No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques

Tanmay Gupta   Alexander Schwing   Derek Hoiem
University of Illinois Urbana-Champaign

{tgupta6, aschwing, dhoiem}@illinois.edu http://tanmaygupta.info/no_frills/

Abstract

We show that for human-object interaction detection a relatively simple factorized model with appearance and layout encodings constructed from pre-trained object detectors outperforms more sophisticated approaches. Our model includes factors for detection scores, human and object appearance, and coarse (box-pair configuration) and optionally fine-grained layout (human pose). We also develop training techniques that improve learning efficiency by: (1) eliminating a train-inference mismatch; (2) rejecting easy negatives during mini-batch training; and (3) using a ratio of negatives to positives that is two orders of magnitude larger than existing approaches. We conduct a thorough ablation study to understand the importance of different factors and training techniques using the challenging HICO-Det dataset [4].

1. Introduction

Human-object interaction (HOI) detection is the task of localizing all instances of a predetermined set of human-object interactions. For example, detecting the HOI "human-row-boat" refers to localizing a "human," a "boat," and predicting the interaction "row" for this human-object pair. Note that an image may contain multiple people rowing boats (or even the same boat), and the same person could simultaneously interact with the same or a different object. For example, a person can simultaneously "sit on" and "row" a boat while "wearing" a backpack.

Recently, increasingly sophisticated techniques have been proposed for encoding position and appearance for HOI detection. For instance, Chao *et al*. [4] encode the configuration of human-object box pairs using a CNN operating on a two channel binary image called the *interaction pattern*. Gkioxari *et al*. [10] predict a distribution over target object locations based on human appearance using a mixture density network [1]. For encoding appearance, approaches range from multitask training of a human-centric
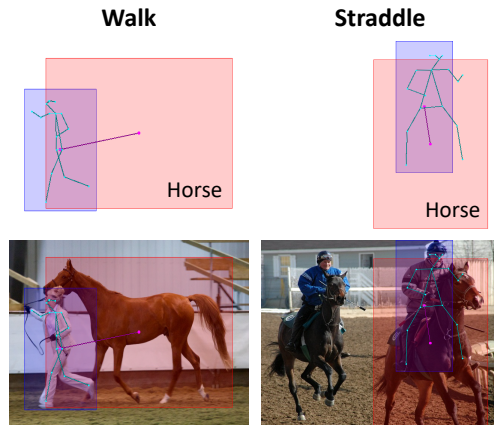

Figure 1: **Outputs of pretrained object and human-pose detectors provide strong cues for predicting interactions**. Top: human and object boxes, object label, and human pose predicted by Faster-RCNN [20] and OpenPose [2] respectively. We encode appearance and layout using these predictions (and Faster-RCNN features) and use a factored model to detect human-object interactions. Bottom: boxes and pose overlaid on the input image.

branch [10] alongside object classification, to using an attention mechanism which gathers contextual information from the image [9].

In this work, we propose a no-frills model for HOI detection. In contrast to sophisticated end-to-end models, we use appearance features from pretrained object detectors, and encode layout using hand-crafted bounding-box coordinate features (optionally human pose keypoints). Our network architecture is also modest, comprising of light-weight multi-layer perceptrons (MLPs) that operate on these appearance and layout features. In spite of these simplifications, our model achieves state-of-the-art performance on the challenging HICO-Det dataset.

Our gains are due to the choice of factorization, direct encoding and scoring of layout, and improved training techniques. Our model consists of human/object detection terms and an interaction term. The interaction term further consists of human and object appearance, box-configuration, and pose or fine-layout factors. We perform a thorough ablation study to evaluate the effect of each factor.

