




# Weakly-supervised Human-object Interaction Detection

Masaki Sugimoto<sup>1</sup><sup>a</sup>, Ryosuke Furuta<sup>2</sup><sup>b</sup> and Yukinobu Taniguchi<sup>1</sup><sup>c</sup>

<sup>1</sup>*Department of Information and Computer Technology, Tokyo University of Science, Tokyo, Japan*

<sup>2</sup>*Institute of Industrial Science, The University of Tokyo, Tokyo, Japan*

**Keywords:** Weakly-supervised Learning, Human-Object Interaction, Object Detection.

**Abstract:** Human-Object Interaction detection is the image recognition task of detecting pairs (a person and an object) in an image and estimating the relationships between them, such as “holding” or “riding”. Existing methods based on supervised learning require a lot of effort to create training data because they need the supervision provided as Bounding Boxes (BBs) of people and objects and verb labels that represent the relationships. In this paper, we extend Proposal Cluster Learning (PCL), a weakly-supervised object detection method, for a new task called weakly-supervised human-object interaction detection, where only the verb labels are assigned to the entire images (i.e., no BBs are given) during the training. Experiments show that the proposed method can successfully learn to detect the BBs of people and objects and the verb labels between them without instance-level supervision.


## 1 INTRODUCTION


Human-Object Interaction detection (HOI detection) is a task to detect the pairs of a person and an object with their Bounding Boxes (BBs) in the input image and to estimate the relationships between them, such as “holding” and “riding”. HOI detection can provide more detailed scene understanding than conventional object detection. The detection results are utilized for enhancing a variety of applications such as automatic caption generation, person identification, and surveillance camera systems. Unfortunately, existing HOI detection methods based on supervised learning require excessive amounts of labor and time to create the training data needed (Gupta and Malik, 2015).


In order to reduce the effort of creating training data, a number of weakly-supervised learning methods have been proposed for object detection (Bilen and Vedaldi, 2016; Tang et al., 2017; Wan et al., 2018; Tang et al., 2018). They are trained to detect instance BBs from image-level labels that represent which classes exist in the image; in other words, their training does not require instance-level BB annotations. (Bearman et al., 2016) showed that it takes 10 seconds per instance to annotate the BB of an object while it takes only 1 second per class to anno-

tate which class is included in the image. However, for HOI detection, although many supervised learning methods have been proposed that learn from a set of BBs and verb labels that represent their relationships, none can be trained using a set of just verb labels assigned to the entire set of images.

Therefore, in this paper, we tackle the new task of weakly-supervised HOI detection, where only labels of person-object relationships (verb labels) are assigned to the images, and no BBs are given for training (Figure 1). To the best of our knowledge, this is the first attempt to tackle weakly-supervised learning for HOI detection. We propose to extend Proposal Cluster Learning (PCL) (Tang et al., 2018), a weakly-supervised object detection method, to realize weakly-supervised HOI detection. The original PCL extracts a feature vector of each object candidate region by ROI pooling (Ren et al., 2015) and calculates its detection score through a subsequent network. In the proposed method, we obtain person regions in advance (e.g., by using an off-the-shelf human detector) and calculate the detection score for each pair of a person and an object candidate by adding their feature vectors. In the experiments on the V-COCO dataset (Gupta and Malik, 2015), the proposed method achieves CorLoc of 10.7% on the training set and mAP of 7.08% on the test set.

<sup>a</sup>  <https://orcid.org/0000-0001-9315-3804>

<sup>b</sup>  <https://orcid.org/0000-0003-1441-889X>

<sup>c</sup>  <https://orcid.org/0000-0003-3290-1041>

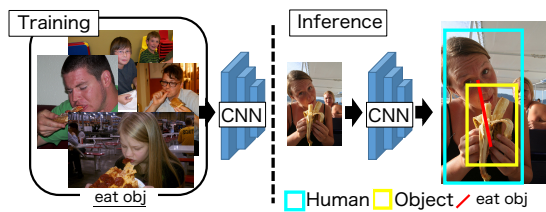


Figure 1: Overview of weakly-supervised human object interaction detection.

## 2 RELATED WORKS

In this section, we summarize weakly-supervised object detection methods and supervised HOI detection methods related to our work because there are no weakly-supervised HOI detection methods.

