Learning to Detect Human-Object Interactions

Yu-Wei Chao[1], Yunfan Liu[1], Xieyang Liu[1], Huayi Zeng[2], and Jia Deng[1]

[1]University of Michigan, Ann Arbor
{ywchao,yunfan,lxieyang,jiadeng}@umich.edu

[2]Washington University in St. Louis*
{zengh}@wustl.edu

Abstract

We study the problem of detecting human-object interactions (HOI) in static images, defined as predicting a human and an object bounding box with an interaction class label that connects them. HOI detection is a fundamental problem in computer vision as it provides semantic information about the interactions among the detected objects. We introduce HICO-DET, a new large benchmark for HOI detection, by augmenting the current HICO classification benchmark with instance annotations. To solve the task, we propose Human-Object Region-based Convolutional Neural Networks (HO-RCNN). At the core of our HO-RCNN is the Interaction Pattern, a novel DNN input that characterizes the spatial relations between two bounding boxes. Experiments on HICO-DET demonstrate that our HO-RCNN, by exploiting human-object spatial relations through Interaction Patterns, significantly improves the performance of HOI detection over baseline approaches.

1. Introduction

Visual recognition of human-object interactions (HOI) (e.g. "riding a horse", "eating a sandwich") is a fundamental problem in computer vision. Successful HOI recognition could identify not only objects but also the relationships between them, providing a deeper understanding of the semantics of visual scenes than just object recognition [19, 32, 12] or object detection [8, 29, 23, 3]. Without HOI recognition, an image can only be interpreted as a collection of object bounding boxes. An AI system can only pick up information such as "A baseball bat is in the right corner" and "A boy is close to the baseball bat", but not "A boy wearing a cap is swinging a baseball bat".

HOI recognition has recently attracted increasing attention in the field of computer vision [10, 34, 33, 6, 25, 4, 5, 28, 13]. While significant progress has been made, the problem of HOI recognition is still far from being solved. A key issue is that these approaches have been evaluated using small datasets with limited HOI categories, e.g. 10 categories in PASCAL VOC [7] and 40 categories in Stanford 40 Actions [35]. Furthermore, these datasets offer only a limited variety of interaction classes for each object category. For example, in Stanford 40 Actions, "repairing a car" is the only HOI category involving the object "car". It is unclear whether a successful algorithm can really recognize the interaction (e.g. "repairing"), or whether it simply recognizes the present object (e.g. "car"). This issue has recently been addressed by [1], which introduced "Humans interacting with Common Objects" (HICO), a large image dataset containing 600 HOI categories over 80 common object categories and featuring a diverse set of interactions for each object category. HICO was used in [1] to provide the first benchmark for image-level HOI classification, i.e. classifying whether an HOI class is present in an image.

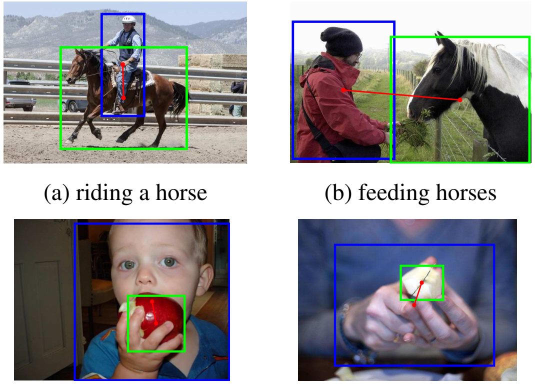

(a) riding a horse
(b) feeding horses
(c) eating an apple
(d) cutting an apple

Figure 1: Detecting human-object interactions. Blue boxes mark the humans. Green boxes mark the objects. Each red line links a person and an object involved in the labeled HOI class.

While the introduction of HICO may facilitate progress in the study of HOI classification, HOI recognition still cannot be fully addressed, since with only HOI classification computers are not able to accurately localize the present interactions in images. To be able to ground HOIs to image regions, we propose studying a new problem: detecting human-object interactions in static images. The goal of HOI

*Work done at the University of Michigan as a visiting student.

detection is not only to determine the presence of HOIs, but also to estimate their locations. Formally, we define the problem of HOI detection as predicting a pair of bounding boxes—first for a person and second for an object—and identifying the interaction class, as illustrated in Fig. 1. This is different from conventional object detection, where the output is only a single bounding box with a class label. Addressing HOI detection will bridge the gap between HOI classification and object detection by identifying the interaction relations between detected objects.

The contributions of this paper are two-fold: (1) We introduce HICO-DET, the first large benchmark for HOI detection, by augmenting the current HICO classification benchmark with instance annotations. HICO-DET offers more than 150K annotated instances of human-object pairs, spanning the 600 HOI categories in HICO, i.e. an average of 250 instances per HOI category. (2) We propose Human-Object Region-based Convolutional Neural Networks (HO-RCNN), a DNN-based framework that extends state-of-the-art region-based object detectors [9, 8, 29] from detecting a single bounding box to a pair of bounding boxes. At the core of our HO-RCNN is the Interaction Pattern, a novel DNN input that characterizes the spatial relations between two bounding boxes. Experiments on HICO-DET demonstrate that our HO-RCNN, by exploiting human-object spatial relations through Interaction Patterns, significantly improves the performance of HOI detection over baseline approaches. The dataset and code are publicly available at `http://www.umich.edu/~ywchao/hico/`.