In contrast to existing work, which needs to train a
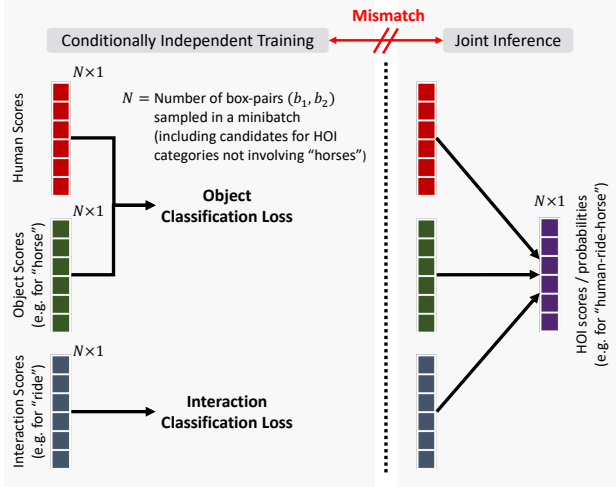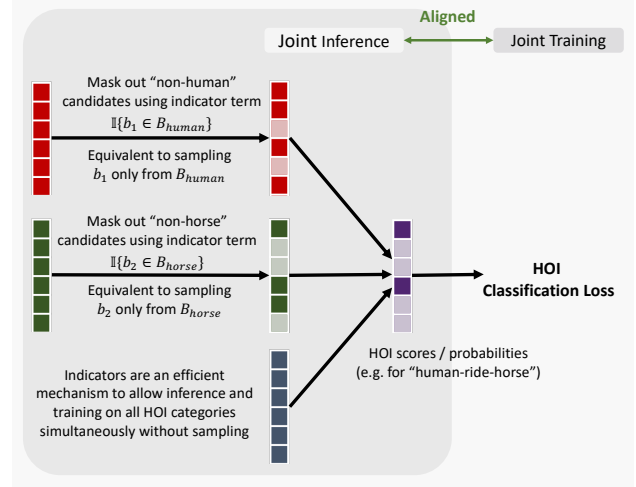
Figure 2: Proposed training techniques. The figure shows training and inference on a single HOI class ("human-ride-horse") for simplicity. **Eliminating training-inference mismatch:** As shown in (a), existing models [10, 9] often train human/object and interaction branches using object and interaction classification losses. The scores produced by these branches are simply multiplied during testing to produce the final HOI scores. Hence, training does not accurately reflect the test objective. Our model, shown in (b), fixes this mismatch by directly optimizing the combined scores using a multi-label HOI classification loss. **Rejecting easy negatives:** In a mini-batch, we treat candidate box-pairs for HOI categories other than "human-ride-horse" as easy negatives and their probability for HOI category "human-ride-horse" is set to 0 both during training and inference using indicator terms. This is implemented efficiently by applying binary masks to predicted probabilities. Specifically, if a human candidate box b_1 does not belong to the set of "human" proposals B_{human}, or if the object candidate box b_2 does not belong to the set of "horse" proposals B_{horse}, the mask entry corresponding to (b_1, b_2) and HOI category "human-ride-horse" is set to 0.

CNN [4] or a mixture density network [10] to encode layout, we use hand-crafted absolute and relative position features computed from bounding boxes or human pose keypoints. Our choice is motivated by the observation illustrated in Fig. 1: pretrained object and pose detectors provide strong geometric cues for interaction prediction.

We also develop the following training techniques for improving learning efficiency of our factored model:

(1) **Eliminating train-inference mismatch.** [10, 9] learn detection and interaction terms via separate detection and interaction losses. During inference, the scores of all factors are simply multiplied to get final HOI class probabilities. Instead, we directly optimizing the HOI class probabilities using a multi-label HOI classification loss (Fig. 2) (Interaction Loss: 15.89 mAP *vs.* HOI Loss: 16.96 mAP).

(2) **Rejecting easy negatives using indicator terms.** Rejecting easy negatives is beneficial not only during test but also during training because it allows the model to focus on learning to score hard negatives. We generate a candidate box-pair (b_1, b_2) using a pre-trained object detector which is then scored by the factor model. If either b_1 is not a "human" candidate (category h) or b_2 is not an object candidate o, then the factor model should predict a 0 probability of (b_1, b_2) belonging to HOI category (h, o, i) for *any* interactions i. This is achieved by including *indicator* terms in our object detection factors and can be implemented efficiently by applying a mask on predicted probabilities constructed from labels predicted by the object detector (Fig. 2) (w/o

indicators: 15.93 mAP *vs.* w indicators: 16.96 mAP).

(3) **Training with large negative to positive ratio.** We construct training mini-batches by sampling a two orders of magnitude larger number of negative box-pairs per positive pair than related work (1000 *vs.* < 10). Higher ratios compared to *object* detector training are expected since the number of negative pairs is quadratic in the number of object proposals as opposed to being linear for object detectors (neg. to pos. ratio 10: 13.40 mAP *vs.* 1000: 16.96 mAP).

In summary, our key contributions are: (1) a simple but competitive model for HOI detection that takes advantage of appearance and layout encodings from a pre-trained object detector (and optionally a pose detector); (2) a comparison of coarse and fine-grained layout encodings; and (3) techniques for enhancing learning efficiency of our model.

2. Related Work

Assessing interactions between humans and objects in images is a challenging problem which has received a considerable amount of attention [25, 24, 8, 7, 6, 17].

Human activity recognition is among the early efforts to analyze human actions in images or videos. Benchmarks such as UCF101 [22] and THUMOS [12] focus on classifying a video sequence into one of 101 action categories. While UCF101 uses carefully trimmed videos, the THUMOS challenge additionally introduced the task of temporal localization of activities in untrimmed videos. Image action recognition benchmarks such as Stanford 40 Actions [26]

and PASCAL VOC 2010 [17] have also been used in the literature. While similar in intent, these action recognition challenges differ from human-object interaction detection in three ways: (1) the tasks are limited to images or videos containing a *single* human-centric action, such as bowling, diving, fencing, *etc*.; (2) the action classes are *disjoint* and often involve interaction with an object unique to the activity (allowing models to cheat by simply recognizing the object); and (3) spatial localization of neither the person nor the object is required.