### 2.1 Weakly Supervised Object Detection

Weakly-supervised object detection is a task to detect instances by learning from image-level labels that represent which classes are included in the image, instead of learning from BBs and class labels assigned to each BB (instance-level labels). Weakly-supervised learning is expected to reduce the time and effort needed in creating training data. There are two types of weakly-supervised object detection methods; one is based on the Multiple Instance Detection Network (Bilen and Vedaldi, 2016) and the other is based on solving the entropy minimization problem (Wan et al., 2018). We focus on the former type here because the proposed method is based on it.

Bilen et al. proposed the Multiple Instance Detection Network (MIDN) (Bilen and Vedaldi, 2016) as the first example of weakly-supervised object detection. MIDN is able to learn object detection by predicting whether each class is present or not in the input image. The prediction scores are calculated from the scores normalized for each object candidate region and the ones normalized for each class. Tang et al. proposed a method called Online Instance Classifier Refinement (OICR) (Tang et al., 2017) based on MIDN. OICR improves the accuracy of object detection by proposing an online training method that generates pseudo-ground truth BBs for each label. Tang et al. improved its performance by a method called Proposal Cluster Learning (PCL) (Tang et al., 2018), which refines the pseudo-ground truth BB generation.

All of the above methods are tailored for object detection. In this paper, we extend PCL to weakly-supervised HOI detection.

### 2.2 Human Object Interaction Detection

HOI detection is an image recognition task proposed by Gupta and Malik (Gupta and Malik, 2015). Typical HOI detection methods first detect human and object areas in the regular way. After that, they predict the relationships between people and objects (verb labels). The purpose of HOI detection is to understand a scene in detail.

Most HOI detection methods learn the appearance features of the person and object regions and their spatial relationships. Gupta and Malik (Gupta and Malik, 2015) proposed an early method of HOI detection, which narrowed down the coordinates of the corresponding object region based on the verb label from the person region. The final outputs are obtained by combining the narrowed search range and the results of object detection. (Gkioxari et al., 2018) proposed a network based on Faster R-CNN, which predicts the BBs of people and objects and the verb labels by considering their relations through utilizing an object detection branch, human-centric branch, and interaction branch. (Gao et al., 2018) proposed a network with attention modules based on existing methods, where the attention maps are generated from the entire feature maps. Recently, (Ulutan et al., 2020) proposed a network that has a spatial attention branch and a graph convolutional network to effectively consider contextual information and explicitly model structural relations. This method outperforms the other state-of-the-art methods by a large margin.

As mentioned above, various methods have been proposed for HOI detection. However, all of them are based on supervised learning, and there are no weakly-supervised methods for HOI detection.

The work most relevant to ours is (Yang et al., 2019). However, their problem setting is different from ours in that their objective is to learn object detection, not HOI detection, by using human keypoints and action labels as strong cues.

## 3 PROPOSAL CLUSTER LEARNING (PCL)

Because the proposed method is based on PCL (Tang et al., 2018), we describe it in detail in this section.

PCL is a weakly-supervised object detection method consisting of Multiple Instance Detection Network (MIDN), Online Instance Classifier Refinement (OICR), and Proposal Cluster Learning. Hereafter, we differentiate “PCL” from “PCL part”, where the former refers to the entire method for weakly-supervised object detection proposed by (Tang et al.

2018), and the latter refers to the Proposal Cluster Learning part of the former.

### 3.1 Learning

Here, we describe the training procedure of MIDN and omit those of OICR part and PCL part because we extend the MIDN part in PCL to weakly-supervised HOI detection. Let  $C$  be the number of classes trained.

#### 1. Extracting the Features of Candidate Region.

First, MCG (Arbeláez et al., 2014) is used to generate  $R$  object candidate regions. For each candidate object region, VGG16 (Simonyan and Zisserman, 2015) and RoI Pooling (Ren et al., 2015) are used to extract the 4,096-dim feature vector  $\mathbf{f}_r (r = 1, \dots, R)$ .

#### 2. Concatenating Feature Vectors.

Next, we transform  $\mathbf{f}_r$  into a  $C$ -dim vector using a fully-connected (fc) layer and create matrix  $\mathbf{X}^a \in \mathbb{R}^{C \times R}$  by concatenating the  $R$  vectors ( $r = 1, \dots, R$ ). In the same manner, we create another matrix  $\mathbf{X}^b \in \mathbb{R}^{C \times R}$  using a different fc layer.

