

Human Object Interaction Detection using Two-Direction Spatial Enhancement and Exclusive Object Prior

Lu Liu^{a,*}, Robby T. Tan^{a,b}

^a*Electrical and Computer Engineering Department, National University of Singapore
E4-06-20, 4 Engineering Drive 3, 117583, Singapore*
^b*Yale-NUS College, Singapore*

Abstract

Human-Object Interaction (HOI) detection aims to detect visual relations between human and objects in images. One significant problem of HOI detection is that non-interactive human-object pair can be easily mis-grouped and misclassified as an action, especially when humans are close and performing similar actions in the scene. To address the mis-grouping problem, we propose a spatial enhancement approach to enforce fine-level spatial constraints in two directions from human body parts to the object center, and from object parts to the human center. At inference, we propose a human-object regrouping approach by considering the object-exclusive property of an action, where the target object should not be shared by more than one human. By suppressing non-interactive pairs, our approach can decrease the false positives. Experiments on V-COCO and HICO-DET datasets demonstrate our approach is more robust compared to the existing methods under the presence of multiple humans and objects in the scene.

Keywords: human-object interaction detection, two-direction spatial enhancement, exclusive object prior, mis-grouped human-object pairs, non-interactive suppression

1. Introduction

Understanding relationships between individual objects is one of the important and challenging problems in visual recognition. To understand what is happening in a

*Corresponding author

Email addresses: lliu@u.nus.edu (Lu Liu), roddy.tan@yale-nus.edu.sg (Robby T. Tan)

scene, computers need to recognize how humans interact with surrounding objects. In this paper, we tackle the challenging task of detecting human-object interactions (HOI), which is to understand the interactions between humans and objects. Given an image, HOI detection aims to detect an interaction triplet $\langle \text{human}, \text{action}, \text{object} \rangle$. This requires to not only localize a human and an object instance, but also recognize the actions/interactions that the human is performing to the object, such as "ride bike" and "eat apple". HOI detection is an important step towards the high-level semantic understanding of human-centric scenes.

HOI detection problem is challenging since it requires to correctly detect and recognize fine-grained human-centric interactions (e.g. "feed horse", "straddle horse" and "jump horse"), and also involves multiple co-occurring actions (e.g. "reading book while sitting on a chair"). Moreover, multiple humans can appear in one scene performing the same interaction to the same object ("throw and catch ball"), or performing the same interaction to the different objects ("people ski in the field"). These complex and diverse interaction scenarios impose significant challenges to HOI detection task.

One of the significant problems in HOI detection is mis-grouping (see Fig. 1), which is no interaction human-object pair can be easily mis-grouped and misclassified as an action. The mis-grouping is one of the main causes of False Positives (FP) in HOI detections. We consider the mis-grouping error mainly comes from the complexity nature of combination, where multiple humans can interact with multiple objects via multiple actions/interactions. Specifically, when multiple objects lie around one human, this human can interact with any of these objects at the same time. When multiple humans and objects simultaneously appear in one scene, it becomes even harder to tell which human is performing which action to which object.

Given a human of interest (red box in Fig. 1) and an action of interest (e.g. ski), we want to find to which object (blue box in Fig. 1) this human is performing this action. The paring correctness depends on the correctness of the HOI scoring function, which is to predict a higher score for the correct human-object pair than the wrong ones. However, when multiple humans appear in the scene, existing HOI detection methods [1, 2, 3, 4, 5] can fail to identify the correct interaction pair, especially when humans are close and performing similar actions (e.g. people skiing in the field, or