2. Related Work

HOI Recognition A surge of work on HOI recognition has been published since 2009. Results produced in these works were evaluated on either action classification [34, 33, 6, 25, 4, 5, 28, 13], object detection [34], or human pose estimation [34, 5]; none of them were directly evaluated on HOI detection. Chao *et al.* [1] recently contributed a large image dataset "HICO" for HOI classification [1, 26]. However, HICO does not provide ground-truth annotations for evaluating HOI detection, which motivates us to construct a new benchmark by augmenting HICO. We also highlight a few other recent datasets. Gupta and Malik [11] augmented MS-COCO [22] by connecting interacting people and objects and labeling their semantic roles. Yatskar *et al.* [36] contributed an image dataset for situation recognition, defined as identifying the activity together with the participating objects and their roles. Both datasets, unlike HICO, do not offer a diverse set of interaction classes for each object category. Lu *et al.* [24] and Krishna *et al.* [18] separately introduced two image datasets for detecting object relationships. While they feature a diverse set of relationships, the relationships are not exhaustively labeled in each image. As a result, follow-up works [2, 21, 20, 37, 38] which benchmark on these datasets can only evaluate their detection result with recall, but not precision. In contrast, we exhaustively labeled all the instances for each positive HOI label in each image, enabling us to evaluate our result with mean Average Precision (mAP).

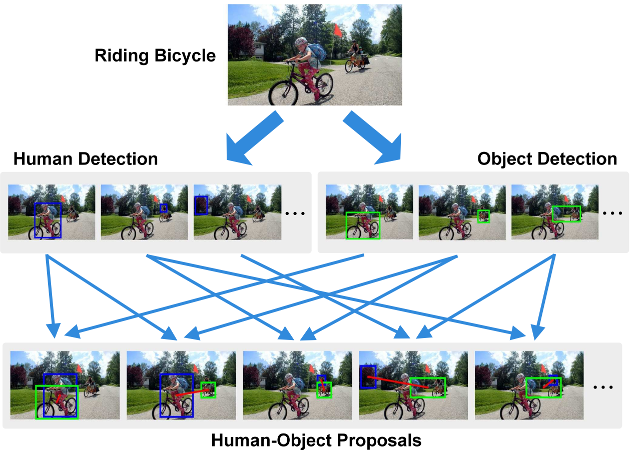


Figure 2: Generating human-object proposals from human and object detections.

Object Detection Standard object detectors [8, 29, 23, 3] only produce a class-specific bounding box around each object instance; they do not label the interaction among objects. Sadeghi and Farhadi [31] proposed "visual phrases" by treating each pair of interacting objects as a unit and leveraged object detectors to localize them. HOI detection further extends the detection of "visual phrases" to localize individual objects in each pair. Our proposed HO-RCNN, built on recent advances in object detection, extends region-based object detectors [9, 8, 29] from taking single bounding boxes to taking bounding box pairs.

Grounding Text Descriptions to Images HOI detection grounds the semantics of subjects, objects, and interactions to image regions, which is relevant to recent work on grounding text descriptions to images. Given an image and its caption, Kong *et al.* [17] and Plummer *et al.* [27] focus on localizing the mentioned entities (e.g. nouns and pronouns) in the image. HOI detection, besides grounding entities, i.e. people and objects, also grounds interactions to image regions. Karpathy and Fei-Fei [16] and Johnson *et al.* [15] address region-based captioning, which can be used to generate HOI descriptions in image regions. However, they are unable to localize individual persons and objects involved in the HOIs.

3. HO-RCNN

Our HO-RCNN detects HOIs in two steps. First, we generate proposals of human-object region pairs using state-of-the-art human and object detectors. Second, each human-

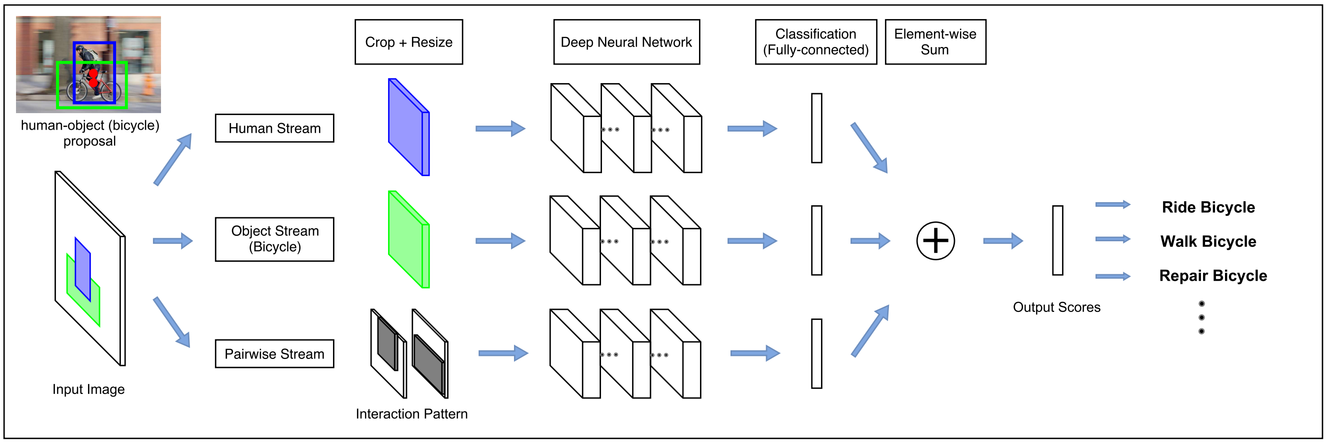


Figure 3: Multi-stream architecture of our HO-RCNN.

object proposal is passed into a ConvNet to generate HOI classification scores. Our network adopts a multi-stream architecture to extract features on the detected humans, objects, and human-object spatial relations.