#### 3. Normalizing each of $\mathbf{X}^a$ and $\mathbf{X}^b$ along Different Directions.

Then, we normalize  $\mathbf{X}^a$  within each candidate region using softmax function  $[\sigma(\mathbf{X}^a)]_{cr} = \frac{\exp(X_{cr}^a)}{\sum_{c'=1}^C \exp(X_{c'r}^a)}$ . In contrast, we normalize  $\mathbf{X}^b$  within each class  $[\sigma(\mathbf{X}^b)]_{cr} = \frac{\exp(X_{cr}^b)}{\sum_{r'=1}^R \exp(X_{c'r'}^b)}$ .

#### 4. Calculating the Object Candidate Scores.

We use the element-wise product  $\mathbf{X}^{score} = \sigma(\mathbf{X}^a) \odot \sigma(\mathbf{X}^b)$  in computing the object candidate scores  $\mathbf{X}^{score} \in \mathbb{R}^{C \times R}$ . Each element  $X_{cr}^{score} (r = 1, \dots, R$  and  $c = 1, \dots, C)$  represents the prediction score of the  $r$ -th object candidate region for the  $c$ -th class.

#### 5. Calculating the Losses.

By taking the sum of the scores of all object candidate regions,  $\phi_c = \sum_{r=1}^R X_{cr}^{score}$ , we can obtain the probability  $\phi_c$  of the presence of class  $c$  in the image. MIDN loss  $L_b^{MIDN}$  is calculated between  $\boldsymbol{\phi} = [\phi_1, \phi_2, \dots, \phi_C]^T \in \mathbb{R}^{C \times 1}$  and the image-level label supervision  $\mathbf{y} = [y_1, y_2, \dots, y_C]^T \in \mathbb{R}^{C \times 1}, y_c \in \{0, 1\}$  by the cross entropy loss function in Eq.(1).

$$L_b^{MIDN} = - \sum_{c=1}^C \{y_c \log \phi_c + (1 - y_c) \log (1 - \phi_c)\}, \quad (1)$$

where  $y_c$  is the label that takes 1 if the  $c$ -th class is included in the image, and 0 otherwise.

In addition to the training of MIDN described above, we also train  $K$  refined classifiers online. The  $k$ -th classifier ( $k = 1, \dots, K$ ) takes  $\mathbf{f}_r (r = 1, \dots, R)$  as input and outputs the refined scores  $\mathbf{X}^{refine(k)} \in \mathbb{R}^{(C+1) \times R}$  through an fc layer, where the  $(C + 1)$ -th class indicates the background class. The  $k$ -th classifier is trained by minimizing the refinement loss,  $L^{refine(k)}$ , between the refined scores and pseudo-ground truth BBs generated in the PCL part (for details, see (Tang et al., 2018)).

### 3.2 Inference

At the time of inference, the final prediction score for each object candidate region and each class  $\mathbf{X}^{detect} \in \mathbb{R}^{(C+1) \times R}$  is calculated as the average of the outputs of the  $K$  refined classifiers as in Eq.(2).

$$\mathbf{X}^{detect} = \frac{1}{K} \sum_{k=1}^K \mathbf{X}^{refine(k)} \quad (2)$$

## 4 PROPOSED METHOD

In HOI detection, it is necessary to detect person-object pairs and to predict the classes that represent their relationships. In this paper, we extend the MIDN part of PCL for HOI detection. An overview of the proposed method is shown in Figure 2. The person regions can be detected by employing an off-the-shelf person detector. However, the experiments in Section 5 assume that all person regions have been detected perfectly and use ground truth BBs as inputs to the proposed method. In this section, we detail the training of the proposed method.

### 4.1 Learning

#### 1. Extracting the Features of Candidate Region.

Object proposal generators such as MCG are used to detect  $R$  candidate object regions. For each candidate object region, VGG16 and RoI Pooling are used to extract the 4,096-dim feature vector  $\mathbf{f}_r (r = 1, \dots, R)$ .

#### 2. Extracting the Features of Human Region.

For each of the  $N$  human regions detected, VGG16 and RoI Pooling are used to extract the 4,096-dim feature vector  $\mathbf{g}_n (n = 1, \dots, N)$  in the same way.

