Detecting Human-Object Interaction with Mixed Supervision

Suresh Kirthi Kumaraswamy
Univ Le Mans, CNRS, IRISA
kirthifame@gmail.com

Miaojing Shi
King's College London
miaojing.shi@kcl.ac.uk

Ewa Kijak
Univ Rennes, Inria, CNRS, IRISA
ewa.kijak@irisa.fr

Abstract

Human object interaction (HOI) detection is an important task in image understanding and reasoning. It is in a form of HOI triplet $\langle human, verb, object\rangle$, requiring bounding boxes for human and object, and action between them for the task completion. In other words, this task requires strong supervision for training that is however hard to procure. A natural solution to overcome this is to pursue weakly-supervised learning, where we only know the presence of certain HOI triplets in images but their exact location is unknown. Most weakly-supervised learning methods do not make provision for leveraging data with strong supervision, when they are available; and indeed a naive combination of this two paradigms in HOI detection fails to make contributions to each other. In this regard we propose a mixed-supervised HOI detection pipeline: thanks to a specific design of momentum-independent learning that learns seamlessly across these two types of supervision. Moreover, in light of the annotation insufficiency in mixed supervision, we introduce an HOI element swapping technique to synthesize diverse and hard negatives across images and improve the robustness of the model. Our method is evaluated on the challenging HICO-DET dataset. It performs close to or even better than many fully-supervised methods by using a mixed amount of strong and weak annotations; furthermore, it outperforms representative state of the art weakly- and fully-supervised methods under the same supervision.

1. Introduction

The task of human-object interaction (HOI) detection is defined as a detection of a triplet $\langle human, verb, object\rangle$, identifying not only the bounding boxes of human and object but also their interaction [14, 49, 19, 36, 4, 57, 8, 10]. It is derived from visual relationship detection (VRD) of triplet $\langle object_1, predicate, object_2\rangle$ [31, 18, 16, 17, 37, 1, 55, 42, 56], but present different challenges: the predicates in VRD can be verbs (e.g. "push"), spatial (e.g. "on top of"), prepositions (e.g. "with"), comparative (e.g. " taller than"), while in HOI they are mainly verbs. On the other hand, *human-centric* interactions are more diverse and complicated, one person can easily interact with multiple objects in the meantime, e.g. "person wearing a jacket and riding a bicycle". This makes HOI a much more fine-grained task than VRD.

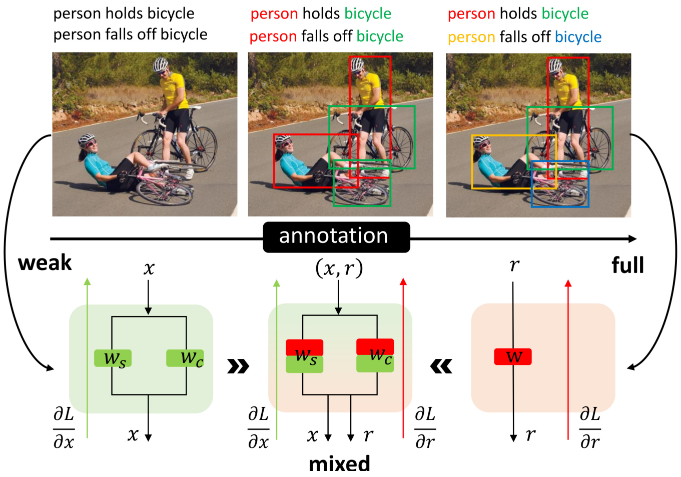

Figure 1: Human-object interaction detection with different levels of supervision. Top: annotation cost increases from image-level (x) labels (left) in weakly-supervised learning to region-level (r) bounding boxes (middle) and their correspondences (right) in fully-supervised learning. Bottom: our proposed mixed-supervised HOI detection pipeline (middle) enables joint learning of weakly- and fully-supervised HOI detection (left and right).

Intensive attention has been drawn to HOI alongside the introduction of new benchmarks, i.e. V-COCO [18], HICO-DET [5], featured with diverse and numerous human-object interactions. For instance, in HICO-DET, there exist 600 HOI classes and 80 common object classes in total. Despite that recent advances have reported significant improvement [19, 36, 4, 57], the annotation cost is exponentially increased in these datasets. Given N humans and M objects in an image, the maximum number of HOI is $N \times M$, where we have to scan each of them and provide instance-level annotations (e.g. bounding boxes) for the real ones (see Fig. 1: top-right). To alleviate this manual labor, we could provide only image-level HOI labels: a set of images are known to contain triplets of a certain HOI class, but the location and correspondence of objects are unknown in images (see Fig. 1: top-left); note the location of objects are assumed known sometimes (top-middle). Both cases can be conceptualized as weakly-supervised HOI detection.

There are few works that learn interactions from weak supervision. One representative is for VRD [54], which designs a weakly-supervised predicate prediction module inspired from the two-branch parallel structure in [3]. This can be easily adapted to HOI detection, as shown in Fig. 1: bottom-left. Nevertheless, in real application, instead of having only one type of labels, we can have a mixture of them: fully-labeled (instance-level), weakly-labeled (image-level), and even unlabeled.