Human-Object Proposals We first generate proposals of human-object region pairs. One naive way to exploit a pool of class-agnostic bounding boxes like other region-based object detection approaches [9, 8, 29]. However, since each proposal is a pairing between a human and object bounding box, the number of proposals will be quadratic in the number of the candidate bounding boxes. To ensure high recall, one usually needs to keep hundreds to thousands of candidate bounding boxes, which results in more than tens of thousands of human-object proposals. Classifying HOIs for all proposals will be intractable. Instead, we assume having a list of HOI categories of interest (e.g. “riding a horse”, “eating an apple”) beforehand, so we can first detect bounding boxes for humans and the object categories of interest (e.g. “horse”, “apple”) using state-of-the-art object detectors. We keep the bounding boxes with top detection scores. For each HOI category (e.g. “riding a horse”), the proposals are then generated by pairing the detected humans and the detected objects of interest (e.g. “horse”) as illustrated in Fig. 2.

Multi-stream Architecture Given a human-object proposal, our HO-RCNN classifies its HOIs using a multi-stream network (Fig. 3), where different streams extract features from different sources. To illustrate our idea, consider the classification of one HOI class “riding a bike”. Intuitively, local information around humans and objects, such as human body poses and object local contexts, are critical in distinguishing HOIs: A person riding a bike is more likely to be in a sitting pose rather than standing; a bike being ridden by a person is more likely to be occluded by the person’s body in the upper region than those not being ridden. In addition, human-object spatial relations are also important cues: The position of a person is typically at the middle top of a bicycle when he is riding it. Our multi-stream architecture is composed of three streams which encode the above intuitions: (1) The *human stream* extracts local features from the detected humans. (2) The *object stream* extracts local features from the detected objects. (3) The *pairwise stream* extracts features which encode pairwise spatial relations between the detected human and object. The last layer of each stream is a binary classifier that outputs a confidence score for the HOI “riding a bike”. The final confidence score is obtained by summing the scores over all streams. To extend to mulitple HOI classes, we train one binary classifier for each HOI class at the last layer of each stream. The final score is summed over all streams separately for each HOI class.

Human and Object Stream Given a human-object proposal, the human stream extracts local features from the human bounding box, and generates confidence scores for each HOI class. The full image is first cropped using the bounding box and resized to a fixed size. This normalized image patch is then passed into a ConvNet that extracts features through a seires of convolutional, max pooling, and fully-connected layers. The last layer is a fully-connected layer of size K, where K is the number of HOI classes of interest, and each output corresponds to the confidence score of one HOI class. The object stream follows the same design except that the input is cropped and resized from the object bounding box of the human-object proposal.

Pairwise Stream Given a human-object proposal, the pairwise stream extracts features that encode the spatial relations between the human and object, and generates a confidence score for each HOI class. Since the focus is on spatial configurations of humans and objects, the input of this stream should ignore pixel values and only exploit informa-

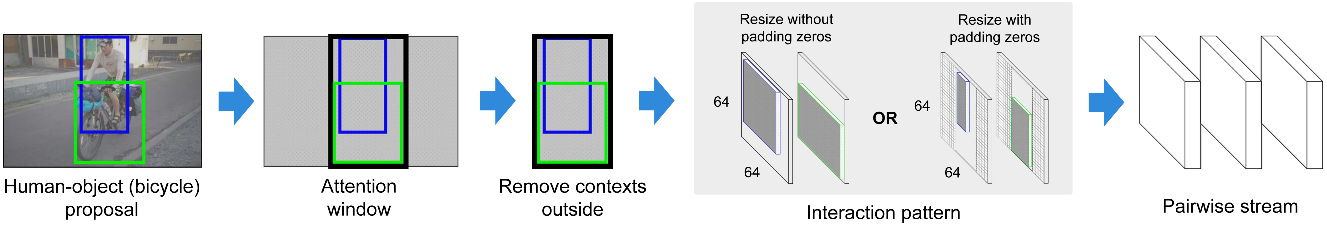

Human-object (bicycle) proposal
Attention window
Remove contexts outside
Resize without padding zeros
Resize with padding zeros
64
64
OR
64
64
Interaction pattern
Pairwise stream

Figure 4: Construction of Interaction Patterns for the pairwise stream.

tion of bounding box locations. Instead of directly taking the bounding box coordinates as inputs, we propose *Interaction Patterns*, a special type of DNN input that characterizes the relative location of two bounding boxes. Given a pair of bounding boxes, its Interaction Pattern is a binary image with two channels: The first channel has value 1 at pixels enclosed by the first bounding box, and value 0 elsewhere; the second channel has value 1 at pixels enclosed by the second bounding box, and value 0 elsewhere. [1] In our pairwise stream, the first channel corresponds to the human bounding box and the second channel corresponds to the object bounding box. Take the input image in Fig. 3 as an example, where the person is "riding a bike". The first (human) channel will have value 1 at the upper central region, while the second (object) channel will have value 1 at the lower central region. This representation enables DNN to learn 2D filters that respond to similar 2D patterns of human-object spatial configurations.