#### 3. Adding the Feature Vectors.

This is the key point of the proposed extension of MIDN part to HOI detection. To obtain feature vectors that represent a person-object pair, we add together the feature vector  $\mathbf{f}_r$  of the  $r$ -th object candidate region and the feature vector  $\mathbf{g}_n$  of the  $n$ -th person

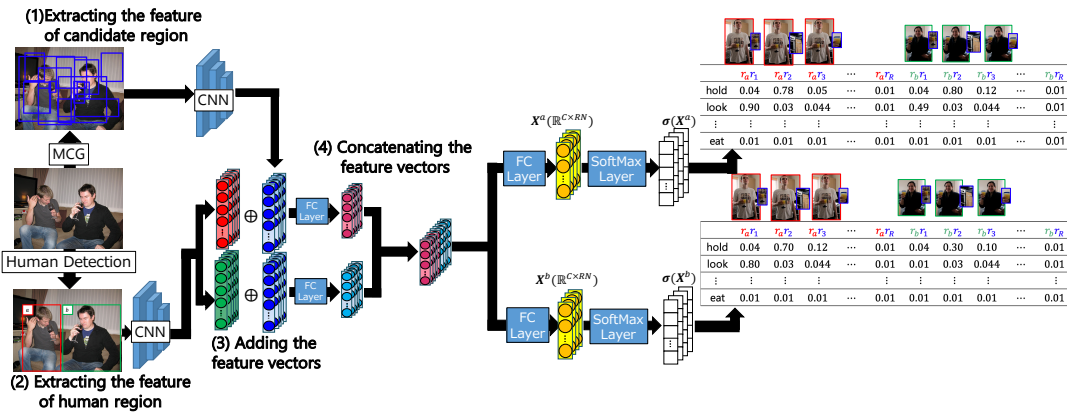


Figure 2: Overview of proposed method.

region  $\mathbf{h}_{\{r,n\}} = \mathbf{f}_r + \mathbf{g}_n$ . Initially we tried concatenating the two vectors instead of adding them, but this yielded lower performance than adding them.

The following flow is the same as 2-5 in Section 3.1.

4. **Concatenating the Feature Vectors.** Two types of matrices  $\mathbf{X}^a, \mathbf{X}^b \in \mathbb{R}^{C \times RN}$  are created by transforming  $\mathbf{h}_{\{r,n\}}$  into a  $C$ -dim vector through an fc layer and concatenating the output vectors for all the pairs of object candidate region ( $r = 1, \dots, R$ ) and person region ( $n = 1, \dots, N$ ).
5. **Normalizing each of  $\mathbf{X}^a$  and  $\mathbf{X}^b$  along Different Directions.** Then, we normalize  $\mathbf{X}^a$  within each candidate region using softmax function  $[\sigma(\mathbf{X}^a)]_{c\{r,n\}} = \frac{\exp(X_{c\{r,n\}}^a)}{\sum_{c'=1}^C \exp(X_{c'\{r,n\}}^a)}$ . In contrast, we normalize  $\mathbf{X}^b$  within each class  $[\sigma(\mathbf{X}^b)]_{c\{r,n\}} = \frac{\exp(X_{c\{r,n\}}^b)}{\sum_{r'=1}^R \sum_{n'=1}^N \exp(X_{c\{r',n'\}}^b)}$ .
6. **Calculating the Object Candidate Scores.** We use the element-wise product  $\mathbf{X}^{score} = \sigma(\mathbf{X}^a) \odot \sigma(\mathbf{X}^b)$  in computing the object candidate scores  $\mathbf{X}^{score} \in \mathbb{R}^{C \times RN}$ . Each element  $X_{c\{r,n\}}^{score}$  ( $r = 1, \dots, R$ ,  $c = 1, \dots, C$  and  $n = 1, \dots, N$ ) represents the prediction score of each pair of a person and an object candidate for the  $c$ -th class.
7. **Calculating the Losses.** By taking the sum of the scores of all pairs,  $\phi_c = \sum_{r=1}^R \sum_{n=1}^N X_{c\{r,n\}}^{score}$ , we can obtain the probability  $\phi_c$  of class  $c$  in the image. We train the network to reduce the same loss function as in Eq.(1).

In addition to the training the proposed method described above, we also train  $K$  refined classifiers online. The  $k$ -th classifier ( $k = 1, \dots, K$ ) takes  $\mathbf{h}_{\{r,n\}}$  as input and outputs the refined scores  $\mathbf{X}^{refine(k)} \in \mathbb{R}^{(C+1) \times RN}$  through an fc layer, where the  $(C+1)$ -th class indicates the background class. The

$k$ -th classifier is trained by minimizing the refinement loss  $L^{refine(k)}$  between the refined scores and pseudo-ground truth BBs generated through the PCL part.