A generalized HOI detection framework for mixed supervision thus becomes necessary. An intuitive solution would be merging the weakly-supervised and fully-supervised HOI detection as a multi-task job, which is nevertheless not straightforward: the different quality of annotations between weakly-labeled and fully-labeled data, as well as their imbalanced ratios should be considered. Fig. 4 illustrates an example: in the HICO-DET dataset, when adding different amounts of fully-labeled data (in red), results are either only slightly better than or even worse than learning with only weakly-labeled data (in grey). This simple combination, at best, does not exploit the full potential that could be derived from fully-supervised data, at worst, decreases the results obtained with weak supervision. This is the first challenge that needs to be solved in the mixed-supervised setting. Furthermore, HOI detection is a fine-grained task requiring the classification of similar interactions such as ”eating”, ”drinking”, ”blowing”. To be able to accurately distinguish them, diverse and hard negatives from similar interactions are essential for the network training. Nevertheless, due to the reason that many samples are weakly-labeled, interactions within them can not be clearly discriminated on the region-level; plus, some interaction classes are not even sufficiently collected. This poses another challenge for HOI detection.

Contributions. We for the first time propose a generalized framework for mixed-supervised HOI detection (MX-HOI):

- We integrate two state-of-the-art pipelines [19] and [54] for fully- and weakly-supervised HOI detection into a mixed-supervised pipeline.
- To tackle the multi-task optimization in the mixed pipeline, we introduce a momentum-independent learning strategy to tackle the adversarial effect between full and weak supervision, by separating their gradient history in momentum learning.
- To tackle the annotation insufficiency in the mixed supervision, we introduce an HOI element swapping strategy to specifically harvest hard negatives across images for the weakly-labeled data.

By conducting our generalized HOI detection framework on the most challenging HICO-DET dataset, we show our method enables HOI detection with a mixed amount of supervision, e.g. with 30% fully labeled data and 70% weakly-labeled data, we are able to retain 93.3% accuracy of the setting of 100% fully-labeled data. Furthermore, 1) our model improves both the state of the art weakly- and fully-supervised HOI detection methods [54, 19] under the same supervision; 2) unlabeled data can also be leveraged into our pipeline following a ”pseudo label” solution [23], where we can use the network trained on labeled data to infer labels of HOI pairs on unlabeled data.

2. Related Work

Visual relationships were originally used to help improve object localization [16, 22, 38], action recognition and pose estimation [9, 37] or semantic segmentation [15]. Relationships that are often modelled between objects include verbs, actions, spatial and prepositions [38, 52, 18, 16, 17, 37, 1, 20, 7]. [31] was the first work to formulate the detection of visual relationships as a separate task. They propose to leverage language priors from semantic word embedding to finetune the likelihood of a predicted relationship. Subsequently, many researchers improved and generalized this model [53, 27, 7]. Triplet metric learning is also adopted to optimize the visual feature connections among semantically related objects in [24, 42]. Attention [20, 55] and spatial locations [56] are some other additive cues to visual relationship detection.

Human-Object Interaction is a concept related to visual relationship. The interactions between humans and objects are mainly focused on verbs, and are much more fine grained (e.g. holding, hitting, throwing, touching) than relationships between generic objects. The study of HOI dates back to [8, 10, 17, 35, 49], when most works were tested on small datasets. This issue was addressed by Chao et al. [5] where they introduced a large dataset (HICO-DET) covering 80 common object categories and 600 HOI categories in total. Many recent works report their performance on this dataset and significant improvement has been achieved [36, 4, 19, 29, 12, 25, 46, 47, 26, 45]. For instance, Qi et al. [36] proposed a graph parsing neural network for HOI and was later extended by [21, 48, 57]; Gupta et al. [19] showed that a simple factorised model with appearance and layout encoding constructed from pretrained object detectors outperforms more sophisticated approaches; additional cues such as language features [12], parts based features [25, 46] are also exploited. Our MX-HOI is built on the recent advance of [19].

Weakly-supervised learning in visual recognition is mostly focused on object detection [3, 40, 44, 39, 43, 51]. One seminal work in weakly-supervised object detection (WSOD) is [3], where they designed a popular two-