While the Interaction Patterns are able to characterize pairwise spatial configurations, there are still two important details to work out. First, the Interaction Patterns should be invariant to any joint translations of the bounding box pair. In other words, the Interaction Patterns should be identical for identitcal pair configurations whether the pair appears on the right or the left side of the image. As a result, we remove all the pixels outside the "attention window", i.e. the tightest window enclosing the two bounding boxes, from the Interaction Pattern. This makes the pairwise stream focus solely on the local window containing the target bounding boxes and ignore global translations. Second, the aspect ratio of Interaction Patterns may vary depending on the attention window. This is problematic as DNNs take input of fixed size (and aspect ratio). We propose two strategies to address this issue: (1) We resize both sides of the Interaction Pattern to a fixed length regardless of its aspect ratio. Note that this may change the aspect ratio of the attention window. (2) We resize the longer side of the Interaction Pattern to a fixed length while keeping the aspect ratio, followed by padding zeros on both sides of the shorter side to achieve the fixed length. This normalizes the size of the Interaction Pattern while keeping the aspect ratio of the attention window. The construction of Interaction Patterns is illustrated in Fig. 4.

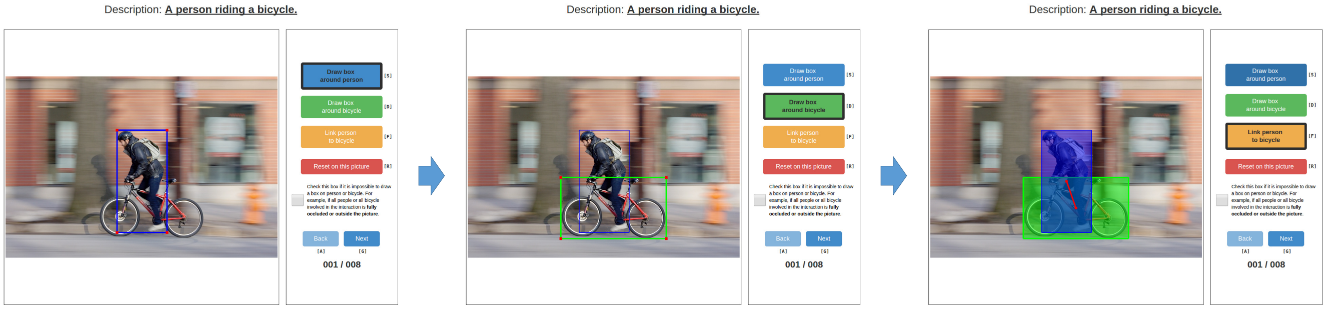
Training with Multi-Label Classification Loss Given a human-object proposal, our HO-RCNN generates confidence scores for a list of HOI categories of interest. As noted in [1], a person can concurrently perform different classes of actions to a target object, e.g. a person can be "riding" and "holding" a bicycle at the same time. Thus HOI recognition should be treated as a mulit-label classification as opposed to the standard K-way classification. As a result, we train the HO-RCNN by applying a sigmoid cross entropy loss on the classification output of each HOI category, and compute the total loss by summing over the individual losses.

4. Constructing HICO-DET

We contribute a new large-scale benchmark for HOI detection by augmenting HICO [1] with instance annotations. HICO currently contains only image-level annotations, i.e. 600 binary labels indicating the presence of the 600 HOI classes of interest (e.g. "feeding a cat", "washing a knife"). We further annotate the HOI instances present in each image, where each instance is represented by a pairing between a human and object bounding box with a class label (Fig. 6).

We collect human annotations by setting up annotation tasks on Amazon Mechanical Turk (AMT). However, there are two key issues: First, given an image and a presented HOI class (e.g. " riding a bike"), the annotation task is not as trivial as drawing bounding boxes around all the humans and objects associated with the interaction (e.g. "bike") — we also need to identify the interacting relations, i.e. linking each person to the objects he is interacting with. Second, although this linking step can be bypassed if the annotator is allowed to draw only one human bounding box followed by one object bounding box each time, such strategy is time intensive. Considering the cases where there are multiple people interacting with one object (e.g. "boarding an airplane" in Fig. 6), or one person interacting with multiple objects (e.g. "herding cows" in Fig. 6), the annotator then

[1] In this work, we apply the *second-order* Interaction Pattern for learning pairwise spatial relations. The Interaction Pattern can be extended to n-th order ($n \in N$) by stacking additional images in the channel axis for learning higher-order relations.



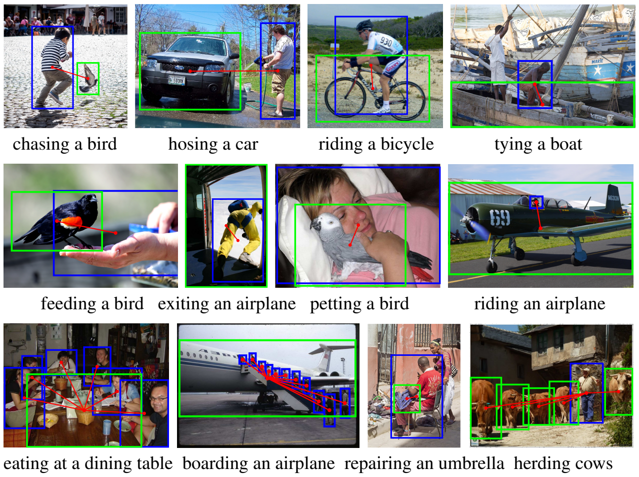
Description: A person riding a bicycle.