Moving from actions to interactions, Chao *et al*. [5, 4] introduce the HICO and HICO-DET datasets. The HICO dataset consists of images annotated with 600 human-object interactions with a diverse set of 117 interactions with 80 COCO [15] object categories. Unlike previous tasks, HOI classification is multi-label in nature since each image may contain multiple humans interacting with the same or different objects. Chao *et al*. [4] extend the HICO dataset with exhaustive bounding box annotations for each of the HOI classes to create HICO-DET. Due to the human-centric nature of the annotation task and the predefined set of objects and interactions, HICO-DET does not suffer from the missing annotation problem (at least to the same extent) that plagues datasets such as Visual Genome [14] and VRD [16] that are used for general visual relationship (object-object interaction) detection.

In a similar effort, Gupta *et al*. [11] augment the COCO dataset [15], annotating people (agents) with one of 26 action labels along with location and labels of objects fulfilling semantic roles for the action. Moreover, for semantic role labeling (SRL), Yatskar *et al*. [27] create an image dataset for situation recognition, which subsumes recognition of activity, participating objects and their roles.

In this work, we choose HICO-DET as a test bed due to its large, diverse, and exhaustive HOI annotations which allows for an accurate and meaningful evaluation.

Existing models for HOI detection. Chao *et al*. [4] propose HO-RCNN, a 3-stream architecture with one stream each for a human candidate, an object candidate, and a geometric encoding of the pair of boxes using the proposed *interaction pattern*. Each stream produces scores for every possible object-interaction category (600 for HICO-DET). The 3 sets of scores are combined using late-fusion to get the final prediction. Note that this approach treats "ride bicycle" and "ride horse" as independent visual entities and does not use the knowledge of "ride" being a common component. In contrast, our approach exploits this compositionality to learn shared visual appearance and geometric representations (*e.g*., "ride" typically involves a human box above an object box). In other words, weight sharing between different HOI classes in our factored model makes it more data efficient than [4] which predicts scores for 600 HOI categories using independent weights in the last 600-

way fully connected layer in each of the 3 streams.

Gkioxari *et al*. [10] propose *InteractNet*, which takes a multitask learning [3] perspective. The idea is to augment the Faster-RCNN [20] object detection framework with a human-centric branch and an interaction branch that are trained jointly alongside the original object recognition branch. To incorporate geometric cues, a Mixture Density Network (MDN) [1] is used to produce parameters of the object location distribution given the human appearance. This distribution is used to score candidate objects for a given human box. The model is trained using an object classification loss for the object branch, interaction classification losses for the human-centric action classification branch and the optional interaction branch, and a smooth L1 loss between the ground truth box-pair encoding and mean predicted by the localization MDN. During inference, predictions from these branches are fused heuristically. In addition to differences in the details of factorization, and appearance and layout encodings used in our model, we introduce training techniques for enhancing learning efficiency of similar factored models for this task. We optimize the final HOI score obtained after fusing the individual factor scores. We also encode more directly a box-pair layout using absolute and relative bounding box features which are then scored using a dedicated factor.

Gao *et al*. [9] follow an approach similar to [10] but introduce an attention mechanism that augments human and object appearance with contextual information from the image. An attention map is computed using cues derived from the human/object appearance encoding and the context is computed as an attention weighted average of convolution features. The model is trained using an interaction classification loss. In contrast, the only sources of contextual information in our model are the ROI pooled region features from the object detector. Adding a similar attention mechanism may further improve performance.

3. No-Frills HOI Detection

In the following, we first present an overview of the proposed model, followed by details of different factors and our training strategy.

3.1. Overview

Given an image x and a set of object-interaction categories of interest, human-object interaction (HOI) detection is the task of localizing all human-object pairs participating in one of the said interactions. The combinatorial search over human and object bounding-box locations and scales, as well as object labels, $\mathcal{O}$, and interaction labels, $\mathcal{I}$, makes both learning and inference challenging. To deal with this complexity, we decompose inference into two stages (Alg. 1). In the **first stage**, object category specific bounding box candidates $B_o \ \forall \ o \in \mathcal{O}$ are

Algorithm 1: Inference on a single image

Input : Image x,
             Set of objects ($\mathcal{O}$), interactions ($\mathcal{I}$), and
                HOI ($\mathcal{H} \subseteq \{h\} \times \mathcal{O} \times \mathcal{I}$) classes of interest,
             Pretrained object (Faster-RCNN) and
                human-pose (OpenPose) detectors
     // Stage 1: Create a set of box candidates for each object (including human)
1 Run Faster-RCNN on $x \ \forall \ o \in \mathcal{O}$ to get 300 region proposals (R_o)
2   with ROI appearance features and detection probabilities
3 **foreach** $o \in \mathcal{O}$ **do**
4   Construct $B_o = \{b \in R_o$ such that
5 b survives NMS (threshold 0.3) and
6 $P_{\text{det}}(l_{\text{det}} = o|b, x) > 0.01\}$
7   Update B_o to keep at most 10 highest ranking detections.
8 **end**
9 Run OpenPose on x to get skeletal-keypoints $k(b) \ \forall \ b \in B_h$
     // Stage 2: Score candidate pairs using the proposed factored model
10 **foreach** $(h, o, i) \in \mathcal{H}$ **do**
11   **foreach** $(b_1, b_2) \in B_h \times B_o$ **do**
12   Compute box configuration features using (b_1, b_2)
13   Compute pose features using $(k(b_1), b_1, b_2)$
14   Compute $P(y_{(h,o,i)} = 1|b_1, b_2, x)$
15   using equations 1, 2, and 3
16   **end**
       Output: Ranked list of $(b_h, b_o) \in B_h \times B_o$ as detections for class (h, o, i) with probabilities. For any $o' \neq o$ probability of $(b_h, b_{o'})$ belonging to class (h, o, i) is predicted as 0.
17 **end**
     // Steps 10-17 are implemented with a single forward pass on a mini-batch of precomputed features