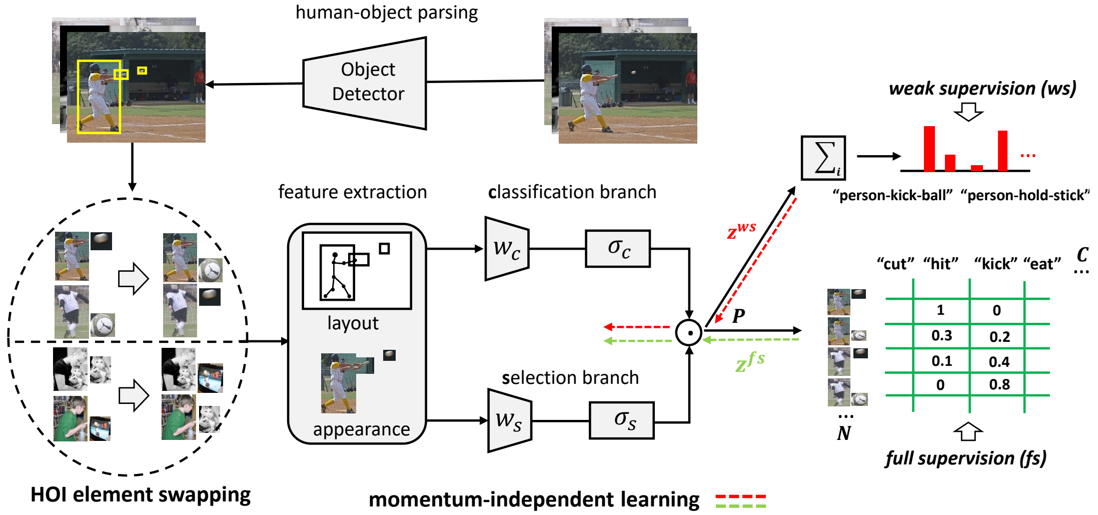

Figure 2: Illustration of our mixed-supervised HOI detection pipeline (MX-HOI). Human and object bounding boxes are obtained via an object detector. Human-object pairs are randomly created within an image and also across another image via the proposed HOI element swapping. HOI detection is realized via a two-branch structure in parallel for interaction classification and selection. Each branch consists of FC layer (w_c/w_s) for score prediction and softmax layer (σ_c/σ_s) for score normalization over rows or columns of the score matrix, respectively. The score matrices in the two branches are of size N (human-object pairs) and C (HOI classes) and are multiplied to produce the final matrix P. Training data with full and weak supervision (fs, ws) are optimized with region-level and image-level ground truth, respectively. We introduce a momentum-independent strategy to enable the mixed-supervised learning with two momentum z^{fs} and z^{ws}. Human-object pairs from two images are optimized in one batch.

branch parallel structure followed by [44, 43, 51]. Weakly-supervised relationship detection is more complex than WSOD as we need to detect individual objects for specific relations. Pretrained object detectors are normally assumed in this scenario [34, 54, 33]. Peyre et al. [34] proposed a weakly-supervised discriminative clustering model to learn relations with only image-level labels; later on, they developed another model for transfer by analogy to obtain visual phrases of never seen relations [33]. Zhang et al. [54] adopted the WSOD module in [3] to do weakly-supervised relationship detection and achieved very good results. In this work, we also adopt the WSOD module following [54] and adapt it to be part of our MX-HOI pipeline.

Mixed-supervised learning normally refers to learning from a mixture of strongly labeled data and weakly-labeled data. For instance, Cinbis et al. [6] considered mixed supervision in object detection where some images are annotated with bounding boxes while some are only with image-level labels. Papandreou et al. [32] studied the problem for semantic image segmentation from a combination of few strongly labeled (pixel-level annotations) and many weakly labeled (image-level labels or bounding boxes) images. Mixed-supervised learning can also be realized as leveraging an existing dataset of fully-labeled training images of non-target classes during the weakly-supervised learning of a new object category, which is connected to transfer learning, see e.g. [11, 41, 39, 50].

3. Mixed-supervised HOI detection

3.1. Preliminary

We build our MX-HOI framework on two state of the art HOI works with full supervision [19] and weak supervision [54], respectively. [19] introduces a no-frills model for HOI detection where they use appearance features from pretrained object detectors, spatial features through box layout, and encode human pose keypoints, as shown in Fig. 2: feature extraction. This is a no-frills detection without relying on attention or graph-based message passing [55, 36]. It uses a factorized multi-layer perceptrons (MLPs) and introduces several new training techniques to improve the MLPs: eliminating a train-inference mismatch, rejecting easy negatives using indicator terms, and training with large negative to positive ratios.

[54] adopts the weakly-supervised object detection pipeline [3] for weakly-supervised predicate prediction (WSPP): it is accomplished via the element-wise multiplication of the predicate selection and classification branch (see Fig. 2). The predicate score is softmax normalized over all candidate human-object box pairs with respect to a predicate class for the selection branch, and over all possible predicate classes with respect to one human-object pair for the classification branch, respectively. The predicate score in [54] is obtained from a position-role sensitive score maps with a pairwise ROI pooling. To integrate it into the no-

frills model above, we use the conventional ROI pooling. Predicate scores are predicted from the FC layers of the two branches, which is similar to the original structure in [3].

3.2. Overview

We introduce a mixed-supervised HOI detection framework (MX-HOI) as shown in Fig. 2: the input for MX-HOI is region proposals output from a pretrained object detector. We follow the same procedure as in [19] to extract both appearance and layout features for the human and object bounding boxes in a pair. Given human-object pairs, their region features are fed into the adapted two-branch predicate prediction structure from [54] (Sec. 3.1). The output of the two branches (matrices) are multiplied element-wise to produce one $N \times C$ matrix P over N human-object pairs (in a batch) and C interaction classes. Each element p_{ij} indicates the probability of the i^{th} human-object pair having j^{th} interaction type. For fully-labeled data, the predicate prediction is optimized on the region-level on matrix P, where a corresponding ground truth matrix is associated with each element being 1 or 0 indicating the true or false for the human-object interaction. For weakly-labeled data, the predicate optimization is on the image-level: P is accumulated over rows ($\sum_i p_{ij}$) to produce a C-dimensional vector where each element signifies the probability of the image containing certain HOI class. Similarly, an image-level ground truth vector with elements 1 or 0 is associated.

This is a multi-task optimization defined jointly with full and weak supervision. The learning is not straightforward: the optimization in the two learning manners is different as one focuses on the image-level and the other on the region-level (x and r in Fig. 1); the error surface toggles between the gradients from weak supervision and full supervision across different batches in the network. We therefore propose a momentum-independent learning strategy (Sec. 3.3). Besides, for the weakly-labeled data in mixed-supervised learning, we introduce an HOI element swapping strategy (Sec. 3.4) to further augment the hard negatives. Loss function is given in the end (Sec. 3.5).