## 4.2 Inference

As in Section 3.2, at the time of inference, the final prediction score for each pair and each class  $\mathbf{X}^{detect} \in \mathbb{R}^{(C+1) \times RN}$  is calculated as the average of the outputs from the  $K$  refined classifiers as in Eq.(2).

## 5 EXPERIMENTS SETTINGS

In this section, we describe the settings of the experiments conducted to evaluate the performance of the proposed method.

### 5.1 Dataset

V-COCO (Gupta and Malik, 2015) is one of the datasets for HOI detection, and consists of 10,310 images from the COCO (Lin et al., 2014) dataset. A BB is annotated to each person and each object. No category names such as racket or skate-board are given to the objects. If the person interacts with the object in the image, one or more labels that represent their relationships are assigned to the pair of the person and the object. Although the original V-COCO dataset has 26 different action labels (e.g., hold, hit, sit, and so on), four labels (smile, stand, run, and walk) are generally not used for HOI detection because they do not signify object interaction. The label (point) is also not used because it rarely occurs in the dataset. Each of three action labels (cut, eat, and hit) is divided into two different verb labels (i.e., cut obj, cut instr, eat obj, eat instr, hit obj, and hit instr) according to

Table 1: Labels used for weakly-supervised HOI detection.

eat obj	drink	talk on phone	ride	read
eat instr	catch	work on computer	carry	ski
throw	skateboard	cut obj	hit instr	hit obj
surf	snowboard	cut instr	lay	kick

whether the interaction is with an object or an instrument. As a result, 24 verb labels are generally used for HOI detection.

In this experiment, we additionally removed some labels through the following process. We first removed the labels assigned to the entire images, not to the BBs of pairs of a person and an object. Second, we removed the pairs of BBs that had more than one verb labels because the proposed method for weakly-supervised HOI detection aims to detect one verb label for each pair of a person and an object. We also removed four labels (hold, look, jump, and sit) because they frequently co-occur with other labels in the dataset. Finally, we removed the images that had no labels. As a result, the number of classes (verb labels) used in this paper was 20 (Table 1). The numbers of images used for training and testing were 4,287 and 3,854, respectively.

## 5.2 Object Proposal Generation

In this paper, we used two types of methods to generate the object candidate regions: MCG and Faster R-CNN. Section 6 reports the performance of the proposed method when using each of them. We employed MCG because it does not need any supervision and is widely used in weakly-supervised object detection. When we used MCG, the number of candidate regions  $R$  was approximately 2,000.

We also used Faster R-CNN trained on the COCO dataset for object detection. It is important to note that the BB and label annotations for HOI detection were not used to train Faster R-CNN while those for object detection were used. It is reasonable to use the detector trained for object detection because the datasets for object detection are easier to create or obtain than those for HOI detection. When we used Faster R-CNN, we set the number of candidate regions  $R = 30$ . When the number of objects detected by Faster R-CNN in an image was less than 30, we randomly added the object candidate regions generated by MCG because  $R$  must be fixed when training the proposed method.

## 5.3 Evaluation Metric

We used the trainval set for training. Similar to weakly-supervised object detection, we report Cor-Loc (Deselaers et al., 2012) on the trainval set and mAP on the test set to evaluate the performance of the proposed method. The predicted tuple (person

BB, object BB, and verb label) is counted to be true when the following three conditions are satisfied: (i) the verb label is correctly predicted, (ii) the intersection over union (IoU) between the detected object region and ground truth BB is more than threshold  $\tau$ , and (iii) IoU between the corresponding person region and ground truth BB is more than threshold  $\tau$ . We set  $\tau = 0.5$ . Condition (iii) is always satisfied in our experiments because we assumed that the person regions were perfectly detected, i.e., we trained and tested the proposed method and the compared method (Ulutan et al., 2020) using the ground truth person BBs.