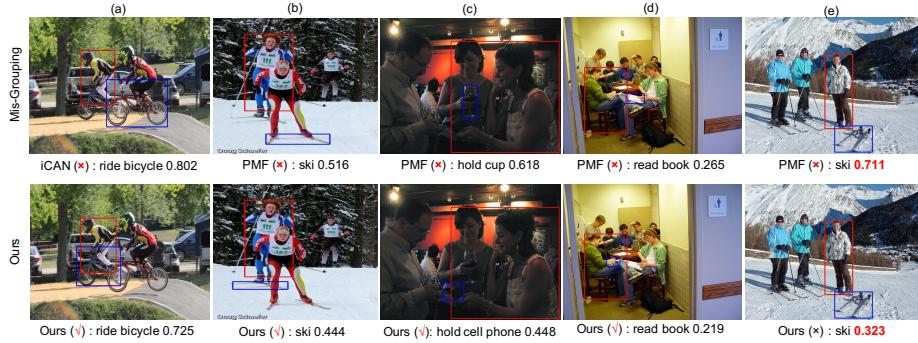


Figure 1: Example of HOI detections using the proposed approach, and iCAN [1], and the recently introduced PMFNet method [2]. The first row shows five typical mis-grouping mistakes made by iCAN [1] and PMFNet [2]. Existing methods fails to correctly pair the agent human (red box) and the target object (blue box). The mis-grouping can happen when the wrong target object is close to the interested human (a-e), or the correct target object is occluded (a-d). Our approach accurately groups the human and the respective object (row 2 a-d), and suppresses the score for the non-interactive pair even when no positive target object exists in the scene (row 2 e). Better viewed in color.

studying in a room in Fig. 1). We consider the spatial affinity, object occlusion and appearance similarity are the main causes that can weaken the confidence score of the correct human-object pair, resulting in a false positive HOI detection result (see Fig. 1 row 1).

To address the mis-grouping problem, InteractNet [3] propose max-object selection to only select one target object, which is the highest scoring object, for each human-action pair to form the final HOI triplet. Intuitively, for example, one human cannot ride more than one bike at the same time. It assumes that there is no such cases as human walking multiple dogs or herding multiple sheep. This max-object selection can filter out the mis-grouped detections made in iCAN [1] shown in Fig. 1 (a)¹. Although this max-object selection can suppress non-interactive target objects for each human, it cannot suppress non-interactive pairs between one object and multiple humans. For instance, two humans skiing one pair of skis dose not comply with the common sense. TIN [6] suppresses the mis-grouped non-interactive pairs by training

¹The score shown in Fig 1 (row 1 a) of iCAN has been normalized from [0, 2] to [0, 1] for comparison.

a binary interactiveness classifier to first filters out non-interactive pairs predicted by the interactiveness network using a threshold, and then predict HOI actions for the predicted interactive pairs. However, this first-stage hard filtering can exclude true positive pairs as well, and the interactiveness threshold setting is hand-crafted and empirical. The recent PMFNet [2] jointly trains the binary interactiveness classifier with the HOI action classifier, and fuse the scores as the final prediction. However, when multiple humans are performing similar actions in one scene, PMFNet [2] fails to correctly pair the agent human and the target object. For instance, in Fig. 1 (b), two humans are skiing together. PMFNet [2] fails to pick out the correct background skis for the background human, because the background skis are partially occluded, and the foreground skis are close to the background human. Moreover, when the human is not interacting with any objects in the scene such as in Fig. 1 (d), it becomes even harder to suppress such non-interactive pair, since the scene is just like skiing, and the skis are close to the human feet.

In this paper, our goal is to suppress non-interactive human-object pairs to address mis-grouping problem. We propose two approaches during training and inference respectively to not only predict a small action score for non-interactive human-object pairs, but also pick out the correct interactive target object for the interested human.

Firstly, we propose a spatial enhancement approach to enforce fine-level spatial constraints in two directions from human body parts to the object center (i.e. center point of the object box), and from object parts to the human center. Spatial relationship is an important cue to infer HOI actions. For example, if a human is riding a bike, the human should be just above the bike, and the central axis of human and object should be close to each other (Fig. 1 (a)). Previous methods [1, 4, 7, 6] only encode the rough spatial locations between the whole human bounding box and object bounding box. Recent methods extract spatial features between human body parts and the object center [2, 8]. However, we argue that purely encoding the object center is not distinguishing enough when the object is large and surrounding the human, such as "lie in bed" and "drive car". Also for objects that have several separate components, such as "a pair of skis", the center point cannot describe the shape and the exact location of the object instance. Hence, we propose symmetric spatial constraints from human parts to

the object center, and from object parts to the human center to enhance the spatial configurations between the target human and object. By encoding the detailed part-level spatial features, the network can learn more discriminative spatial features, which can help predict high confidence scores for interactive human-object pairs and low scores for non-interactive pairs.