3.3. Momentum-independent Learning

In the context of mixed-supervised learning, network weights are updated by either weak or full supervision, depending on the samples within the mini-batch. Most recent gradient descent based optimizers use momentum-based weights update. Let w_t and $\nabla f(w_t)$ be respectively the weight and the gradient at iteration t, and α be the step size, the momentum-based update rule is given by:

$$w_t = w_{t-1} - z_t; \;\; z_t = \beta z_{t-1} + \alpha \nabla f(w_{t-1}) \tag{1}$$

where β is the momentum parameter (usually $\beta \geq 0.9$) and z_t is the momentum, which is dependent on all the previous gradient values.

Using momentum-based gradient descent can however be a problem in the mixed-supervised learning. In the fully-supervised case, the ground truth is directly given on the instance level such that the gradient of the loss function will accordingly backpropagate to the specific regions of the human and object in a pair. In the weakly supervised case, the ground truth is instead only given on the image-level, and predictions on all possible human-object pairs are aggregated together to the image-level for loss computation; at the backpropagation time, the gradient is distributed among all the human-object pairs. As a result, the gradients for full and weak supervision are computed on different error surfaces and are not compatible. Using one momentum to record both will make the network weight optimization toggles between the two sources of gradients across mini-batches. This indeed leads to an adversarial effect of the mixed-supervised learning (see the ablation study in Sec. 4.2).

To mitigate this, we propose to bootstrap the mixed-supervised learning with two independent momentum z_t^{ws} and z_t^{fs} to record the gradient history of weak and full supervision separately. z_t^{ws} will be used and updated only with weakly-labeled samples in the mini-batch, while z_t^{fs} will be instead used for the fully-labeled samples. w_t however is remained to be shared in the network such that the weakly- and fully-supervised pipeline are jointly optimized:

$$\begin{aligned} w_t = w_{t-1} - z_t^{ws}; \quad z_t^{ws} = \beta z_{t-1}^{ws} + \alpha \nabla f(w_{t-1}) \\ w_t = w_{t-1} - z_t^{fs}; \quad z_t^{fs} = \beta z_{t-1}^{fs} + \alpha \nabla f(w_{t-1}) \end{aligned} \tag{2}$$

3.4. HOI element swapping

HOI detection is a fine-grained task. To accurately classify similar interactions, diverse and hard negatives are needed. In the fully-supervised case, where region-level ground truth are available, this can be achieved via choosing the false positive class of large confidence score or false positive region of large intersection-over-union (IoU) with ground truth. While in the weakly-supervised case, where only image-level ground truths are available, conventional manners of finding negatives no longer apply. We instead introduce an HOI element swapping way to collect diverse and hard negatives across images.

Suppose that two images im_1 and im_2 contain one human h and one object o inside each, respectively. The standard way to create the candidate human-object pairs is to group (h_1, o_1) and (h_2, o_2) within each image (see Fig. 3). To further augment negatives, a simple way is to lower the threshold from the RPN to produce more proposals; but this is inefficient as many proposals with low confidence scores are not good detection of humans or objects, and we do not have ground truth bounding boxes to distinguish them in the weakly-supervised setting. Hence, we propose to keep a fair detection of humans and objects within each

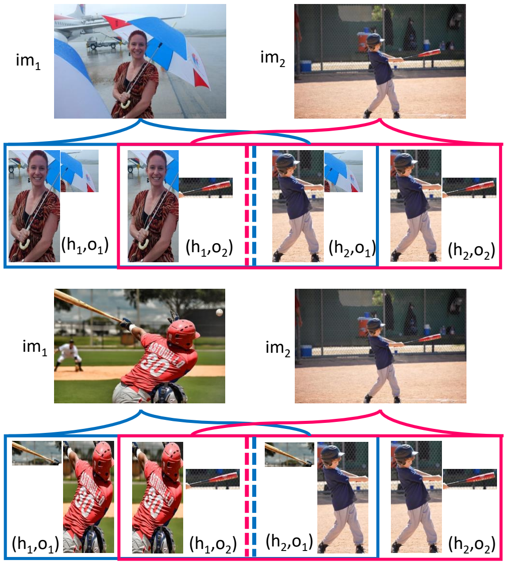

Figure 3: Illustration of HOI element swapping. Top: Object classes from the two images are different. Bottom: same object class in the two images. Swapped pairs (h_i, o_j) are negatives due to wrong object class (top) or wrong spatial layout (bottom).

image and augment the human-object pairs across images by swapping their HOI elements: given (h_1, o_1) from im_1 and (h_2, o_2) from im_2, we mix the human proposal from the im_1 with the object proposal from im_2 and vice versa: (h_1, o_2) and (h_2, o_1); this gives us two more mixed human-object pairs. One image may contain more than one human or object. Considering there are H_1 humans and O_1 objects in im_1, H_2 humans and O_2 objects in im_2, selecting all humans and objects from the two images will produce $(H_1 + H_2) \times (O_1 + O_2)$ pairs in total for the two images. which is far too much. In practice, we remove those easy negatives with low confidence scores such that the number of human-object pairs is kept the same to that of the original number in two images, i.e. $H_1 \times O_1 + H_2 \times O2$.

By doing this, we can obtain more diverse combinations of HOI pairs, where many swapped HOI pairs coming from two images might look like positive HOI pairs playing the role of hard negatives. For instance, in Fig. 3: bottom, the object class from two images is the same in particular, yet the swapped human-object pairs (h_1, o_2) and (h_2, o_1) are still negatives due to the wrong spatial layout.

