1,2,\*</sup> Gyujeong Lee,<sup>1,\*</sup> Jisoo Jeong,<sup>1,\*</sup> Nojun Kwak<sup>1,†</sup>

<sup>1</sup>Seoul National University, <sup>2</sup>NAVER WEBTOON Corp.  
 {daesik.kim, regulation.lee, soo3553, nojunk}@snu.ac.kr

## Abstract

In this work, we introduce a novel weakly supervised object detection (WSOD) paradigm to detect objects belonging to rare classes that have not many examples using transferable knowledge from human-object interactions (HOI). While WSOD shows lower performance than full supervision, we mainly focus on HOI as the main context which can strongly supervise complex semantics in images. Therefore, we propose a novel module called RRPN (relational region proposal network) which outputs an object-localizing attention map only with human poses and action verbs. In the source domain, we fully train an object detector and the RRPN with full supervision of HOI. With transferred knowledge about localization map from the trained RRPN, a new object detector can learn unseen objects with weak verbal supervision of HOI without bounding box annotations in the target domain. Because the RRPN is designed as an add-on type, we can apply it not only to the object detection but also to other domains such as semantic segmentation. The experimental results on HICO-DET dataset show the possibility that the proposed method can be a cheap alternative for the current supervised object detection paradigm. Moreover, qualitative results demonstrate that our model can properly localize unseen objects on HICO-DET and V-COCO datasets.

## Introduction

In a decade, object detection has become one of the most successful fields in computer vision with various applications (Ren et al. 2015; Dai et al. 2016; Redmon et al. 2016; Liu et al. 2016). Most of the successful models have emerged after the release of large scale datasets (e.g. PASCAL VOC, MS-COCO (Everingham et al. 2010; Lin et al. 2014)) with bounding box annotations. Given input images, conventional object detection models can localize boxes with the corresponding class scores. Thus, they normally require manually annotated bounding boxes containing accurate coordinate values and object labels for training.

However, annotating bounding boxes is time-consuming and labor-intensive. It can also be difficult to expand the volume of a dataset by adding more object classes or adding

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

\*All of authors equally contributed to this work.

†Corresponding author.

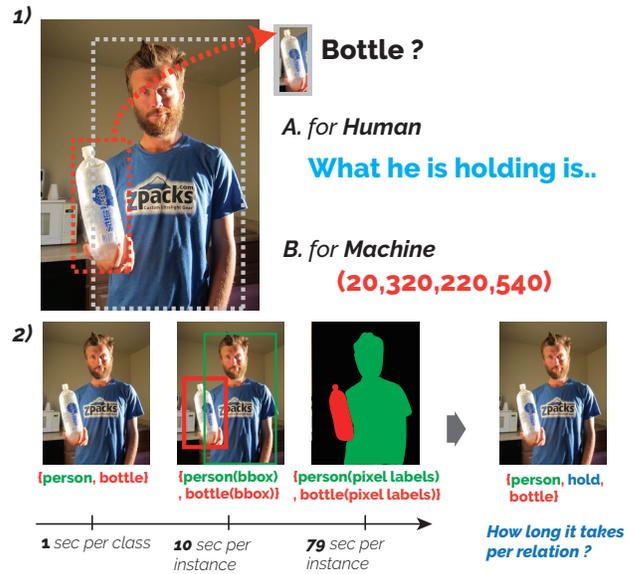


Figure 1: 1) Two different types of description of an object. A is human’s way of identifying an object while B is for machines. 2) Manually annotating time for three tasks. Bearman et al. (Bearman et al. 2016) estimated annotation times for image level, bounding box and pixel level. At rightmost, annotating time of relation sentence can be similar to that of image-level since only action verb “hold” is added.

more images. Therefore, researches to reduce those costs in various ways have drawn attentions these days.

Weakly supervised object detection (WSOD) has been proposed to tackle the aforementioned problems (Zhu et al. 2017; Shi, Caesar, and Ferrari 2017; Jie et al. 2017). It is to detect objects within images by weak supervision such as image-level labels. At the cost of lowered annotation cost, WSOD performs worse than full supervision.

To overcome a limitation of weak supervision, some approaches (Shi, Caesar, and Ferrari 2017) rely on another type of full supervision with transfer learning. Transferred knowledge from a source domain could support weak super-

vision in a target domain. However, annotating other types of labels such as segmentation mask is also expensive.

Our main intuition is that supervision of machines is totally different from that of humans. For example, Fig. 1(1) shows different ways of identifying objects between humans and machines. While we should provide accurate coordinate values of object boxes for machines, humans usually recognize new objects from contexts. Contexts also can reinforce supervision without much additional efforts.

Especially, how objects are related to human actions can be practical and advantageous since information about a human can be a proper evidence for recognizing contexts in an image. Moreover, humans can easily express contexts with sentences as shown in Fig. 1(1), so that linguistic labels can be a key to reduce annotation cost for humans as shown in Fig. 1(2). Compared to other annotating costs (Bearman et al. 2016), the cost of annotating a relation sentence such as “person, hold, bottle” can be almost similar to that of image-level annotation. Thus, we propose a novel paradigm to learn unseen objects based on human-object interaction (HOI).

Our key idea is to exploit transferable knowledge from HOI contexts annotated by language as is in (Chao et al. 2018). Specifically, we propose a novel module that predicts object locations from HOI. Since the actual coordinate values can not be specified, we use an attention map as localization results to connect it with a bounding box. Moreover, in order to train a full object detector (e.g. Faster-RCNN (Ren et al. 2015)) in an end-to-end fashion, we design a new module as an add-on type.