## 5.4 Implementation Details

We implemented the proposed method on PyTorch. We used the SGD optimizer to train it. The number of iterations was set to  $2.5 \times 10^4$ . We set the mini-batch size of SGD to 1. We set the learning rate to  $5.0 \times 10^{-4}$  at the start of training and changed it to  $5.0 \times 10^{-5}$  in the last  $1.0 \times 10^4$  iterations. The momentum and weight decay were set to 0.9 and  $5.0 \times 10^{-4}$ , respectively. In training the proposed method, we first trained PCL using only the object regions and verb labels (i.e., without person regions) as an object detection task, see Section 3. Then, we trained the proposed method as described in Section 4 using its weights as initial values. Due to the GPU memory limitation, five detected persons were used as input to the proposed method even if more than five persons were detected in one image.

## 5.5 Baseline Method

We implemented a baseline method for the comparisons because there are no existing methods for weakly-supervised HOI detection. In training the baseline method, we trained PCL using only the object regions and verb labels (i.e., without person regions) as an object detection task, where the training parameters were completely same as those in Section 5.4.

The inference procedure of the baseline method is as follows. First, object regions are detected by the trained PCL as an object detection task, where a verb label is predicted for each detected object region. Second, Euclidean distances between center coordinates of every pair of a person and the detected object region are calculated. Finally, the output tuples (person BB, object BB, and verb label) are obtained by choosing the closest person region to each detected object region.

## 6 RESULTS

### 6.1 Results on Trainval Set

Table 2 compares the CorLoc of the proposed method with the baseline method for each verb label on the trainval set. The third column in Table 2 shows the results when we used MCG to generate the object candidate regions. We observed that some labels such as read, kick, and hit obj were successfully detected. However, the proposed method did not work well for other labels such as catch, throw, surf, and so on. CorLoc averaged over all the verb labels (mean CorLoc) was 10.7%. Figure 3 shows cases in which the proposed method was successful. The yellow rectangles and textbox indicate the predicted object regions and labels, respectively. The dark blue rectangle shows the ground truth object regions. The red line represents the person-object pairs, and the light blue rectangle shows the corresponding person regions. We observed that the proposed method successfully detected the corresponding object regions and verb labels for the target persons.

The second column in Table 2 shows the results of the baseline method, which pairs the detected object regions and the closest person regions. The mean CorLoc of the baseline method was 3.44%, significantly lower than that of the proposed method. From the results, we observed that the proposed method successfully learned to detect HOI from appearance features.

The fourth column in Table 2 shows the results of the proposed method when we used Faster R-CNN as the object proposal generator. Compared with the results yielded with MCG, the performance was significantly improved because the object candidate regions generated by Faster R-CNN were much more accurate than those generated by MCG. In particular, we observed that some labels such as cut instr, hit instr, and skateboard were successfully detected although they were seldom detected by the proposed method with MCG.

### 6.2 Results on Test Set

Figure 4 visualizes the successful cases of the proposed method with MCG on the test set, and the



Figure 3: Successful cases of the proposed method with MCG on the trainval set.

Table 2: Comparisons of CorLoc on the trainval set.

verb	Baseline	Ours (MCG)	Ours (Faster R-CNN)
read	0.332%	21.8%	58.4%
kick	0.00%	51.6%	87.0%
drink	0.00%	2.33%	33.6%
eat instr	0.00%	0.00%	7.20%
cut obj	2.17%	25.0%	44.2%
cut instr	0.00612%	0.00208%	20.4%
hit obj	0.00%	32.6%	69.2%
catch	0.00%	0.00221%	64.6%
throw	0.00%	0.00%	0.0253%
ride	11.3%	7.13%	4.13%
ski	0.0218%	0.00%	0.0236%
lay	49.28%	58.4%	57.0%
talk on phone	0.00%	0.123%	40.0%
hit instr	0.00%	0.00%	14.7%
snowboard	0.128%	0.00162%	0.178%
eat obj	0.0424%	0.466%	55.0%
work on computer	4.29%	15.2%	32.2%
carry	1.10%	0.216%	12.7%
skateboard	0.00165%	0.000488%	32.7%
surf	0.113%	0.00110%	0.188%
Average	3.44%	10.7%	31.7%



Figure 4: Successful cases of the proposed method with MCG on the test set.

third column in Table 3 shows average precision (AP) on the V-COCO test set. Similar to the results on the trainval set (Table 2), we observed the proposed method successfully detected some labels such as read, kick and hit obj, but failed to detect other labels such as catch, throw and surf.