The augmented human-object pairs as shown in Fig. 3 can be hard negatives for both images. To efficiently optimize the learning on two images, we propose to aggregate all the human-object pairs from two images to form one image-level HOI label vector, where the corresponding ground truth is the HOI labels from both images. Apart from efficiency, another benefit of doing this, compared to optimizing the image separately, is that the positive human-object pairs from one image could also serve as negatives for the other image if they are of different HOI labels or as positives if they are of the same HOI label.

3.5. Loss function

Loss function is defined within each mini-batch depending on the supervision $\mathbb{S}$ of the input samples, which can be full supervision ($\mathbb{S} = \mathbb{FS}$) or weak supervision ($\mathbb{S} = \mathbb{WS}$):

$$\mathcal{L}_{\text{mini-batch}} = \sum_{j=1}^{C}\Big(1(\mathbb{S}=\mathbb{FS})\frac{1}{N}\sum_{i=1}^{N} BCE(y_{ij}, p_{ij}) + 1(\mathbb{S}=\mathbb{WS}) BCE(y_j, p_j)\Big) \tag{3}$$

BCE is the binary cross-entropy; p_{ij} is the probability of j^{th} HOI class for the i^{th} human-object pair, where there are N pairs and C classes in total; p_j is the probability of the j^{th} HOI class for the given images in the mini-batch. The former is defined for the fully-labeled data with region-level ground truth, while the latter is defined for the weakly-labeled data with image-level ground truth only. In practice, we feed the features of human-object pairs from two images in each mini-batch. Referring to Sec. 3.2, the ground truth for fully-labeled data is given in a form of a matrix, where each element y_{ij} indicates whether the i^{th} human-object pair with j^{th} HOI class is true or false. $y_{ij} = 1$ if the human and object boxes in the i^{th} pair have an IoU greater than 0.5 with a ground truth box-pair of the j^{th} HOI class. The ground truth for the weakly-labeled data is given in a form of C-dimensional vector where its element $y_j = 1$ if the j^{th} HOI class occurs in any of the two images.

4. Experiments

4.1. Experimental setup

Datasets. Ever since its introduction in [4], the HICO-DET has become the defacto standard dataset for human object interaction detection. The dataset has a total of 47,776 images: 38,118 (80%) are used for training and 9,658 (20%) for testing. Each image is provided with the $\langle human, object, predicate\rangle$ triplets which include human and object bounding boxes and HOI classes. It covers 80 object categories and 117 interactions, which result into 600 HOI classes in total. These classes are subdivided into 138 *rare* ones, whose training samples are less than 10 images; 462 *non-rare* ones, whose training samples are more than 10 images. On average, 1.67 HOI triplets are annotated in each image. HICO-DET is a much bigger dataset compared to the previous V-COCO dataset [18]. In line with recent works on HOI detection [19, 2], we evaluate our method on the large-scale HICO-DET to offer comprehensive study and in-depth analysis on it.

Data splitting. The training images are randomly split with different ratios of weakly- and fully-labeled data (denoted as WS and FS). The default WS/FS ratio is set to 70/30 and 30/70 where 70% (30%) data from the training set are weakly-labeled and the rest are fully-labeled. We also evaluate different WS/FS ratios ranging between 100/0 and 0/100 in the experiments. Some more settings regarding unlabeled data and class-split are also presented in the end.

Implementation details and evaluation protocol. Following [4], the human and object detection results are taken from the top scoring output of a Faster-RCNN pretrained on MS-COCO [28]. Each human is paired with all the objects within an image. Faster-RCNN produces numerous candidate bounding boxes. For each object, we filter the 30-top performing boxes depending on the detection scores. For fully-labeled data, ground truth HOI triplets are provided with human/object bounding boxes and their interaction. For weakly-labeled data, only image-level HOI labels are provided meaning that the real correspondence from a human detection to an object is not given in the image. For unsupervised data, no HOI labels are provided. The network is trained with a mini-batch containing the set of region proposal pairs in two images, which are randomly selected from either the fully-labelled set or the weakly-labeled set. This is done once before the training for efficiency. Two images from the weakly-labeled set are applied with element swapping. The learning rate is 1e-3 and 1e-4 for weakly- and fully-labeled data, respectively. We train 40,000 iterations in total.

Evaluation of HOI detection employs the widely used mean average precision (mAP) metric where a prediction is considered correct only if its HOI class label is correct, and its human and object bounding boxes have an Intersection over Union (IoU) larger than 0.5 with their respective ground truth bounding boxes.

4.2. Ablation study

In this section, we first justfiy the importance of our proposed new elements MIL and HES in order to enable a meaningful mix-supervised HOI detection. Next, we give the result of MX-HOI generalizing over different WS/FS ratios from 0/100 to 100/0.

Using weak and strong annotations. To start with our ablation study, we first train our HOI detector with weak annotations only; next, we train the detector with both weak and strong annotations. We illustrate the mAP on the test set in Fig. 4: by default 70% (30%) data are chosen as weakly-labeled and the rest are as fully-labeled. The results show that without using our proposed MIL (w/o MIL), adding FS data can perform even worse than using WS data only: for example using 70% of the training data, all weakly-labeled (WS/FS = 70/0), it (*weak only*, grey) yields a mAP of 14.68; when adding fully-labeled data (WS/FS = 70/30), the detector performs an even worse mAP 13.17 (red). This illustrates well the adversarial effect between the weakly- and fully-labeled data. In order to tackle the problem, we first tried a sequence training (ST) strategy [40, 30], where all fully-labeled data (resp. all weakly-labeled data) are presented in the network with some epochs before the weakly-labeled data (resp. fully-labeled data) are added in. We denote the strategy as w/ ST-F when the fully-labeled data are trained first, or w/ ST-W when the weakly-labeled data are trained first. The results are improved in this manner but not too much (see Fig 4). Next, we introduce our momentum-independent learning strategy to specifically enable the mixed-supervised HOI detection.