Check this box if it is impossible to draw a box on person or bicycle. For example, if all people or all bicycle involved in the interaction is **fully occluded or outside the picture**.

Draw box around person [S]
Draw box around bicycle [D]
Link person to bicycle [F]
Reset on this picture [R]
Back [A]
Next [G]
001 / 008

Step 1 Step 2 Step 3

Figure 5: Our data annotation task for each image involves three steps.



chasing a bird, hosing a car, riding a bicycle, tying a boat

feeding a bird, exiting an airplane, petting a bird, riding an airplane

eating at a dining table, boarding an airplane, repairing an umbrella, herding cows

Figure 6: Sample annotations of our HICO-DET.

has to repeatedly draw bounding boxes around the shared persons and objects. [2] To efficiently collect such annotations, we adopt a *three-step* annotation procedure (Fig. 5). For each image, the annotator is presented with a sentence description, such as "A person riding a bicycle", and asked to proceed with the following three steps:

Step 1: Draw a bounding box around each person. The first step is to draw bounding boxes around each person involved in the described interaction (e.g. each person riding bicycles). Note that the annotators are explicitly asked to ignore any person not involved in the described interaction, (e.g. any person not riding a bicycle), since those people do not participate in any instances of "riding a bicycle".

Step 2: Draw a bounding box around each object. The second step is to draw bounding boxes around each object involved in the described interaction, (e.g. each bicycle being ridden by someone). Similar to the first step, the annotator should ignore any object that is not involved in the described interaction (e.g. any bicycles not being ridden by someone).

	HICO-DET			
	#image	#positive	#instance	#bounding box
Train	38118	70373	117871 (1.67/pos)	199733 (2.84/pos)
Test	9658	20268	33405 (1.65/pos)	56939 (2.81/pos)
Total	47776	90641	151276 (1.67/pos)	256672 (2.83/pos)

Table 1: Statistics of annotations in our HICO-DET.

Step 3: Linking each person to objects. The final step is to link a person bounding box to an object bounding box if the described interaction is taking place between them (e.g. link a person to a bicycle if the person is riding the bicycle). Note that one person can be linked to multiple objects if he is interacting with more than one objects (e.g. "herding cows" in Fig. 6), and one object can be linked with multiple people if it is the case that more than one person are interacting with it (e.g. "boarding an airplane" in Fig. 6).

Note that in some rare cases, the involved person or object may be invisible, even though the presence of the HOI can be inferred from the image. (e.g. We can tell a person is "sitting on a chair" although the chair is fully-occluded by the person's body.) If the involved person or object is completely invisible in the image, the annotator is asked to mark those images as "invisible". Among all 90641 annotation tasks (each corresponds to one positive HOI label for one image in HICO), we found that there are 1209 (1.33%) tasks labeled as "invisible". Since our instance annotations are built upon HICO's HOI class annotations, our HICO-DET also has a long-tail distribution in the number of instances per HOI class as in HICO. By keeping the same training-test split, we found that there are 2 out of 600 classes ("jumping a car" and "repairing a mouse") which have no training instances due to the invisibility of people or objects. As a

[2] Although we formulate HOI detection as localizing interactions between a single person and a single object, actual interactions can be more complex such as the one-versus-many and many-versus-one cases. However, these types of interactions can be decomposed into multiple instances of person-object interaction pairs. Our goal is to detect all the decomposed person-object pairs in such cases.

result, we added 2 new images to our HICO-DET so we have at least one training instance for each of the 600 HOI classes. Tab. 1 shows the statistics of the newly collected annotations. We see that each image in HICO-DET has on average more than one (1.67) instance for each positive HOI label. Note that the total number of bounding boxes (256672) is less than twice the total number of instances (151274). This is because different instances can share people or objects, as shown in Fig. 6.

5. Experiments

Evaluation Setup Following the standard evaluation metric for object detection, we evaluate HOI detection using mean average precision (mAP). In object detection, a detected bounding box is assigned a true positive if it overlaps with a ground truth bounding box of the same class with intersection over union (IoU) greater than 0.5. Since we predict one human and one object bounding box in HOI detection, we declare a true positive if the minimum of human overlap IoU_h and object overlap IoU_o exceeds 0.5, i.e. $\min(\text{IoU}_h, \text{IoU}_o) > 0.5$. We report the mean AP over three different HOI category sets: (a) all 600 HOI categories in HICO (Full), (b) 138 HOI categories with less than 10 training instances (Rare), and (c) 462 HOI categories with 10 or more training instances (Non-Rare). All reported results are evaluated on the test set.

Following the HICO classification benchmark [1], we also consider two different evaluation settings: (1) *Known Object* setting: For each HOI category (e.g. “riding a bike”), we evaluate the detection only on the images containing the target object category (e.g. “bike”). The challenge is to localize HOI (e.g. human-bike pairs) as well as distinguishing the interaction (e.g. “riding”). (2) *Default* setting: For each HOI category, we evaluate the detection on the full test set, including images both containing and not containing the target object category. This is a more challenging setting as we also need to distinguish background images (e.g. images without “bike”).