The objective of this paper is to make our model learn additional rare classes with weak verbal supervision annotated easily by human. During the first stage, strong supervision on non-rare classes teach our model to localize a proper location with a human pose and an action verb. In the next stage, only weak supervision with transferred knowledge keeps training an object detector for unseen rare classes.

Our main contributions can be summarized as follows:

- We define a new weakly-supervised object detection scheme which mainly relies on interactions between a human and objects without box annotations.
- We propose a novel module called RRPN (Relational Region Proposal Network) to localize boxes by using the location of a human and verb embeddings.
- The proposed RRPN is designed as a universal add-on type that can be easily adapted into existing models such as Faster-RCNN.

Our experiments validate that our model outperforms baselines on the HICO-DET (Chao et al. 2018) dataset and can effectively transfer the knowledge to other dataset such as V-COCO (Gupta and Malik 2015).

## Related works

### Weakly Supervised Object Localization and Detection

Most of the weakly supervised object localization and detection methods have been proposed based on an image-level supervision. With cheaper but weaker annotations,

studies (Bilen and Vedaldi 2016; Diba et al. 2017; Kantorov et al. 2016; Oquab et al. 2015; Tang et al. 2017; Jie et al. 2017) mainly tried to enhance performance by multiple instance learning (MIL). In MIL, a bag is defined as a collection of regions in an image. It is labeled as positive if at least one object is positive and labeled as negative if all the objects are negative.

Tang et al. 2018 have proposed proposal cluster learning algorithm to learn refined instance classifier. Yang et al. 2019 have proposed activity-driven WSOD, which also exploits action classes as contextual information to localize objects without box annotations. However, those works has generated box proposals using Selective Search (Uijlings et al. 2013), which is a rule-based algorithm. Since the box proposal method cannot be trained, there is a fundamental limitation which proper proposals hardly exist in novel data. Uijlings, Popov, and Ferrari 2018 have addressed a new WSOD framework that revisits knowledge transfer for training object detectors on target classes. Since this work has optimized a box proposal network for target classes by MIL, box generators can be insufficiently trained with rare classes due to a lack of contextual information. Our method resolves the aforementioned issues with a novel box proposal module that can transfer knowledge using an HOI dataset.

**Human-object interaction** Visual recognition of HOI is crucial for comprehending a scene in an image. Early work studied mutual context of human pose and objects (Yao and Fei-Fei 2010) and Bayesian model (Gupta and Davis 2007; Gupta, Kembhavi, and Davis 2009) with handcraft features. Recently, with success of deep learning, Chao et al. 2015 introduced a new large-scale benchmark, “Humans Interacting with Common Objects” (HICO), for HOI recognition, which was expanded for detection problems in HICO-DET (Chao et al. 2018). In order to solve HICO-DET datasets, various approaches have been proposed. In (Chao et al. 2018), combined features from human proposals and object regions were used to solve HOI detection. Gkioxari et al. 2018 proposed a human and object detector-based approach estimating a density map based on Faster-RCNN architecture. A recent approach (Qi et al. 2018) generates the HOI graph and propagates message between nodes to infer relationships in a parsing graph. In this paper, rather than directly solving HOI problems, we exploit contextual information in HICO-DET to construct a weakly-supervised object detector.

## Algorithm Overview

An overview of our algorithm is illustrated in Fig. 2. Let  $D = \{(I_i, y_i)\}_{i=1}^N$  is the data set, where  $y_i$  is the label of the image  $I_i$  and  $N$  is the number of images. The image label  $y_i$  is organized as a tuple as shown below:

$$y_i = \{(H_{verb}^j, O_{bbox}^j, O_{cls}^j)\}_{j=1}^M \quad (1)$$

where,  $O_{bbox}^j, O_{cls}^j$  are the bounding box and the class of an object in the image,  $H_{verb}^j$  is the action verb corresponding to the object, and  $M$  is the number of tuples in the image  $I_i$ . To evaluate the proposed method, we divided  $D$  into two sub-categories based on the number of objects in a class: non-rare (source classes)  $D_S$  and rare (target classes)  $D_T$ .

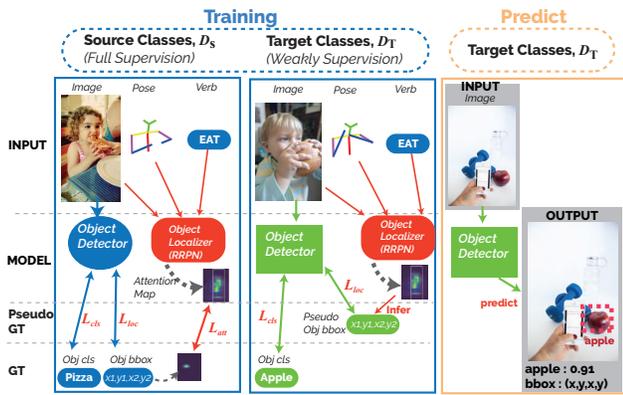


Figure 2: Overview of our algorithm. 1) During the training phase for source classes, RRPN is also trained to predict attention map from human-object interaction. 2) In the target-class training phase, an object detector is trained using the ground truth class label, and the box label provided by the trained RRPN. In other words, our problem focuses solely on solving the weakly supervised object detection problem on the target classes. 3) As a result, the trained object detector for target classes can infer box coordinates and object classes with only an image input.