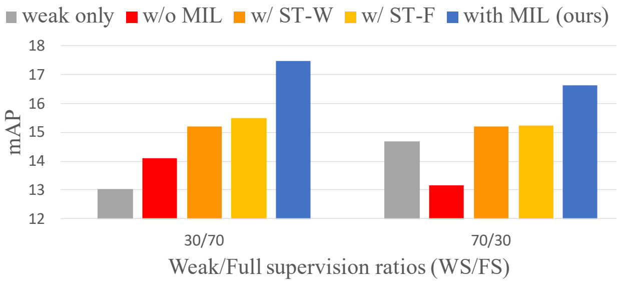

Figure 4: Using weak, strong and mixed annotations for HOI detection. MIL: momentum-independent learning. ST-W: sequence training with weakly-labeled data first; ST-F: sequence training with fully-labeled data first;

Method	WS/FS	Rare	Non-Rare	WS/FS	Rare	Non-Rare
weak only	70/0	11.84	15.72	30/0	8.68	14.11
w/o MIL	70/30	8.88	14.45	30/70	9.17	15.74
w/ ST-F	70/30	10.41	16.69	30/70	10.52	17.00
w/ ST-W	70/30	10.17	16.71	30/70	11.03	16.43
with MIL (ours)	70/30	**12.36**	**17.91**	30/70	**12.79**	**18.80**

Table 1: Ablation of momentum-independent learning (MIL) in MX-HOI on HICO-DET dataset. mAP is reported.

Momentum-independent learning (MIL). To tackle the inconsistency of gradients in the network backpropagation, we introduce two independent momentum to record the gradient history of full and weak supervision separately as they are computed on different error surfaces and are functionally different in the network. This is conceptualized as momentum-independent learning and is a key element of our MX-HOI pipeline. Having a look at Fig. 4, ours (w/ MIL) significantly increases the mAP to e.g. 16.63 for WS/FS = 70/30 and 17.47 for WS/FS = 30/70. This demonstrates the importance of our proposed MIL to enable a meaningful mixed-supervised HOI detection. Some more detailed comparisons between our MIL and other variants on the rare and non-rare classes are shown in Table 1.

HOI element swapping (HES). Referring to Sec. 3.4, HOI

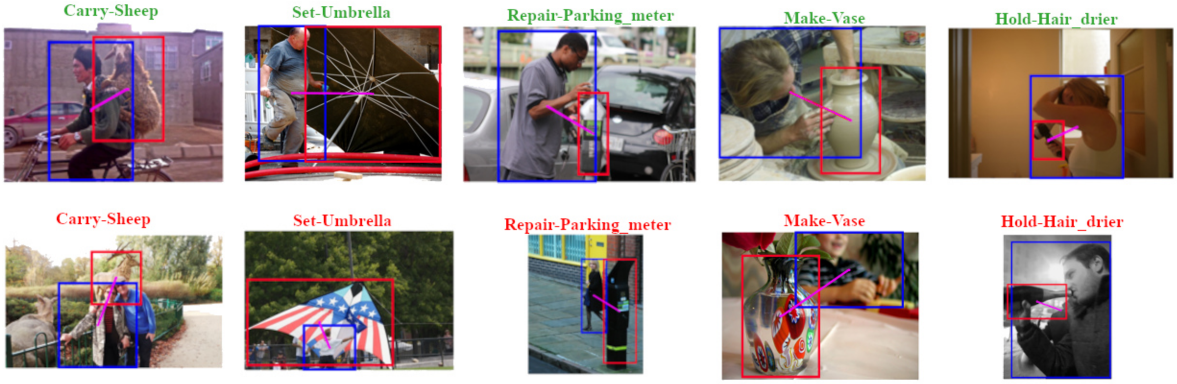

Figure 5: Examples of correct detections (top) and incorrect detections (bottom) with MX-HOI. The classes shown here are Carry-Sheep, Set-Umbrella, Repair-Parking-meter, Make-Vase and Hold-Hairdryer.

Method	WS/FS	Full	Rare	Non-Rare
ours (w/o HES)	100/0	15.14	10.65	16.48
ours	100/0	**16.14**	**12.06**	**17.50**
ours (w/o HES)	70/30	15.82	10.39	17.41
ours	70/30	**16.63**	**12.36**	**17.91**
ours (w/o HES)	30/70	16.73	12.00	18.14
ours	30/70	**17.47**	**12.79**	**18.80**

Table 2: Ablation of HOI element swapping (HES) in MX-HOI on HICO-DET dataset. mAP is reported.

element swapping is introduced for hard negative harvest on weakly-labeled data. To verify its effectiveness, we ablate it in Table 2 by comparing with MX-HOI without HES. We vary WS/FS from 100/0 to 30/70 and show that on different mixed levels, HES always helps the HOI detection. For instance, when WS/FS = 70/30, MX-HOI yields +0.81% improvement over MX-HOI (w/o HES).