Training HO-RCNN We first generate human-object proposals using state-of-the-art object detectors. Since HICO and MS-COCO [22] share the same 80 object categories, we train 80 object detectors using Fast-RCNN [8] on the MS-COCO training set. As detailed in Sec. 3, we generate proposals for each HOI category (e.g. “riding a bike”) by pairing the top detected humans and objects (e.g. “bike”) in each image. In our experiments, we adopt the top 10 detections for human and each object category, resulting in 100 proposals per object category per image.

We implement our HO-RCNN using Caffe [14]. For both the human and object streams, we adopt the CaffeNet architecture with weights pre-trained on the ImageNet classification task [30]. To train on HICO-DET, we run SGD with a global learning rate 0.001 for 100k iterations, and then lower the learning rate to 0.0001 and run for another 50k iterations. We use an *image-centric* sampling strategy similar to [8] for mini-batch sampling: Each mini-batch of size 64 is constructed from 8 randomly sampled images, with 8 randomly sampled proposals for each image. These 8 proposals are from three different sources. Suppose a sampled image contains interactions with “bike”, we sample: (a) 1 *positive example*: human-bike proposals that have $\min(\text{IoU}_h, \text{IoU}_o) \geq 0.5$ with at least one ground-truth instance from a category involving “bike”. (b) 3 *type-I negatives*: non-positve human-bike proposals that have $\min(\text{IoU}_h, \text{IoU}_o) \in [0.1, 0.5)$ with at least one ground-truth instance from a category involving “bike”. (c) 4 *type-II negatives*: proposals that do not involve “bike”.

	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
HO	5.73	3.21	6.48	8.46	7.53	8.74
HO+vec0 (fc)	6.47	3.57	7.34	9.32	8.19	9.65
HO+vec1 (fc)	6.24	3.59	7.03	9.13	8.09	9.45
HO+IP0 (fc)	7.07	4.06	7.97	10.10	8.38	10.61
HO+IP1 (fc)	6.93	3.91	7.84	10.07	8.43	10.56
HO+IP0 (conv)	7.15	4.47	7.95	10.23	8.85	10.64
HO+IP1 (conv)	**7.30**	**4.68**	**8.08**	**10.37**	**9.06**	**10.76**

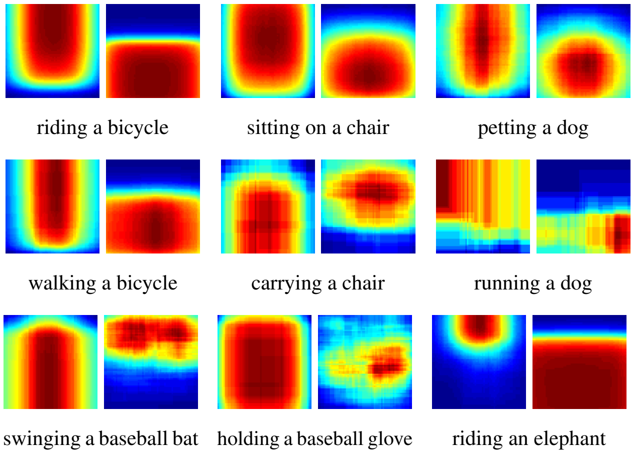
	Default		
	Full	Rare	Non-Rare
HO+vec1 (fc) vs. HO	< 0.001	0.132	< 0.001
HO+IP1 (conv) vs. HO	< 0.001	0.001	< 0.001
HO+IP1 (conv) vs. HO+vec1 (fc)	< 0.001	0.001	< 0.001

	Known Object		
	Full	Rare	Non-Rare
HO+vec1 (fc) vs. HO	< 0.001	0.077	< 0.001
HO+IP1 (conv) vs. HO	< 0.001	0.005	< 0.001
HO+IP1 (conv) vs. HO+vec1 (fc)	< 0.001	0.049	< 0.001

Table 2: Performace comparison of difference pairwise stream variants. Top: mAP (%). Bottom: p-value for the paired t-test.

Ablation Study We first perform an ablation study on the pairwise stream. We consider the two different variants of the Interaction Patterns described in Sec. 3, i.e. without padding (IP0) and with padding (IP1), each paired with two different DNN architectures: a fully-connected network (fc) and a convolutional network (conv). [3] We also report baselines that use the same fc architecture but take the 2D vector from human's center to object's center (vec0: without padding, vec1: with padding). Tab. 2 (top) reports the mAP of using the human and object stream alone (HO) as well as combined with different pairwise streams (vec0 (fc), vec1 (fc), IP0 (fc), IP1 (fc), IP0 (conv), IP1 (conv)). Note that

[3] The architecture is detailed in the supplementary material.



riding a bicycle

sitting on a chair

petting a dog

walking a bicycle

carrying a chair

running a dog

swinging a baseball bat

holding a baseball glove

riding an elephant

Figure 7: Average Interaction Patterns for different HOI categories obtained from ground-truth annotations. Left: average for the human channel. Right: average for the object channel.

for all the methods, the Default setting has lower mAPs than the Known Object setting due to the increasing challenge in the test set, and the rare categories have lower mAP than the none-rare categories due to sparse training examples. Although the mAPs are low overall (i.e. below 11%), we still observe in both settings that adding a pairwise stream improves the mAP. Among all pairwise streams, using Interaction Patterns with the conv architecture achieves the highest mAP (e.g. for IP1 (conv) on the full dataset, 7.30% in the Default setting and 10.37% in the Known Object setting). To demonstrate the signficance of the improvements, we perform *paired t-test*: We compare two methods by their AP difference on each HOI category. The null hypothesis is that the mean of the AP differences over the categories is zero. We show the p-values in Tab. 2 (bottom). While the 2D vector baselines outperform the HO baseline in mAP, the p-value is above 0.05 on rare categories (e.g. 0.13 for "HO+vec1 (fc) vs. HO" in the Default setting). On the other hand, "HO+IP1 (conv) vs. HO" and "HO+IP1 (conv) vs. HO+vec1 (fc)" both have all p-values below 0.05, suggesting that using Interaction Patterns with the conv architecture has a significant improvement not only over the HO baseline, but also over the 2D vector baseline.