Note that, there is no object class duplicates but all action verbs are overlapped between two subcategory datasets.

For  $D_S$ , we normally train the first object detector (blue circle in Fig. 2) with full supervision using  $(O_{bbox}^j, O_{cls}^j)$ . Along with training of the object detector, we also train an object localizer (red circle) called RRPN with newly defined inputs. Since the RRPN should learn how to localize an object only with the information on a human and an action verb, we use the image  $I_i$ , the verb  $H_{verb}^j$ , and the pose of the human  $H_{pose}^j$  as inputs. The  $H_{verb}^j$  simply comes from  $y_i$  but the human pose  $H_{pose}^j$  is extracted from an image  $I_i$  with an existing human pose estimation method. As a result, the RRPN predicts an attention map  $\hat{A}_i^j$  of an object location in the  $i$ -th image from human’s action and appearance. We optimize losses regarding the object class and the location using  $O_{bbox}^j$  and  $O_{cls}^j$  for the object detector, but create a Gaussian map of  $O_{bbox}^j$  and use it as a ground truth in the training of the attention map of the RRPN. In this phase, since the ground truth bounding box location is available, the RRPN can learn common knowledge between objects and human actions.

For  $D_T$ , we assume that only object class information  $O_{cls}^j$  and the action verb  $H_{verb}^j$  are available but the bounding box information is not. To fill the absence of  $O_{bbox}^j$ , we exploit learned knowledge inferred by the RRPN with the same kinds of inputs as the training phase for the source classes. Since the output of RRPN is an attention map, we extract a coordinate by thresholding it and generate pseudo bounding box  $\hat{O}_{bbox}^j$ . Then, we normally train the second object detector (green rectangle in Fig. 2) for  $D_T$ . Since we already have used all action verbs to train the RRPN in the

previous phase and transfer the same parameters in the training phase for the target classes, it can infer an object location with a human pose and an action verb. In Fig. 2, after the RRPN already learned to localize unseen object “Apple” with verb “EAT” and grabbing pose in the training phase of  $D_S$ , it can infer a proper location as a pseudo ground truth  $\hat{O}_{bbox}^j$ . In conclusion, we use weak supervision by human actions to train a full object detector.

Eventually, the trained object detector in the second phase can predict objects in  $D_T$  only with an input image  $I_i$  as shown in Fig. 2. Although we have not shown the real location of “Apple”, it is possible to predict the class score and the coordinate of an “Apple” object.

In this scheme, we can additionally train new object detector for unseen rare classes without bounding box annotations. Moreover, since we already trained the RRPN with strong supervisions, we need smaller amount of data in target classes compared to other WSOD algorithms. Our experiments validate that our target domain which contains extremely rare object classes is trained successfully by our method.

## Architecture

Fig. 3 depicts the overall architecture of the proposed algorithm. The proposed algorithm consists of two modules, including the RRPN and the object detector. More precisely, it means that RRPN can be combined with the conventional architecture such as the Faster-RCNN. RRPN is a multi-stage encoder-decoder network, which is responsible for predicting an object-location-centric attention map  $A_i^j$  from a multi-domain integrated feature map. The other module, object detector, is a conventional object detector which is trained for a given input image using the ground truth label  $O_{cls}$  and the bounding box  $O_{bbox}$ .

In order to exploit the knowledge of interaction, we train  $D_T$  after the training of  $D_S$  is done. Since, however, we do not account for the continual learning,  $D_S$  and  $D_T$  do not share parameters for the object detector. While the object detector is trained for  $D_S$  with supervision, at the same time, RRPN is also trained to learn the knowledge from interactions between a human and an object through action verbs. Then, the object detector is trained for  $D_T$  without object bounding boxes, i.e. in a weakly supervised way, using the transferred knowledge from  $D_S$ .

### Training on the Source classes $D_S$

$D_S$  are object classes on which data can be easily acquired. Training on  $D_S$  is a standard supervised object detection procedure by using ground truth class and box labels for all the objects. The main purpose of training on  $D_S$  is to predict an object location from a human-object interaction. Therefore, RRPN is also trained at the same time as the training of the object detector. The detailed training procedure for Faster-RCNN is applied in the same way as the original paper. The training procedure of RRPN is as follows.

**Relational Region Proposal Network (RRPN)** RRPN is designed to be universally applicable to various task’s models, including other object detectors, in an add-on manner,