Additionally, we also apply HES on fully-labeled data, i.e. WS/FS = 0/100, and obtain the mAP 17.04, 13.35, 18.11 on full, rare and non-rare classes, respectively, which actually harms the performance on non-rare classes while helps a bit on rare classes comparing to 17.82, 12.91, and 19.17 in Table 3. Rare classes do not have adequate training samples, HES can help provide hard negatives; while for non-rare classes, hard negatives can be directly mined via the given bounding box ground truth. Overall, we did not find HES to be effective for fully-labeled data in general.

WS/FS variations. We offer the results of MX-HOI with different WS/FS: 100/0, 80/20, 70/30, 50/50, 30/70, 20/80 and 0/100 in Table 3. One can see that the performance increases with an increase of FS for both rare and non-rare classes, as more fully-labeled data are added into the training. The overall full mAP increases from 16.14 to 17.82. In Table 3, we also show the result of fixing either WS, or FS to 30% while varying the other: the performance increases along with the training (sub-)set size; but the improvement margin is bigger when fixing WS and increasing FS compared to fixing FS and increasing WS. All these results make perfect sense for MX-HOI: adding more labeled data, regardless WS or FS, increases its performance; FS data in general provides more help than WS data. Examples of MX-HOI with WS/FS=70/30 are given in Fig. 5.

Supervision (WS/FS)	Full	Rare	Non-Rare
100/0	16.14	12.06	17.50
80/20	16.49	12.28	17.81
70/30	16.63	12.36	17.91
50/50	17.08	12.58	18.17
30/70	17.47	12.79	18.80
20/80	17.60	12.85	18.95
0/100	17.82	12.91	19.17
30/30	16.05	11.64	17.37
30/50	16.84	11.81	18.47
30/70	17.47	12.79	18.80
50/30	16.34	12.04	17.69
70/30	16.63	12.36	17.91

Table 3: Different ratios of WS/FS in MX-HOI on HICO-DET dataset (top). Fixing WS (resp. FS) ratio and varying the other (bottom). mAP is reported.

4.3. Comparison to state of the art

In Table 4, we first compare our method to its lower and upper bounds denoted respectively by WS-No-Frills and No-Frills in the following. WS-No-Frills is an adaption of a representative weakly-supervised relationship detection module [54] onto the SOTA HOI detection pipeline [19], which produces mAP 15.14, 10.65, and 16.48 on full, rare and non-rare classes. Our MX-HOI under the same setting WS/FS=100/0 improves the result to 16.14, 12.06, 17.50 due to the adoption of HOI element swapping. In fully-supervised setting (WS/FS = 0/100), our MX-HOI also improves the No-Frills [19] by +0.6%, this is attributed to our

Methods	WS/FS	Full	Rare	Non-rare
WS-No-Frills	100/0	15.14	10.65	16.48
MX-HOI		**16.14**	**12.06**	**17.50**
MX-HOI	70/30	16.63	12.36	17.91
MX-HOI	30/70	17.47	12.79	18.80
MX-HOI	0/100	17.82	12.91	19.17
No-Frills [19]		17.18	12.17	18.68
VSGNet [45]		**19.80**	**16.05**	**20.91**
PMFNet [46]		17.46	15.65	18.00
TIN [26]		17.22	13.51	18.32
GCN-HOI [48]		14.70	13.26	15.13
GPNN [36]		13.11	9.34	14.23
ICAN [13]		12.80	8.53	14.07

Table 4: Comparison with the state-of-the-art methods on HICO-DET test set (mAP).

two-branch softmax [3] in No-Frills (This softmax is applied in subbranches before the final classification head). Despite MX-HOI is introduced for mix-supervised HOI detection, it also improves the SOTA bounds as side benefits.

In the mixed-supervised setting, MX-HOI retains 93.3 % accuracy of the SOTA No-Frills by using a mixture of 30% fully-labeled data and 70% weakly-labeled data. To compare with it, we implement a naive *multi-stage training pipeline*: it first trains the model on 30% fully-labeled data, then infers the HOI class probabilities on the human-object pairs in the rest 70% weakly-labeled images; the HOI triplet with the largest probability is selected as pseudo ground truth for each given HOI label on the image-level in weakly-labeled data. These selected HOI triplets are mixed with existing fully-labeled data to train the network again. The network remains a fully-supervised pipeline in this manner. This process would repeat several rounds until the convergence of the model. We obtain mAP 15.23, 10.63, and 16.61 under the setting of WS/FS = 70/30, which is much lower than our MX-HOI (16.63, 12.36 and 17.91).

We also compare MX-HOI with other recent arts [48, 13, 36, 45, 46, 26] using 100% supervision. One can see that with WS/FS = 70/30, MX-HOI performs very close to the SOTA.

4.4. More settings

Unlabeled data can also be added into the whole framework: we first obtain all possible human-object pairs in an unlabeled image (US) from the detection result. Given the trained model of the mix-supervised pipeline, we can estimate the marginal HOI class probability for every human-object pair in the unlabeled image. If the probability is larger than a threshold (e.g. 0.5), we take the predicted HOI class as the pseudo ground truth for this human-object pair and add it into the network training in the next cycle. The loss function in (3) now includes another term for the unlabeled data, which is formulated similarly to the fully-labeled data with pseudo ground truth. The loss weights among the three terms remain 1. This process iterates for several cycles until the convergence of the network.

	WS/FS/US	Full	Rare	Non-Rare
MX-HOI	30/40/0	16.08	12.05	17.29
MX-HOI	30/40/30	16.53	11.63	17.79

Table 5: Adding unlabeled data (US) into MX-HOI.