The second column in Table 3 shows AP of the baseline method. Similar to Table 2, the proposed method which detects HOI based on appearance features, achieved superior performance on the test set, compared to the baseline based on distances between object and person regions. The fourth column shows the results of the proposed method with Faster R-CNN. The mean of AP on all labels (mAP) was 29.6%. For comparison, in the last column of Table 3, we show the AP of VSGNet (Ulutan et al., 2020), which is a state-of-the-art method for supervised HOI detection. We ran the publicly available code of VSGNet<sup>1</sup> on the same trainval and test set. Although the mAP of the proposed method with Faster R-CNN was not as high as that of VSGNet, the proposed method outperformed on some labels such as kick, cut obj, and hit obj even though our method was trained with weakly-supervised learning.

### 6.3 Discussions

Some failure cases of the proposed method with MCG on the trainval set are shown in Figures 5, 6, and 7. Figure 5 shows the examples where the regions pre-

<sup>1</sup><https://github.com/ASMIftexhar/VSGNet>

Table 3: Comparisons of AP on the test set.

verb	Baseline	Ours (MCG)	Ours (Faster R-CNN)	VSGNet
read	0.00%	3.28%	25.1%	55.1%
kick	0.00%	47.5%	95.3%	79.3%
drink	0.00%	2.65%	25.1%	48.5%
eat instr	0.00%	0.0368%	8.04%	30.5%
cut obj	0.784%	24.6%	45.6%	27.8%
cut instr	0.00717%	0.0115%	23.1%	33.9%
hit obj	0.00%	24.3%	83.3%	42.1%
catch	0.00%	0.172%	69.3%	12.8%
throw	0.00%	0.0126%	1.25%	59.1%
ride	11.0%	9.58%	4.03%	4.95%
ski	0.0189%	0.00257%	0.703%	4.73%
lay	47.8%	16.2%	21.6%	33.8%
talk on phone	0.00%	0.692%	39.7%	77.1%
hit instr	0.00%	0.106%	16.9%	72.4%
snowboard	0.204%	0.145%	1.61%	39.0%
eat obj	0.116%	1.10%	52.2%	61.7%
work on computer	5.32%	10.8%	31.5%	44.4%
carry	0.209%	0.363%	10.7%	9.59%
skateboard	0.00%	0.0129%	35.2%	57.3%
surf	0.0463%	0.0227%	2.75%	32.3%
Average	3.28%	7.08%	29.6%	41.3%

dicted by the proposed method were larger than the ground truth regions. Figure 6 shows the cases where the proposed method failed to detect extremely small objects. Also, we found that the proposed method with MCG frequently predicted extremely large BBs (almost entire image) as shown in Figure 7. As shown in Figure 8, we observed that many of these failure cases were resolved by using Faster R-CNN instead of MCG.

Figure 9 shows the AP of the proposed method with MCG for each label as a function of the IoU threshold  $\tau$ . We observed that the performance hardly changed on some labels such as cut instr, throw, and hit instr because the proposed method rarely predicted these labels. In contrast, AP increased for other labels such as kick, cut obj, and hit obj when the threshold was set lower than 0.5. This is because localization errors on these labels were frequent while the proposed method correctly predicted the labels.

Figure 10 plots AP versus IoU threshold  $\tau$  on the V-COCO test set when we used Faster R-CNN. AP decreased for many types of labels as threshold  $\tau$  increased. From the results, we observed that using Faster R-CNN yielded many localization errors rather than label prediction errors, unlike MCG.

## 7 CONCLUSIONS

In this paper, we tackled weakly-supervised HOI detection, where only image-level supervision (i.e., verb labels) without BBs is used in training. We proposed to extend PCL, which is a weakly-supervised object detection method, to HOI detection. The proposed method obtains a feature vector that represents a person-object pair by adding the feature vector extracted for the person and the one for the object.

We conducted experiments on the V-COCO dataset. The results showed that when we employed MCG to generate the object candidate regions, the

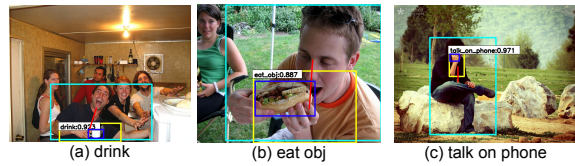


Figure 5: Examples of failure cases where BBs larger than ground truth objects are predicted.



Figure 6: Examples of failure cases where target persons or objects are extremely small.