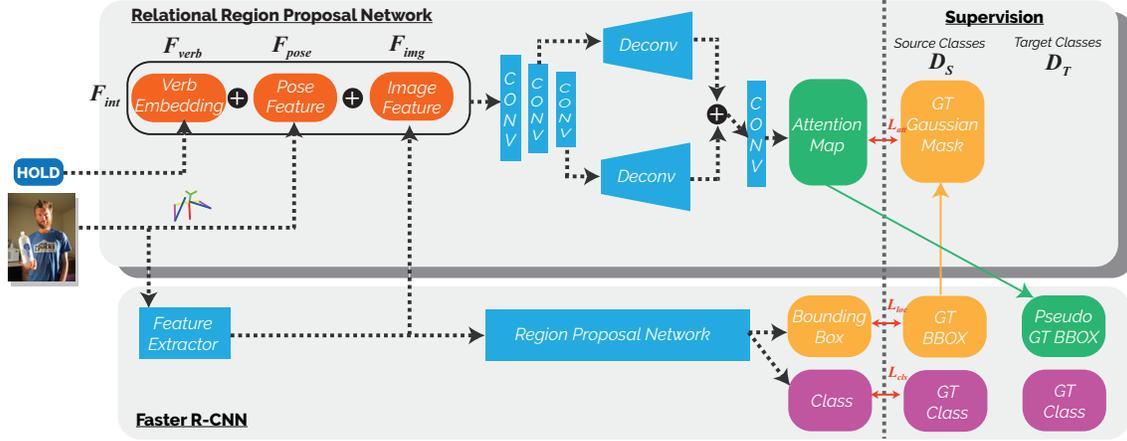


Figure 3: Overall network architecture of the proposed algorithm. Relational Region Proposal Network (RRPN) at the top is mounted on a basic Faster-RCNN model at the bottom. In RRPN, a combined feature  $F_{int}$  from verb, pose and image produces an attention map through a network which has four blocks. With source classes, the RRPN is trained with a Gaussian mask from the ground truth bounding box. However, with target classes, the RRPN generates a pseudo ground truth bounding box so that Faster-RCNN can be optimized.

and can share the backbone network with other model for image features to improve memory efficiency.

As mentioned above, RRPN predicts attention map  $\tilde{A}_i^j$  for a given image  $I_i$  using a multi-domain feature map  $F_{int}^{i,j} \in \mathbb{R}^{C \times H \times W}$  as an input, where  $C, W, H$  are the depth, width and height of the feature map.  $F_{int}$  is obtained by

$$F_{int}^i = \{F_{int}^{i,j}\}_{j=1}^M = \{(F_{img}^j \oplus F_{pose}^j \oplus F_{verb}^j)\}_{j=1}^M, \quad (2)$$

where,  $F_{img}, F_{pose}, F_{word}$  are the image feature, pose feature, and verb feature obtained by their corresponding models  $f_{img}(I_i, \theta_{img}), f_{pose}(I_i, \theta_{pose}), f_{word}(H_{act}^j, \theta_{verb})$ , and  $\oplus$  is the matrix concatenation. Here,  $\theta_x$  is the corresponding model parameters. The convolution operation for  $F_{int}^i$  computes the object existence probability for a combination of  $F_{img}^j, F_{pose}^j$ , and  $F_{verb}^j$  at a specific location on  $I_i$ .

As in (2), we used three feature maps to utilize contexts from various domains in a given dataset, and each feature map has its own contribution. The pose and word feature are responsible for the visual context of the human’s location and action, and the distinguishable linguistic context for the human’s action, respectively. The image feature is responsible for representing the whole scene as well as the object of interest. The details for each feature maps are as follows:

**Pose feature** We use the well-known human pose estimation model, OPENPOSE (Cao et al. 2017), to extract pose features. OPENPOSE predicts the location of human body joints using image or video as an input. The output consists of channels corresponding to each joint and a channel representing background information. In this paper, we used a pose estimation model with 19 channels including 18 joints and 1 background. In order to feed distinct information of human pose to the RRPN, we exploit the 18 channels except for the background channel as the pose feature.

**Verb feature** The widely used GloVe-twitter-27B-25d model (J. Pennington and Manning 2014) is applied as the

word embedding model for the verb. Since a word is embedded into a vector, one needs to convert it into a tensor form for integration with other features. While  $F_{img}^j$  and  $F_{pose}^j$  may have different spatial-wise activations depending on  $I_i$ ,  $F_{verb}^j$  must have the same value regardless of positions. In designing  $F_{verb}^j$ , we also take this consideration into account. In order to match the spatial dimension with others, the verb feature is copied to every spatial position. So that dimension of  $F_{verb}^j$  is converted from  $\mathbb{R}^{25}$  to  $\mathbb{R}^{25 \times H \times W}$ . By stacking a depth-wise word vector at all spatial positions, we can conduct a convolution operation using the same verbal information at all position of  $F_{img}^j$ . Note that, among the HICO-DET datasets, tuples with ‘No interaction’ verb labels were excluded from training and validation phases for accurate evaluation of the proposed algorithm.

**Image feature** The proposed algorithm makes use of the representative two-stage object detection model, Faster-RCNN. It consists of a feature extractor, a back-bone network, and a region proposal network (RPN). The output feature map of the back-bone network of the Faster-RCNN is used as the image feature for the RRPN. When training on the target classes, the parameters of the backbone network are reused, but the parameters of RPN are reset.

Multi-domain feature map  $F_{int}^{i,j}$  is then fed into the network to predict attention map  $A_i^j$ . In order to robustly detect objects in various sizes, we designed the network architecture which has four blocks as in (Wang and Shen 2018): an Encoder block  $f_{en}$ , two decoder blocks  $f_{de1}, f_{de2}$  and an attention block  $f_{att}$ .  $f_{en}$  takes  $F_{int}^{i,j}$  as an input and outputs two feature maps with different spatial dimensions. Then, each output feature map feeds into  $f_{de1}$  and  $f_{de2}$ , respectively. The output feature maps of  $f_{de1}$  and  $f_{de2}$  having the same spatial dimension are concatenated and inputted to the

attention block resulting in an attention map  $\tilde{A}_i^j$  as

$$\tilde{A}_i^j = f_{att}[f_{de1}\{f_{en}^1(F_{int}^{i,j})\} \oplus f_{de2}\{f_{en}^2(F_{int}^{i,j})\}]. \quad (3)$$

The output of RRPN is an attention map which emphasizes the location where the object is likely to be located. To train attention maps, we create a Gaussian map  $A_i^j$ , as a ground truth attention map, using  $O_{bbox}^j$ . RRPN is trained using  $A_i^j$  as the label. We use pixel-wise binary cross entropy loss (BCE)  $L_{att}$  between  $(A_i^j, \tilde{A}_i^j)$ .