---

selected using a *pre-trained* object detector such as Faster-RCNN (using non-maximum suppression and thresholding on class probabilities). For each HOI category, *i.e.*, for each triplet $(h, o, i) \in \mathcal{H}$, a set of candidate human-object box-pairs is constructed by pairing every human box candidate $b_h \in B_h$ with every object box candidate $b_o \in B_o$. In the **second stage**, a factored model is used to score and rank candidate box-pairs $(b_h, b_o) \in B_h \times B_o$ for each HOI category. Our factor graph consists of human and object appearance, box-pair configuration (coarse layout) and human-pose (fine-grained layout) factors. The factors operate on appearance and layout encodings constructed from outputs of pretrained object and human-pose detectors. The model is parameterized to share representations and computation across different object and interaction categories to efficiently score candidate box-pairs for all HOI categories of interest in a single forward pass.

3.2. Factored Model

For an image x, given a human-object candidate box pair (b_1, b_2), human pose keypoints $k(b_1)$ detected inside b_1 (if any), and the set of box candidates for each object category, the factored model computes the probability of occurrence of human-object interaction (h, o, i) in (b_1, b_2) as follows:

$$
\begin{aligned}
&P(y_{(h,o,i)} = 1|b_1, b_2, x, k(b_1), B_h, B_o) \quad (1) \\
&= P(y_h = 1, y_o = 1, y_i = 1|b_1, b_2, x, k(b_1), B_h, B_o) \\
&= P(y_h = 1|b_1, x, B_h) \cdot P(y_o = 1|b_2, x, B_o) \cdot \\
&\quad P(y_i = 1|b_1, b_2, k(b_1), x).
\end{aligned}
$$

Here, $y_h \in \{0, 1\}$ is a random variable denoting if b_1 is labeled as a human, $y_o \in \{0, 1\}$ denotes if b_2 is labeled as object category o, and $y_i \in \{0, 1\}$ denotes if the interaction assigned to the box-pair is i. The above factorization assumes that human and object class labels depend only on the individual boxes, the image, and the set of bounding-box candidates for the respective classes, while the interaction label depends on the box-pair, pose, and the image. B_h and B_o are used in the detection terms to compute the indicator functions for easy negative rejection. For brevity, we refer to the left hand side of Eq. (1) as $P(y_{(h,o,i)} = 1|b_1, b_2, x)$. We now describe the terms in detail.

3.2.1 Detector Terms
The first two terms in Eq. (1) are modeled using the set of candidate bounding boxes for each object class and classification probabilities produced by a pretrained object detector. For any object category $o \in \mathcal{O}$ (including h), the detector term is computed via

$$
P(y_o = 1|b, x, B_o) = \mathbb{1}(b \in B_o) \cdot P_{\text{det}}(l_{\text{det}} = o|b, x), \quad (2)
$$

where the P_{det} term corresponds to the probability of assigning object class o to region b in image x by the object detector. The indicator function checks if b belongs to B_o, the set of candidate bounding boxes for o, and sets the probability to 0 otherwise. Thus, easy negatives for class (h, o, i), *i.e.*, pairs (b_1, b_2) where either $b_1 \notin B_h$ or $b_2 \notin B_o$, are assigned 0 probability. Easy negative rejection is not only beneficial during test but also during training since model capacity is not wasted on learning to predict a low probability of class (h, o, i) for box-pairs that belong to any of the classes in the set $\{(h, o', i') \mid o' \in \mathcal{O} \setminus \{o\}, \ i' \in \mathcal{I}\}$.

3.2.2 Interaction Term
The interaction term refers to the probability of entities in b_1 and b_2 engaging in interaction $i \in \mathcal{I}$. To utilize appearance and layout information, the interaction term $P(y_i = 1|b_1, b_2, k(b_1), x)$ is further factorized as follows:

$$
\begin{aligned}
\sigma(&\phi_{\text{human}}(i|b_1, x) + \phi_{\text{object}}(i|b_2, x) + \\
&\phi_{\text{boxes}}(i|b_1, b_2) + \phi_{\text{pose}}(i|b_1, b_2, k(b_1))),
\end{aligned} \quad (3)
$$