We show the average Interaction Patterns obtained from the ground-truth annotations of different HOIs in Fig. 7. We see distinguishable patterns for different interactions on the same object category. For example, a chair involved in "sitting on a chair" is more likely to be in the lower region of the Interaction Pattern, while a chair involved in "carrying a chair" is more likely to be in the upper region.

We also separately evaluate the output of human, object, and pairwise stream. Tab. 3 shows the mAP of each stream on HO+IP1 (conv). The object stream outperforms the other two in the Default setting. However, in the Known Object setting, the pairwise stream achieves the highest mAP, demonstrating the importance of human-object spatial relations for distinguishing interactions.

	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
Human	0.70	0.08	0.88	2.44	2.14	2.53
Object	**2.11**	**1.19**	**2.39**	3.09	**2.98**	3.13
Pairwise	0.30	0.06	0.37	**3.21**	2.80	**3.33**

Table 3: mAP (%) of each stream on HO+IP1 (conv).

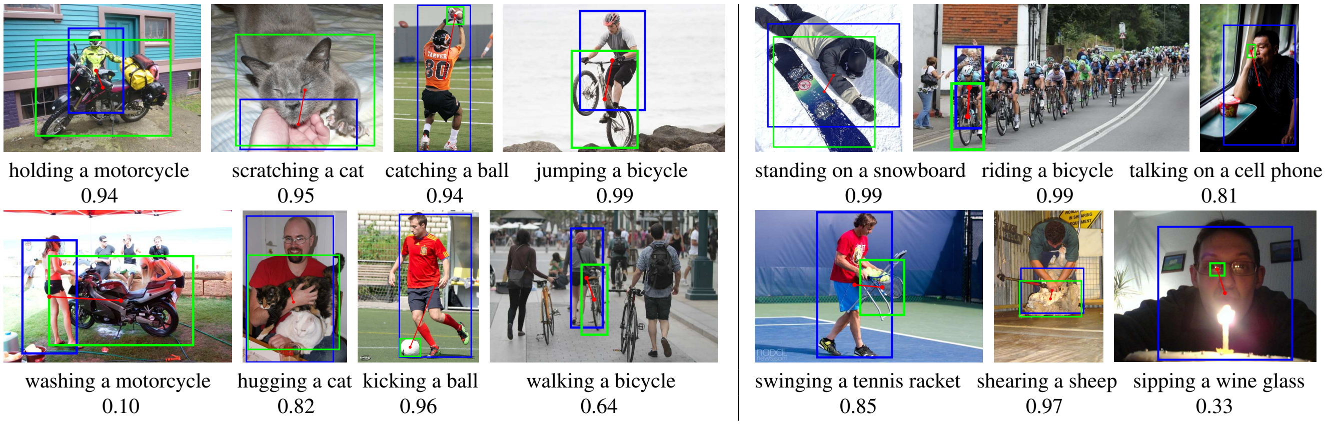
	Default			Known Object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
HO	5.73	3.21	6.48	8.46	7.53	8.74
HO+S	6.07	3.79	6.76	8.09	6.79	8.47
HO+IP1 (conv)	7.30	4.68	8.08	10.37	**9.06**	10.76
HO+IP1 (conv)+S	**7.81**	**5.37**	**8.54**	**10.41**	8.94	**10.85**

	Default		
	Full	Rare	Non-Rare
HO+S vs. HO	0.002	0.024	0.016
HO+IP1 (conv)+S vs. HO+IP1 (conv)	< 0.001	0.028	< 0.001

Table 4: Performance comparison of combining object detection scores. Top: mAP (%). Bottom: p-value for the paired t-test.

Leveraging Object Detection Scores So far we assume the human-object proposals always consist of true object detections, so the HO-RCNN is only required to distinguish the interactions. In practice, the proposals may contain false detections, and the HO-RCNN should learn to generate low scores for all HOI categories in such case. We thus add an extra path with a single neuron that takes the raw object detection score associated with each proposal and produces an offset to the final HOI detection scores. This provides a means by which the final detection scores can be lowered if the raw detection score is low. We show the effect of adding this extra component (HO+S and HO+IP1 (conv)+S) in Tab. 4 (top) and the signifiance of the improvements in Tab. 4 (bottom). The improvement is significant in the Default setting, since the extra background images increase the number of false object detections.