The total loss for training on the source classes including RRPN and object detector is shown below:

$$L_{total} = L_{det} + \lambda L_{att}, \quad L_{det} = L_{cls} + L_{loc} \quad (4)$$

where,  $\lambda$  is a hyper-parameter balancing between the two losses and  $L_{det}$  is the loss for the Faster-RCNN. In the object detector point of view, the proposed algorithm on  $D_S$  is trained in the same way as the conventional supervised object detection algorithms.

### Training On the Target classes $D_T$

The object detector for  $D_T$  should be trained without  $O_{bbox}$ . Therefore, we define this problem as a weakly supervised object detection (WSOD) problem. We use  $\tilde{O}_{bbox}$  as an alternative to the missing  $O_{bbox}$  utilizing RRPN learned in the source classes training phase. It is expected that the trained RRPN can predict locations of unseen objects i.e.  $D_T$ , since it is trained to predict the object location using a human pose, an action (verb) and an image feature. The training process on  $D_T$  using the trained RRPN is as follows:

The  $F_{pose}$ ,  $F_{verb}$ , and  $F_{img}$  are fed into the trained RRPN. We apply a threshold to obtain a pseudo bounding box from the output attention map as

$$\tilde{A}_i^j = \begin{cases} 1, & \text{if, } \tilde{A}_i^j > \delta \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

where,  $\delta$  is a pre-defined threshold. The largest bounding box containing a valid value in  $\tilde{A}_i^j$  is called  $\tilde{O}_{bbox}^j$ . The pseudo ground truth bounding boxes  $\{\tilde{O}_{bbox}^j\}_{j=1}^M$  obtained from the attention maps  $\{\tilde{A}_i^j\}_{j=1}^M$  of all tuples in the image  $I_i$  are collected together and used as bounding box labels  $\tilde{O}_{bbox}$  for training an object detector. In this step, a different type of object detector from the one trained in  $D_S$  training phase can be used for training. The object detector is trained to minimize detection loss using  $O_{cls}$  and  $\tilde{O}_{bbox}$ .

## Experiment

In this section, we evaluate the performance of the proposed WSOD algorithm. To the best of our knowledge, no previous studies have been conducted on the relationship between object detection and HOI. Nevertheless, we conducted the performance comparison with prior works on HICO-DET.

### Dataset and Pre-processing

HICO-DET dataset consists of 47,776 images (38,118 training and 9,658 testing) classified into 117 actions (verb) and

80 object classes, and the object classes are the same as MS-COCO dataset. The ground truth labels consist of a tuple of  $(H_{act}, O_{bbox}, O_{cls})$  as in (1). Note that, the RRPN is trained based on tuples, so images containing multiple tuples are fed multiple times. The total number of tuples is 151,276 (117,871 training and 33,405 testing), and we use 131,560 tuples (102,450 training and 29,110 testing) excluding the tuples corresponding to the action label ‘no-interaction’.

In order to construct the problem environment, the whole dataset is divided into source and target datasets according to the frequency of the object class. Our basic experiment is set up with 116 verbs excluding ‘no interaction’. The number of object classes are 70 for  $D_S$  and 10 for  $D_T$ . In order to more clearly show the effectiveness of RRPN, in the experiment for qualitative result, we use 5 verbs and 10  $D_S$  and 70  $D_T$ . Other hyperparameters remain the same as the basic set up.

We also verified the proposed algorithm on V-COCO dataset for qualitative analysis. The purpose of evaluation on the V-COCO dataset is to show that the knowledge can be transferred from one dataset to other. Details on both datasets are described in the supplementary material.

### Metrics

We use mean Average Precision (mAP) and Recall as evaluation metrics. Because RRPN produces one bounding box for one tuple (action), Recall is used to measure how accurate the location of an object corresponding to an action is. In other words, Recall evaluates the objectness of  $\tilde{A}$  predicted by RRPN, and is calculated as the ratio of tuples for which IoU > 0.5. On the other hand, the object detector detects all the objects in an image at once. Therefore, we use the mAP in measuring the performance of Faster-RCNN which are the standard metrics for object detectors.

$D_S$  and  $D_T$  in Recall are the performance of the RRPN’s agent after the training on  $D_S$ . When training on  $D_T$ , RRPN is fixed and not trained. Note that Recall is measured on test set for  $D_S$  and on both training and test set for  $D_T$ .

### Comparison with prior works

We conduct experiments to compare with prior works on HICO-DET as shown in Table 1. First two columns represent overall results of original algorithms in AD (Yang et al. 2019) and PCL (Tang et al. 2018). However, both results are only able to show performance of all object classes with an entire dataset. Since our method is designed for transfer learning, we experiment to validate PCL on each of source and target domains. As a result, our best model with image, pose and verb has 17.19% which is 4 times better than the result of PCL on  $D_T$ . Moreover, our model only with the image feature outperforms PCL on  $D_T$ . Although a direct apple-to-apple comparison is difficult, we can see that our method is far better than the compared methods.