where σ is the Sigmoid function and each ϕ is a learnable deep net factor. We now describe each of these factors along with the network architecture and appearance and layout encodings the factors operate with:
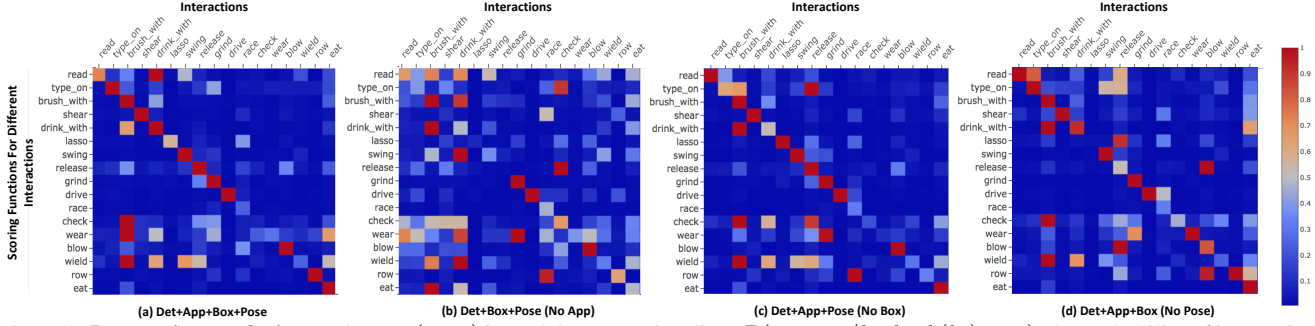
Figure 3: **Interaction confusions.** Element (m, n) in each heatmap visualizes $P(y_{i_m} = 1 | b_1, b_2, k(b_1), o, x)$, the probability of interaction $i_m \in \mathcal{I}$ for box-pair (b_1, b_2), averaged across all box pairs with ground truth interaction $i_n \in \mathcal{I}$. Each row m is independently normalized and exponentiated to highlight the interactions most confused with interaction i_m.

Appearance. Factors ϕ_{human} and ϕ_{object} predict the interaction that the human and the object are engaged in, based on visual appearance alone. The appearance of a box in an image is encoded using Faster-RCNN [20] (Resnet-152 backbone) average pooled *fc7* features extracted from the RoI. By design, this representation captures context in addition to content within the box. The 2048 dimensional *fc7* features are fed into a multi-layer perceptron (MLP) with a single 2048 dimensional hidden layer with Batch Normalization [13] and ReLU [18]. The output layer has 117 neurons, one per interaction category in $\mathcal{I}$.

Box Configuration. Object label and the absolute and relative positions and scales of the human and object boxes are often indicative of the interaction, even without the appearance (*e.g.*, a human box above and overlapping with a 'horse' box strongly suggests a 'riding' interaction). ϕ_{boxes} captures this intuition by predicting a score for each interaction given an encoding of the bounding boxes and the object label. The bounding boxes are represented using a 21-dimensional feature vector. We encode the *absolute position and scale* of both the human and object boxes using box width, height, center position, aspect ratio, and area. We also encode *relative configuration* of the human and object boxes using relative position of their centers, ratio of box areas and their intersection over union.

We also experiment with conditioning on the object label encoded as a $|\mathcal{O}| (= 80)$ dimensional one hot vector to allow the model to learn that only certain interactions are feasible for a given object. For example, it is possible to "clean" or "eat at" a "dinning table" (o) but not to "drive" or "greet" it. The object label encoding is concatenated with 21-dimensional position features and their log absolute values, and passed through an MLP with 2 hidden layers, 112 ($= 2 \times 21 + 80$) dimensional each (same as the input feature dimension), with Batch Normalization and ReLU. In practice, conditioning on o did not affect results significantly and hence is not included in Eq. 1 and Eq. 3 for simplicity.

Human Pose. We supplement the coarse layout encoded by bounding boxes with more fine-grained layout information provided by human pose keypoints. We use Open-Pose [2, 23, 21] to detect 18 keypoints for each person in

the image. A human candidate box is assigned a keypoint-skeleton if the smallest bounding box around the keypoints has 70% or more of its area inside the human box. Similar to box features, we encode both absolute human pose and the relative location with respect to the object candidate box. The *absolute pose features* ($18 \times 3 = 54$) consist of keypoint coordinates normalized to the human bounding box frame of reference and confidence of each keypoint predicted by OpenPose. The *relative pose features* ($18 \times 5 = 90$) consist of offset of the top left and bottom right corners of the object box relative to each keypoint and keypoint confidences. The absolute and relative pose features and their log values are concatenated along with a one-hot object label encoding before being passed through ϕ_{pose}. ϕ_{pose} is also an MLP with 2 hidden layers with $368 (= 2 \times (54 + 90) + 80)$ neurons each. Both hidden layers are equipped with Batch Normalization and ReLU. The output layer has 117 neurons.

Each factor eliminates some interaction confusions. Heatmaps in Fig. 3 show the interactions that are confused by models with all factors and models with one factor missing at a time. Comparing heatmap **b** with **a** shows the role of the appearance factor in reducing confusion between interactions. For instance, without *App*, "eat" is confused with "brush with" and "drink with," but not in the final model. Similarly, **c** and **d** can be compared with **a** for the effects of *Box* and *Pose* factors respectively.