Error Analysis We hypothesize that the low AP classes suffer from excessive false negatives. To verify this hypothesis, we compute the recall of the human-object proposals for each HOI category. Tab. 5 shows the mean recall on the training set by varying the numbers of used human (object) detections. When adopting 10 human (object) detections, we see a low mean recall (46.75%), which expalins the low mAPs in our results. Although the mAPs can be potentially improved by adopting more human (object) detections, the number of human-object proposals will increase quadratically, making the evaluation of all proposals infeasible. This thus calls for better approaches to construct



holding a motorcycle 0.94
scratching a cat 0.95
catching a ball 0.94
jumping a bicycle 0.99
washing a motorcycle 0.10
hugging a cat 0.82
kicking a ball 0.96
walking a bicycle 0.64

standing on a snowboard 0.99
riding a bicycle 0.99
talking on a cell phone 0.81
swinging a tennis racket 0.85
shearing a sheep 0.97
sipping a wine glass 0.33

Figure 8: Qualitative examples of detections from our HO-RCNN. We show the HOI class and the output probability below each detection. Left: true positives. Right: false positives (left/middle/right: incorrect interaction class/inaccurate bounding box/false object detection).

	Number of human (object) detections			
	10	20	50	100
Full	46.75	51.56	57.17	60.37
Rare	54.15	58.62	64.98	68.40
Non-Rare	44.54	49.45	54.84	57.97

Table 5: Mean recall (%) of human-object proposals on the training set.

	Default		
	Full	Rare	Non-Rare
Random	1.35×10^{-3}	5.72×10^{-4}	1.62×10^{-3}
Fast-RCNN [8] (union)	1.75	0.58	2.10
Fast-RCNN [8] (score)	2.85	1.55	3.23
HO	5.73	3.21	6.48
HO+IP1 (conv)	7.30	4.68	8.08
HO+IP1 (conv)+S	**7.81**	**5.37**	**8.54**
	Known Object		
	Full	Rare	Non-Rare
Random	0.19	0.17	0.19
Fast-RCNN [8] (union)	2.51	1.75	2.73
Fast-RCNN [8] (score)	4.08	2.37	4.59
HO	8.46	7.53	8.74
HO+IP1 (conv)	10.37	**9.06**	10.76
HO+IP1 (conv)+S	**10.41**	8.94	**10.85**

Table 6: Comparison of mAP(%) with prior approaches.

high-recall human-object proposals in future studies.

Comparison with Prior Approaches To compare with prior approaches, we consider two extensions to Fast-RCNN [8]. (1) Fast-RCNN (union): For each human-object proposal, we take their attention window as the region proposal for Fast-RCNN. This can be seen as a "single-stream" version of HO-RCNN where the feature is extracted from the tightest window enclosing the human and object bounding box. (2) Fast-RCNN (score): Given the human-object proposals obtained from the object detectors, we train a classifier to classify each HOI category by linearly combining the human and object detection scores. Note that this method does not use any features from the human and object regions nor their spatial relations. We also report a baseline that randomly assigns scores to our human-object proposals (Random). Tab. 6 shows the mAP of the compared methods and different vairants of our HO-RCNN. In both settings, Fast-RCNN (union) performs worse than all other methods except the random baseline. This suggests that the feature extracted from the attention window is not suitable for distinguishing HOI, possibly due to the irrelevant contexts between the human and object when the two bounding boxes are far apart. Fast-RCNN (score) performs better than Fast-RCNN (union), but still worse than all our HO-RCNN variants. This is because object detection scores alone do not contain sufficient information for distinguishing interactions. Finally, our HO+IP1 (conv)+S and HO+IP1 (conv) outperform all other methods in both the Default and the Known Object setting. Fig. 8 shows qualitative examples of the detected HOIs from our HO-RCNN. We show both the true positives (left) and false positives (right).

6. Conclusion

We study the detection of human-object interactions in static images. We introduce HICO-DET, a new large benchmark, by augmenting the HICO classification benchmark with instance annotations. We propose HO-RCNN, a novel DNN-based framework. At the core of HO-RCNN is the Interaction Pattern, a novel DNN input that characterizes the spatial relations between two bounding boxes. Experiments show that HO-RCNN significantly improves the performance of HOI detection over baseline approaches.

Acknowledgement This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. OSR-2015-CRG4-2639.

References

[1] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 1, 2, 4, 6

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

[3] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*. 2016. 1, 2

[4] V. Delaitre, J. Sivic, and I. Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*. 2011. 1, 2

[5] C. Desai and D. Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*. 2012. 1, 2

[6] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshop on Structured Models in Computer Vision*, 2010. 1, 2

[7] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 1

[8] R. Girshick. Fast R-CNN. In *ICCV*, 2015. 1, 2, 3, 6, 8

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 3

[10] A. Gupta, A. Kembhavi, and L. S. Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *TPAMI*, 31(10):1775–1789, Oct 2009. 1

[11] S. Gupta and J. Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[13] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang. Recognising human-object interaction via exemplar based modelling. In *ICCV*, 2013. 1, 2

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 6

[15] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 2

[16] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2

[17] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 2

[18] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 2

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*. 2012. 1

[20] Y. Li, W. Ouyang, X. Wang, and X. Tang. ViP-CNN: Visual phrase guided convolutional neural network. In *CVPR*, 2017. 2

[21] X. Liang, L. Lee, and E. P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *CVPR*, 2017. 2

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*. 2014. 2, 6

[23] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*. 2016. 1, 2

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

[25] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 1, 2

[26] A. Mallya and S. Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*. 2016. 2

[27] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2

[28] A. Prest, C. Schmid, and V. Ferrari. Weakly supervised learning of interactions between humans and objects. *TPAMI*, 34(3):601–614, March 2012. 1, 2

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*. 2015. 1, 2, 3

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 6

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

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1

[33] B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 1, 2

[34] B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 1, 2

[35] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 1

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

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

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