### Comparison with different feature combination

We experiment to verify the performance of different feature combinations. We train and test the RRPN using the same types of feature for both  $D_S$  and  $D_T$  in each experiment. Table 2 shows the performances of RRPN and Faster-RCNN as

Table 1: Comparison of the mAP with other WSOD algorithms on HICO-DET. (PCL\* is tested by ourselves, § is trained on the entire dataset and † is trained on  $D_S$  and  $D_T$  separately. I :  $F_{img}$ , P :  $F_{pose}$ , V :  $F_{verb}$ , W : Weakly supervised object detection,  $\lambda = 10$ ,  $\delta = 0.1$ )

Methods	AD	PCL	PCL*		Ours (I)	Ours (I+P+V)
$D_S(W)$	-	-	4.80 <sup>§</sup>	5.01 <sup>†</sup>	-	-
$D_T(W)$	-	-	0.01 <sup>§</sup>	4.75 <sup>†</sup>	9.57	17.19
Total	5.39	3.62	4.42 <sup>§</sup>	-	-	-

Table 2: Performance comparison of different feature combination. (Notations,  $\lambda$ , and  $\delta$  are the same as Table 1)

I	P	V	Recall@.5 (RRPN)		mAP@.5 (Faster-RCNN)		
			$D_S$ (%)	$D_T$ (%)	$D_S$	$D_T(W)$	$D_T(S)$
✓	✓	✓	<b>47.69</b>	<b>28.64</b>	<b>30.34</b>	<b>17.19</b>	29.37
✓			42.00	22.75	23.57	9.57	22.07
✓	✓		41.42	22.13	24.17	10.07	25.15
✓		✓	46.34	23.84	29.97	16.34	<b>30.28</b>

Recall and mAP, respectively, using different combinations of features.  $D_T(W)$  in mAP is the results of our WSOD, and  $D_S$  and  $D_T(S)$  are the results of full supervision.

Our full model combining all three features in the top of Table 2 shows the highest performance in both Recall and mAP among all combinations. The mAP of  $D_T(W)$  has 17.19%, which is 7.62% better than image-only model and the Recall of the  $D_T$  is 28.64% which is about 5% higher than other combinations. Moreover, the mAP score of our full model is only 4.88% lower than image-only fully supervised model in  $D_T(S)$ . Compared to models of full supervision  $D_T(S)$ , we believe that the mAP score of our full model can meaningfully show that it can be trained despite weak supervision of rare classes.

In the middle, using  $F_{img}$  alone, Recall for  $D_T$  has 22.75% and mAP ( $D_T(W)$ ) are much lower than mAP ( $D_T(S)$ ). It means that RRPN could not be trained solely by  $F_{img}$ . The two results in the bottom are the performance for combined features. When  $F_{pose}$  is combined with  $F_{img}$ , Recall degrades and mAP increases slightly. It can show that  $F_{pose}$  that is extracted from an image is redundant unless it interacts with a verb. Combining  $F_{verb}$  with  $F_{img}$ , however, Recall and mAP significantly increase and mAP( $D_T(S)$ ) also increases. It is interesting that it might be more effective for not only RRPN but also Faster-RCNN when using combination of features from other domain.

### Comparison with different $\lambda$ and $\delta$

In experiments in Table 3, we focus on verifying the effect of shared parameters such as  $\lambda$  in (4) and  $\delta$  in (5).

In top of Table 3, according to the change of the  $\lambda$  in (4), the ratio of the loss weight in RRPN is determined. When  $\lambda$  is zero, due to untrained RRPN, Recall and mAP for  $D_T(W)$  have the lowest score while mAP for  $D_S$  has the highest score. On the other hand, Recall and mAP are the highest at  $\lambda = 10$  with performance improvements of 17.45% and 15.58% compared to  $\lambda = 0$ , respectively. On the contrary,

Table 3: Comparison of quantitative result of  $\lambda$  and  $\delta$  (Notations are the same as Table 1)

parameter		Recall@.5 (RRPN)		mAP@.5 (Faster-RCNN)		
$\lambda$	$\delta$	$D_S$ (%)	$D_T$ (%)	$D_S$	$D_T(W)$	$D_T(S)$
0	0.1	11.64	11.19	<b>31.06</b>	1.61	25.57
1	0.1	41.51	24.46	23.87	9.27	25.43
5	0.1	46.37	23.37	30.10	14.38	26.11
10	0.1	<b>47.69</b>	<b>28.64</b>	30.34	<b>17.19</b>	<b>29.37</b>
15	0.1	46.65	23.07	23.32	15.85	25.45
20	0.1	43.17	26.00	22.68	15.75	22.92
10	0.05	<b>48.22</b>	<b>29.41</b>	30.27	9.41	25.04
10	0.10	47.69	28.64	30.34	<b>17.19</b>	29.37
10	0.15	39.67	17.96	<b>30.40</b>	14.01	26.66
10	0.20	34.14	16.72	30.22	13.24	<b>30.39</b>

on some levels of  $\lambda$ , we can see that the performance degradation for not only  $D_T(W)$  but also  $D_S$ .