3.3. Training

Since more than one HOI label might be assigned to a pair of boxes, the model is trained in a fully supervised fashion using the multi-label binary cross-entropy loss. For each image in the training set, candidate boxes for each HOI category ($B_h \times B_o$ for class (h, o, i)) are assigned binary labels based on whether both the human and object boxes in the pair have an intersection-over-union (IoU) greater than 0.5 with a ground truth box-pair of the corresponding HOI category. During training, the jth sample in a mini-batch consists of a box pair (b_1^j, b_2^j), HOI category $l_j \in \mathcal{H}$ for which the box pair is a candidate ($(b_1, b_2) \in B_h \times B_o$ are considered candidates for HOI class (h, o, i)), binary label y^j to indicate match (or not) with a ground truth box pair of

Models	Full	Rare	Non-Rare
HO-RCNN [4]	7.81	5.37	8.54
VSRL [11] (impl. by [10])	9.09	7.02	9.71
InteractNet [10]	9.94	7.16	10.77
GPNN [19]	13.11	9.34	14.23
iCAN [9]	14.84	10.45	16.15
Det	8.32	6.84	8.76
Det + Box	12.54	10.40	13.18
Det + Human App	11.12	8.82	11.80
Det + Object App	11.05	7.41	12.13
Det + App	15.74	11.35	17.05
Det + Human App + Box	15.63	**12.45**	16.58
Det + Object App + Box	15.68	10.47	17.24
Det + App + Box	**16.96**	11.95	**18.46**
Det + Pose	11.09	8.04	12.00
Det + Box + Pose	14.49	11.86	15.27
Det + App + Pose	15.50	10.14	17.10
Det + App + Box + Pose	**17.18**	**12.17**	**18.68**

Table 1: **Results on HICO-Det test set.** Det, Box, App, and Pose correspond to object detector terms, appearance, box configuration, and pose factors respectively. Each row was both trained and evaluated with specified factors. *Full* (all 600 classes), *Rare* (classes with < 10 training instances), and *Non-Rare* (rest) denote different subsets of HOI classes. **Best** and **second best** numbers are highlighted in color.

class l_j, detection scores for human and object category corresponding to class l_j, and input features for each factor ϕ. Pairs of boxes which are candidates for more than one HOI category are treated as multiple samples during training. Since the number of candidate pairs per image is 3 orders of magnitude (typically > 1000) larger than the number of positive samples (typically < 3), random sampling would leave most mini-batches with no positives. We therefore select all positive samples per image and randomly sample 1000 negatives per positive. Given a mini-batch of size N constructed from a single image x, the loss is

$$\mathcal{L}_{\text{mini-batch}} = \frac{1}{N|\mathcal{H}|} \sum_{j=1}^{N} \sum_{l \in \mathcal{H}} \mathbb{1}(l = l^j) \cdot \text{BCE}(y^j, p_l^j), \quad (4)$$

where $\text{BCE}(y, p)$ is the binary cross entropy loss and $p_l^j = P(y_l = 1 | b_1^j, b_2^j, x)$ is the probability of HOI class l computed for the jth sample using Eq. 1. In our experiments, we only learn parameters of the interaction term (*i.e.*, MLPs used to compute factors ϕ_{human}, ϕ_{box}, and ϕ_{pose}).

4. Experiments

Dataset. HICO-Det [4] and V-COCO [11] datasets are commonly used for evaluating HOI detection models. V-COCO is primarily used for legacy reasons since the early HICO [5] dataset only had image-level annotations. HICO-Det was created to extend HICO with bounding box annotations specifically for the HOI detection task. HICO-Det is both larger and more diverse than V-COCO. While HICO-Det consists of $47,776$ images annotated with 117 interactions with 80 objects resulting in a total of 600 HOI categories, V-COCO only has 26 interactions with a training

Neg./Pos.	Indicators	HOI Loss	Interaction Loss	mAP
10	✓	✓	✗	13.40
50	✓	✓	✗	15.51
100	✓	✓	✗	16.30
500	✓	✓	✗	**17.06**
1000	✓	✓	✗	**16.96**
1500	✓	✓	✗	16.62
1000	✗	✓	✗	15.93
1000	✓	✗	✓	15.89

Table 2: **Training techniques evaluated using Det + App + Box model.** The results highlight the importance of: (1) large *negative to positive ratio* in mini-batches; (2) using *indicators* during training to only learn to rank candidates selected specifically for a given HOI category instead of all detection pairs; (3) directly optimizing the *HOI classification* loss instead of training with an *interaction classification* loss and then combining with object detector scores heuristically. **Best** and **second best** numbers are shown in color.

set $1/12$ the size of HICO-Det's. Exhaustive annotations for each HOI category also make HICO-Det more suitable for AP based evaluation than VRD [16] which suffers from missing annotations. VRD also contains "human" as one among many subjects which makes evaluation of the impact of fine-grained human pose less reliable due to a small sample size. Hence, HICO-Det is best for evaluating our contributions.

HICO-Det contains $38,118$ training and $9,658$ test images annotated with 600 HOI categories. We use an 80-20 split of the training images to generate our actual training and validation sets. HOI categories consist of 80 object categories (same as COCO classes) and 117 interactions. Each image contains on average 1.67 HOI detections.