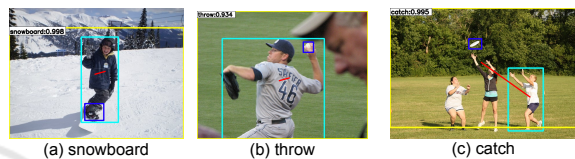


Figure 7: Examples of failure cases where predicted BBs are extremely large.

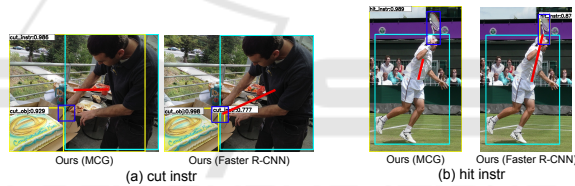


Figure 8: Successful cases achieved by replacing MCG with Faster R-CNN.

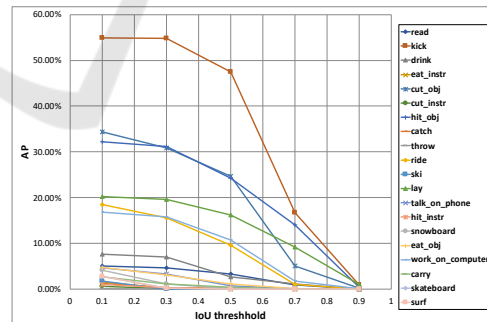


Figure 9: AP vs. IoU threshold on V-COCO test set (ours with MCG).

proposed method achieved 10.7% and 7.08% in terms of mean CorLoc and mAP, respectively. We also evaluated the performance when we used Faster R-CNN for object candidate generation. The mean CorLoc and mAP were significantly improved to 31.7% and 29.6%, respectively. Although BB supervision for object detection was used to train Faster R-CNN, this is a natural setting because datasets for object detection

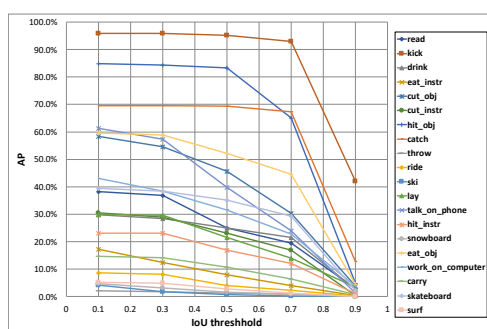


Figure 10: AP vs. IoU threshold on V-COCO test set (ours with Faster R-CNN).

are more commonly available than those for HOI detection. The proposed method with Faster R-CNN attained better performance on some labels than a state-of-the-art method for HOI detection based on supervised learning.

In future work, we will evaluate the performance when we use the person regions detected by a person detector as inputs to the proposed method because we assumed in this study that the person regions were perfectly detected. In order to improve performance, it is also our future work to extend another part of PCL such as pseudo-ground truth BB generation for HOI detection because we extended only the MIDN part in this paper.

## ACKNOWLEDGEMENTS

This study was supported by JSPS KAKENHI Grant Number JP17K06608 and JP20K12115.

## REFERENCES

- Arbeláez, P., Pont-Tuset, J., Barron, J. T., Marques, F., and Malik, J. (2014). Multiscale combinatorial grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 328–335.
- Bearman, A., Russakovsky, O., Ferrari, V., and Fei-Fei, L. (2016). What’s the point: Semantic segmentation with point supervision. In *Proceedings of the European Conference on Computer Vision*, pages 549–565.
- Bilen, H. and Vedaldi, A. (2016). Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854.
- Deselaers, T., Alexe, B., and Ferrari, V. (2012). Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100(3):275–293.
- Gao, C., Zou, Y., and Huang, J.-B. (2018). ican: Instance-centric attention network for human-object interaction detection. In *Proceedings of the British Machine Vision Conference*.
- 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*, pages 8359–8367.
- Gupta, S. and Malik, J. (2015). Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.
- 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 *Proceedings of European Conference on Computer Vision*, pages 740–755.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
- Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., and Yuille, A. (2018). PCL: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):176–191.
- 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*, pages 2843–2851.
- Ulutun, O., Iftekhar, A. S. M., and Manjunath, B. S. (2020). VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13617–13626.
- Wan, F., Wei, P., Jiao, J., Han, Z., and Ye, Q. (2018). Min-entropy latent model for weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1306.
- Yang, Z., Mahajan, D., Ghadiyaram, D., Nevatia, R., and Ramanathan, V. (2019). Activity driven weakly supervised object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2917–2926.