This can be understood as an effect of parameter sharing for image feature extractor between RRPN and an object detector. As mentioned earlier, RRPN is a universal add-on type module which can be adapted to various computer vision tasks. To effectively utilize these advantages, we share the backbone network of RRPN and the object detector in consideration of memory efficiency. Therefore, the RRPN and the object detector affect each other through the backbone network during training.

In bottom of Table 3, according to the  $\delta$  in (5), the size of the pseudo bounding box is determined. A small  $\delta$  makes the size of the boxes increase, while a large  $\delta$  makes the box small or disappear. As  $\delta$  increases, partial information of the object is trained. For example, in the case of an apple, only the central part of the apple is trained with high  $\delta$ , which causes many false positive. On the other hand, lowering the threshold of the box, Faster-RCNN is trained not only with an object but also with backgrounds. It is interesting that  $\delta$  affect differently to both metrics where Recall gets higher when  $\delta$  gets smaller but mAP get the highest score when  $\delta = 0.1$ . We believe that the RRPN can easily learn objectness with a larger box due to small  $\delta$ , but classification of objects could be more difficult due to inaccurate localization. Therefore, too small or too large  $\delta$  causes a degradation of mAP, and we have found the suitable value,  $\delta = 0.1$ , through the experiments by selecting the value with the highest performance in  $D_T(W)$ .

### Qualitative results

Fig. 4 shows the qualitative results of the proposed algorithm on  $D_T$ . The first column indicates input images and the last column indicates output attention maps inferred by the corresponding actions. We can see that RRPN predicts an accurate attention map on unseen object classes in  $D_T$ . Furthermore, it can be seen that the pattern of the predicted attention map differs depending on the verb. For example, while 'hold' shows a strong activation value near the human hand, 'ride' tends to activate at the bottom of a person. Based on this, we can confirm that the object location can be estimated based on the interaction between the verb and

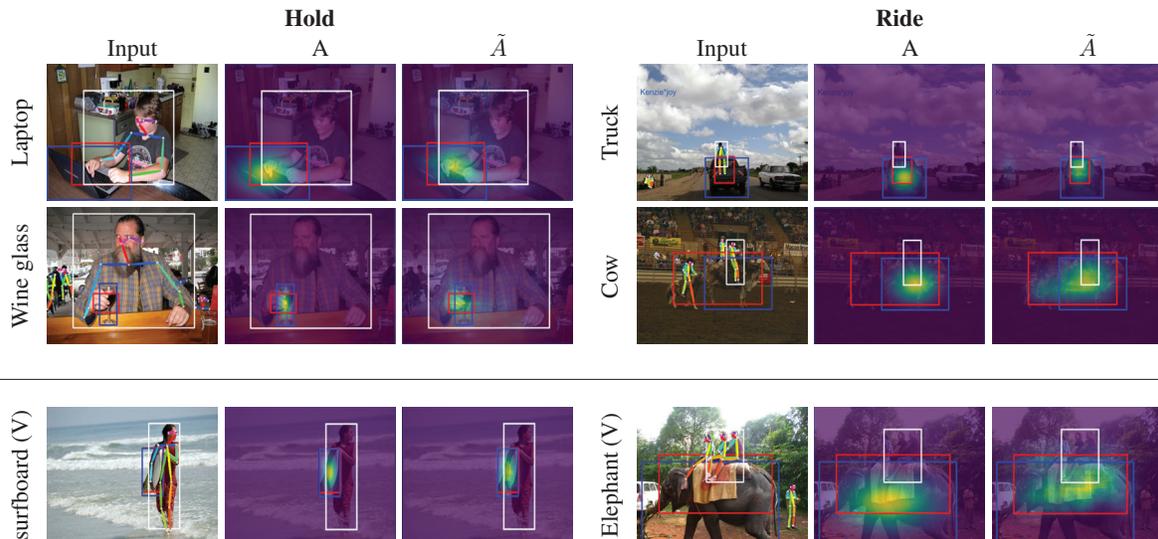


Figure 4: (Left) Input image with pose, (middle) ground truth Gaussian attention mask (A) in yellow, and (Right) predicted attention map ( $\hat{A}$ ). Red box is pseudo object box, blue box is ground truth and white box indicates the human in action. (V) the last row is the result on the V-COCO dataset. Note that a white box is used solely for visually representing an acting human in an image and is not used in training on  $D_T$ .

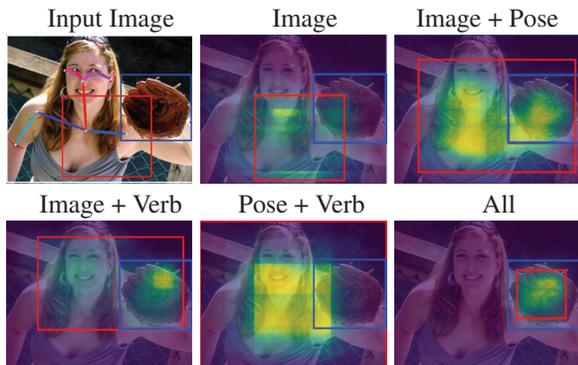


Figure 5: Comparison of predicted attention maps trained only by the image feature and by various integrated features with [glove, hold]. The predicted attention maps show different activations depending on the role of each feature map.

the pose. The role of the pose can be found in the example of [Truck, Ride]. Despite that two trucks exist in an image, the activation of a truck on which the human is riding shows stronger than the other. This can be seen as a contribution of  $F_{pose}$  to the object localization. We also verified the performance of RRPN on the object from a different dataset, V-COCO. The RRPN is trained using  $D_S$  of HICO-DET and predicts the attention map of  $D_T$  of V-COCO. The bottom row shows the predicted attention map on V-COCO. We can see that the proposed algorithm can also predict the object location accurately on images even from other datasets.