In addition to comparing to state-of-the-art, our experiments include factor ablation study (Tab. 1), impact of the proposed training techniques (Tab. 2), visualization of performance distribution across object and interaction categories (Fig. 4), and examples of top ranking detections and failure cases (Fig. 5).

4.1. Comparison to State-of-the-art

Tab. 1 shows that our final models *Det+App+Box* and *Det+App+Box+Pose* outperform existing approaches. We now highlight the key strengths of our approach in comparison to existing models.

Appearance does not need to be relearned. All existing approaches learn a task specific appearance encoding by either fine-tuning appearance encoding branches [10, 9, 11] or training a CNN from scratch [4]. We only use ROI pooled features from Faster-RCNN pretrained on MS-COCO [15].

Layout is directly encoded and scored. We directly encode layout using absolute and relative position features which are scored using ϕ_{box} (an MLP). Our formulation is easier to learn than *InteractNet* which predicts a distribution over target object locations using human appearance features alone. The explicit representation also makes our layout terms more efficient to learn than *HO-RCNN* which
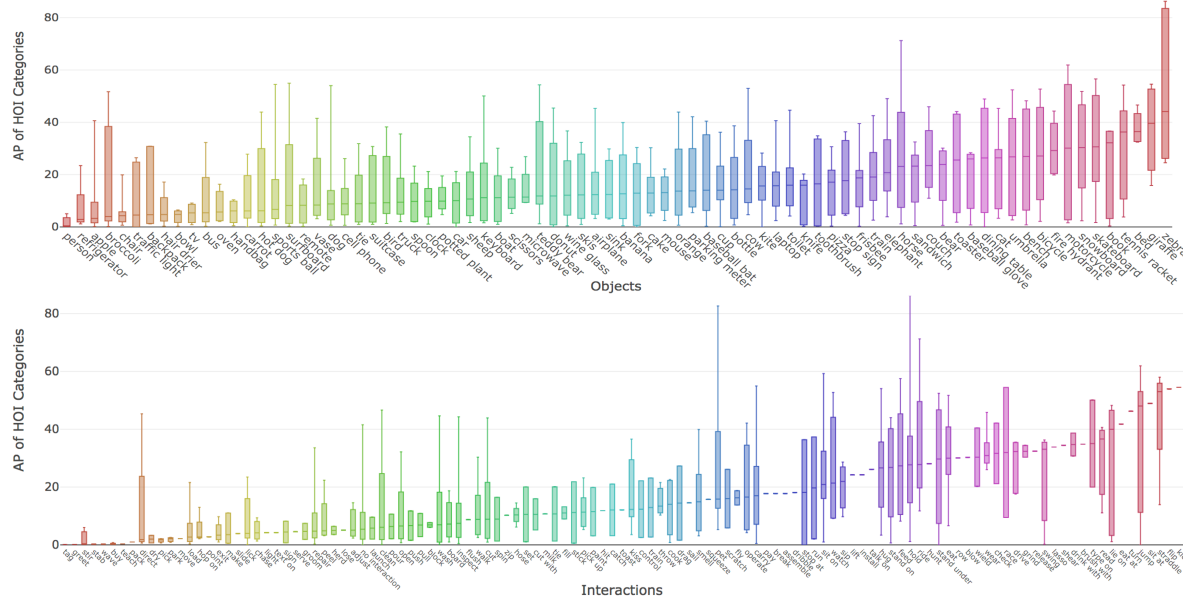
Figure 4: **Spread of performance** (range and quartiles) across interactions with the same object (top) and across objects for a given interaction (bottom). The horizontal axis is sorted by median AP.

needs to learn to encode layout (*interaction pattern*) using a CNN.

Weight sharing for learning efficiency. Weight sharing in our factored model (also in *InteractNet* and *iCAN*) makes it more data efficient than *HO-RCNN* which predicts scores for 600 HOI categories using independent weights in the last 600-way fully connected layer. In other words, *HO-RCNN* treats "ride-bike" and "ride-horse" as independent visual entities and does not benefit from the knowledge of "ride" being a common component.

ROI pooling for context. *iCAN* follows an approach similar to *InteractNet* but augments region appearance features with contextual features computed using an attention mechanism. While our model demonstrates strong performance with only ROI pooled detector features as the source of contextual information, we may further benefit from an attention mechanism similar to *iCAN*. *GPNN* also attempts to benefit from global context reasoning through message passing over an inferred graph. While in theory, such an approach *jointly* infers all HOI detections in an image (as opposed to making predictions for one candidate box-pair at a time), the advantages of this approach over simpler fixed graph approaches like our factor model and *iCAN* remains to be demonstrated.

Our model also benefits from improved training techniques which are discussed next.

4.2. Training Techniques

Tab. 2 shows the effect of proposed training techniques.
Training with large negative to positive ratio. Increasing the ratio of negative to positive box-pairs in a mini-batch during training leads to a dramatic increase in performance (neg. to pos. ratio 10: 13.40 mAP *vs.* 1000: 16.96 mAP).