Table 5 shows the result of WS/FS/US being 30/40/30, where 30%, 40% and 30% percent data are respectively weakly-, fully- and un-supervised (see Sec. 4.1). Results using unlabeled data improves performance when compared to using only WS/FS with 30/40 ratio.

Class-split: Instead of randomly splitting the dataset images for weak and full supervision, we can randomly split the whole HOI classes into 50% vs. 50%. Images from the first 50% classes are trained with full supervision, the second 50% with weak supervision. If we train two models separately on the two sets, we got mAP 13.3 and 11.6 on the test set of each own part of classes, respectively. If we train one model over the two sets jointly using MX-HOI, the mAP increases to 14.8 and 13.10. Despite the two sets are of different classes, training them together with more data benefit the performance of both in our pipeline.

5. Conclusion

We present a mixed-supervised HOI detection framework (MX-HOI) which employs two state-of-the-art fully- and weakly-supervised pipelines. Within this framework, we first introduce a momentum-independent strategy to tackle the adversarial effect of full and weak supervision by separating their gradient history in momentum learning. Second, we introduce an HOI element swapping strategy to harvest hard negatives across images for weakly-labeled data. Unlabeled data can also be leveraged using a “pseudo label” solution where class labels on HOI pairs are provided by the trained mixed-supervised pipeline. Extensive experiments on the large-scale HICO-DET dataset show that, with only 30% fully-labeled data and 70% weakly-labeled data, our MX-HOI is able to retain 93.3% accuracy of the setting of 100% fully-labeled data. Future work will be focused on developing a stronger weakly-supervised HOI detection pipeline to integrate it into our MX-HOI framework.

Acknowledgements: This work was partially supported by the READ-IT project, funded by the JPI Cultural Heritage under the European Union Horizon 2020 R&I program (grant agreement No. 699523). Miaojing Shi was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61828602.

References

[1] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016.

[2] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. In *AAAI*, 2020.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[5] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):189–203, 2016.

[7] Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. Context-dependent diffusion network for visual relationship detection. In *ACM MM*, 2018.

[8] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In *NeurIPS*, 2011.

[9] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012.

[10] Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshops*, 2010.

[11] Thomas Deselaers, Alexe Bogdan, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100:275–293, 2012.

[12] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020.

[13] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018.

[14] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[15] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *International Journal of Computer Vision*, 80(3):300–316, 2008.

[16] Abhinav Gupta and Larry S Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008.

[17] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

[18] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[19] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019.

[20] Chaojun Han, Fumin Shen, Li Liu, Yang Yang, and Heng Tao Shen. Visual spatial attention network for relationship detection. In *ACM MM*, 2018.

[21] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *ICCV*, 2019.

[22] M Pawan Kumar and Daphne Koller. Efficiently selecting regions for scene understanding. In *CVPR*, 2010.

[23] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, 2013.

[24] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, 2017.

[25] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020.

[26] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019.

[27] Xiaodan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 2017.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[29] Y Liu, Q Chen, and A Zisserman. Amplifying key cues for human-object-interaction detection. *ECCV*, 2020.

[30] Yuting Liu, Miaojing Shi, Qijun Zhao, and Xiaofang Wang. Point in, box out: Beyond counting persons in crowds. In *CVPR*, 2019.

[31] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.

[32] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[33] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019.

[34] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *ICCV*, 2017.

[35] Alessandro Prest, Cordelia Schmid, and Vittorio Ferrari. Weakly supervised learning of interactions between humans and objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):601–614, 2011.

[36] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[37] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Charles Rosenberg, and Li Fei-Fei. Learning semantic relationships for better action retrieval in images. In *CVPR*, 2015.

[38] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *CVPR*, 2011.

[39] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *ICCV*, 2017.

[40] Miaojing Shi and Vittorio Ferrari. Weakly supervised object localization using size estimates. In *ECCV*, 2016.

[41] Zhiyuan Shi, Parthipan Siva Siva, and Tao Xiang. Transfer learning by ranking for weakly supervised object annotation. In *BMVC*, 2015.

[42] Xu Sun, Yuan Zi, Tongwei Ren, Jinhui Tang, and Gangshan Wu. Hierarchical visual relationship detection. In *ACM MM*, 2019.

[43] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(1):176–191, 2018.

[44] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *CVPR*, 2017.

[45] Oytun Ulutan, ASM Iftekhar, and Bangalore S Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020.

[46] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019.

[47] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020.

[48] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019.

[49] Wanru Xu, Jian Yu, Zhenjiang Miao, Lili Wan, and Qiang Ji. Prediction-cgan: Human action prediction with conditional generative adversarial networks. In *ACM MM*, 2019.

[50] Yukuan Yang, Fangyu Wei, Miaojing Shi, and Guoqi Li. Restoring negative information in few-shot object detection. In *NeurIPS*, 2020.

[51] Zhaohui Yang, Miaojing Shi, Yannis Avrithis, Chao Xu, and Vittorio Ferrari. Training object detectors from few weakly-labeled and many unlabeled images. *arXiv preprint arXiv:1912.00384*, 2019.

[52] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 2016.

[53] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *ICCV*, 2017.

[54] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. PPR-FCN:: Weakly supervised visual relation detection via parallel pairwise R-FCN. In *ICCV*, 2017.

[55] Sipeng Zheng, Shizhe Chen, and Qin Jin. Visual relation detection with multi-level attention. In *ACM MM*, 2019.

[56] Hao Zhou, Chongyang Zhang, and Chuanping Hu. Visual relationship detection with relative location mining. In *ACM MM*, 2019.

[57] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019.