Fig. 5 depicts the comparison of predicted attention map between different feature combinations on [glove, hold]. As

described in section 5, the pattern of the resulting attention map can be changed by the combination of features. Since “Glove” is an unseen class, backbone has no information to extract reliable feature, so that RRPN cannot predict the location of an object accurately using only  $F_{img}$ . However, if RRPN is trained using more than two features including  $F_{img}$ , RRPN can infer the location of an object based either on  $F_{pose}$  or on  $F_{verb}$ . Specifically,  $F_{img} + F_{verb}$  predicted a more distinguishable attention map for an object, compared to  $F_{img} + F_{pose}$  feature map. Since  $F_{img}$  and  $F_{pose}$  are extracted from the same image, some of the information can be redundant between two features. On the other hand,  $F_{verb}$  is able to provide useful information to  $F_{int}$  because it is extracted from a different domain, language. Consequently, the location of an object can be predicted precisely when we use all three features. On the contrary, if RRPN trained using only  $F_{pose}$  and  $F_{verb}$  without  $F_{img}$ , the output attention map only activates around the human. Thus, it can be understood that  $F_{verb}$  plays a role of providing supplementary information to  $F_{img}$  about the object of interest.

## Conclusion

In this paper, we proposed a novel weakly-supervised scheme for object detection problems. We introduced the RRPN which can universally localize objects in an image with information on human poses and action verbs. Using transferable knowledge from the RRPN, we can continuously train any object detector for unseen objects with weak verbal supervision describing HOI. We validated our method based on the results on HICO-DET dataset and the performances show the possibility of our method for a new WSOD training scheme. Our work shows sufficient potentials to overcome the inefficiency of the supervised training scheme

in recent deep learning. Also, we can develop our method in the direction to the continual learning since we already suggested a novel method to transfer common knowledge to localize objects with HOI.

## Acknowledgments

This work was supported by Next-Generation Information Computing Development Program through the NRF of Korea (2017M3C4A7077582) and Promising-Pioneering Researcher Program through Seoul National University(SNU) in 2015.

## References

- Bearman, A.; Russakovsky, O.; Ferrari, V.; and Fei-Fei, L. 2016. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, 549–565. Springer.
- Bilen, H., and Vedaldi, A. 2016. Weakly supervised deep detection networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Cao, Z.; Simon, T.; Wei, S.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*.
- Chao, Y.-W.; Wang, Z.; He, Y.; Wang, J.; and Deng, J. 2015. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, 1017–1025.
- Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 381–389. IEEE.
- Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, 379–387.
- Diba, A.; Sharma, V.; Pazandeh, A.; Pirsiavash, H.; and Van Gool, L. 2017. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 914–922.
- Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2):303–338.
- Gkioxari, G.; Girshick, R.; Dollár, P.; and He, K. 2018. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8359–8367.
- Gupta, A., and Davis, L. S. 2007. Objects in action: An approach for combining action understanding and object perception. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.
- Gupta, S., and Malik, J. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.
- Gupta, A.; Kembhavi, A.; and Davis, L. S. 2009. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(10):1775–1789.
- J. Pennington, R. S., and Manning, C. 2014. Glove: Global vectors for word representation. *conference on empirical methods in natural language processing* 1532–1543.
- Jie, Z.; Wei, Y.; Jin, X.; Feng, J.; and Liu, W. 2017. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1377–1385.
- Kantorov, V.; Oquab, M.; Cho, M.; and Laptev, I. 2016. Context-locnet: Context-aware deep network models for weakly supervised localization. In *European Conference on Computer Vision*, 350–365. Springer.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.
- Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *CVPR*, 685–694.
- Qi, S.; Wang, W.; Jia, B.; Shen, J.; and Zhu, S.-C. 2018. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 401–417.
- Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *NIPS*.
- Shi, M.; Caesar, H.; and Ferrari, V. 2017. Weakly supervised object localization using things and stuff transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, 3381–3390.
- Tang, P.; Wang, X.; Bai, X.; and Liu, W. 2017. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2843–2851.
- Tang, P.; Wang, X.; Bai, S.; Shen, W.; Bai, X.; Liu, W.; and Yuille, A. L. 2018. Pcl: Proposal cluster learning for weakly supervised object detection. *TPAMI*.
- Uijlings, J.; van de Sande, Gevers, T.; and Smeulders, A. 2013. Selective search for object recognition. In *International Journal of Computer Vision*, 154–171.
- Uijlings, J.; Popov, S.; and Ferrari, V. 2018. Revisiting knowledge transfer for training object class detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1101–1110.
- Wang, W., and Shen, J. 2018. Deep visual attention prediction. *IEEE Transaction on Image Processing* 27(5):2368–2378.
- Yang, Z.; Mahajan, D.; Ghadiyaram, D.; Nevatia, R.; and Ramanathan, V. 2019. Activity driven weakly supervised object detection. *CVPR*.
- Yao, B., and Fei-Fei, L. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 17–24. IEEE.
- Zhu, Y.; Zhou, Y.; Ye, Q.; Qiu, Q.; and Jiao, J. 2017. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 1841–1850.