Note that related work [4, 10] uses low ratios (typically < 10), similar to those used for training object detectors. For HOI detection, since the number of negative pairs is quadratic in the number of object proposals as opposed to linear for object detectors, higher ratios are expected.
Eliminating train-inference mismatch. Training the model using interaction classification loss on the probabilities predicted by the interaction term, as done in [10, 9], is suboptimal in comparison to training using HOI classification loss (15.89 *vs.* 16.96 mAP) even though the same set of parameters are optimized by both losses. This is because the latter calibrates the interaction term relative to the detection terms. A similar approach is used in [4] but without the strong weight sharing assumptions of our factor model.
Rejecting easy negatives. To allow the model to focus on learning to rank hard negatives correctly, we introduce indicator terms in our factor model. The indicator functions ensure that the factor model predicts zero probability of an HOI category (h, o, i) for box-pair (b_1, b_2) if $b_1 \notin B_h$ or $b_2 \notin B_o$. Tab. 2 shows that removing the indicator terms during training causes a drop in mAP from 16.96 to 15.93 (indicators still used during inference).

4.3. Factor Ablation Study

To identify the role of different sources of appearance and spatial information in our model, we train models with subsets of available factors.

The role of individual factors can be assessed by comparing *Det*, *Det+Box*, *Det+App*, and *Det+Pose*. Note that appearance terms lead to largest gains over *Det* followed by *Box* and *Pose*. We further analyze the contribution of human and object appearance towards predicting interactions. Interestingly, while *Det+Human App* and *Det+Object App*

Figure 5: **Qualitative results** showing top ranking **true** and **false** positives for different HOI categories with predicted probability. The blue and red boxes correspond to human and objects detected by a pretrained Faster-RCNN detector respectively. Pose skeleton consists of 18 keypoints predicted by the pretrained OpenPose detector and assigned to the human box.

perform comparably (11.12 and 11.05), the combination outperforms either of them with an mAP of 15.74, showing that the human and object appearance provide complementary information. Note that an mAP of 11.12 (= max(11.12, 11.05)) or less would indicate completely redundant or noisy signals. A similar sense of complementary information can be observed in Tab. 1 for *App-Box*, *App-Pose*, and *Box-Pose* pairs.

While *Det+Box+Pose* improves over *Det+Box*, *Det+App+Pose* and *Det+App* perform comparably. Similarly, *Det+App+Box+Pose* only slightly improves the performance of *Det+App+Box*. This suggests that while it is useful to encode fine-grained layout in addition to coarse layout, human appearance encoded via object detectors already captures human pose to some extent.

Another way of understanding the role of factors is to consider the drop in performance when a particular factor is removed from the final model. Relative to *Det+App+Box+Pose*, performance drops are 2.69, 1.68, and 0.22 mAP for *App*, *Box* and *Pose* factors respectively.

4.4. Performance Distribution

Fig. 4 visualizes the distribution of performance of our model across interactions with a given object and across objects for a given interaction. The figure shows that for most objects certain interactions are much easier to detect than others (with the caveat that AP computation for any class is sensitive to the number of positives for that class in the test set). A similar observation is true for different objects

given an interaction. In addition, we observe that interactions which can occur with only a specific object category (as indicated by absence of box) such as "kick-ball" and "flip-skateboard" are easier to detect than those that tend to occur with more than one object such as "cut" and "clean" and could have drastically different visual and spatial appearance depending on the object.

4.5. Qualitative Results

Qualitative results (Fig. 5) demonstrate the advantages of building HOI detectors on the strong foundation of object and pose detectors. False positives are more commonly due to incorrect interaction prediction than incorrect object/pose detection. Notice that cues for preventing false positives could be as subtle as gaze direction as in the case of "inspect airplane" and "watch bird."

5. Conclusion

We propose a no-frills approach to HOI detection which is competitive with existing literature despite its simplicity. This is achieved through appropriate factorization of the HOI class probabilities, direct encoding and scoring of layout, and improved training techniques. Our ablation study shows the importance of human and object appearance, coarse layout, and fine-grained layout for HOI detection. We also evaluate the significance of the proposed training techniques which can easily be incorporated into other factored models.

References

[1] Christopher M Bishop. Mixture density networks. 1994. 1, 3

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 5

[3] Rich Caruana. Multitask learning. In *Learning to learn*. Springer, 1998. 3

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2, 3, 6, 7

[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. 3, 6

[6] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In *NIPS*, 2011. 2

[7] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012. 2

[8] Chaitanya Desai, Deva Ramanan, and Charless Fowlkes. Discriminative models for static human-object interactions. In *CVPR Workshops*, 2010. 2

[9] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 1, 2, 3, 6, 7

[10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018. 1, 2, 3, 6, 7

[11] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3, 6

[12] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos in the wild. *CVIU*, 2017. 2

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 3

[15] 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. 3, 6

[16] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 3, 6

[17] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 2, 3

[18] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 5

[19] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 6

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 3, 5

[21] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 5

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[23] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 5

[24] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010. 2

[25] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010. 2

[26] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2

[27] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *CVPR*, 2016. 3