To further suppress non-interactive pairs, we propose a human-object regrouping approach by considering the prior knowledge of the exclusive object of an action. For instance, as shown in Fig. 1 (b), existing method PMFNet [2] predicts both humans in foreground and background are skiing the foreground skis. However, this detection result does not comply with the common sense, because one pair of skis cannot be ridden by more than one human at the same time. Motivated by this observation, our idea is to take the object exclusion property into consideration. When predicting an object-exclusive action (e.g. "ski", and "read book"), the prior knowledge tells us the object should be exclusive to one human, and cannot be shared by multiple humans at the same time. Hence, if there are multiple humans sharing the same object instance, we assign this object to the max-scoring human, and reselect the next scoring target object for the other non-max humans.

Compared to previous methods [7, 3, 1, 4, 7, 6, 2, 8, 5], our approach enhances the spatial configurations between human and object, and takes advantage of action-specific exclusive object prior knowledge. Our approach can correct the mis-grouping errors (Fig. 1 row 1) made in existing methods [1, 2] and accurately groups the human and the corresponding object under the presence of multiple humans and objects with occlusions in the scene (Fig. 1 row 2). Hence, our approach can decrease the false positives in HOI detection results.

Our main contributions are three fold: (1) To address the mis-grouping problem, we propose a two-direction spatial enhancement approach to enforce fine-level spatial constraints between human keypoints and object parts. (2) we propose an action-specific exclusive object prior knowledge to regroup the object with the corresponding interactive human. (3) Extensive experiments on V-COCO [9] and HICO-DET [10] datasets demonstrate our approach is more robust compared to the existing methods under the presence of multiple humans and objects in the scene.

2. Related work

Object detection. Object detection [11, 12, 13, 14, 15] is an essential building block for scene understanding. Our work uses the off-the-shelf Mask RCNN [16, 17] to detect humans and object instances. Given the detected instances, our method aims to recognize interactions between all pairs of human and object instances.

Image Human Action Recognition. There has been great progress in understanding human actions from images in recent years [18, 19, 20]. Compared to action recognition from videos [21, 22], which highly relies on motion, action recognition from a single image depends on static cues, such as human poses [23], body parts [18, 19], and contexts [24]. R*CNN [24] discover possible interactive objects around the human as context cues to infer human actions. Zhao et al. [19] incorporate mid-level body part actions to infer body actions. Liu et al. [20] address the misleading object problem by jointly predicting human action and location, to drive the attentions of the network to the target human. Compared to general image action recognition, HOI detection also utilizes the similar visual cues, such as human poses, body parts, and contexts. Differently, HOI further requires to detect individual instances, and identify the corresponding interactive objects in the human-centric scene.

Visual Relationship Detection. Visual Relation Detection (VRD) aims to detect pairs of objects and classify their relationships simultaneously. Lu et al. [25] propose to learn language priors from semantic word embedding to refine relationship detection. Vtrans [26] has a visual translation network to model relation as a translation vector from subject to object. [27] exploits undetermined relationships, including unlabeled pairs, irrelevant pairs, and falsely detected object pairs, to help predict visual relationships. Recently, [28] leverages language models to learn separate visual-semantic embeddings for subject, object and predicate. Peyre et al. [29] propose to learn an additional embedding space for visual phrase classes. They further develop a model using analogy embedding transfer to infer unseen relationships. Different from general relationships detection between arbitrary objects, in our task, we focus on human-centric relationship detection, which aims to infer how humans interact with objects.

Human-Object Interaction Detection. Gupta and Malik [9] first explore the prob-

lem of visual semantic role labelling and construct V-COCO dataset for the role labelling task. In this problem, the goal is to detect the agent (human), an object, and predict the interaction class between them. Later, Chao et al. [7] introduce a new benchmark dataset HICO-DET, and propose a multi-stream network to aggregate human, object and spatial configurations for HOI detection. Gkioxari et al. [3] propose InteractNet to estimate action-specific density map over target locations based on detected human appearance. Some works [30, 4, 8] model the scene as a graph reasoning problem and use graph convolutional network to predict the HOI. [31, 32] adopt compositional learning strategy for zero-shot learning to predict unseen human-object combinations. Context-based methods mine the contextual information in the image to infer HOI. iCAN [1] designs an instance-centric attention module to highlight important regions in the image as contextual information for facilitating interaction prediction. DCA [5] extracts context-aware appearance features using squeeze-and-excitation operation. Part-based approaches [2, 33, 8] utilize human poses to extract detailed human body part features. Our work is, in particular, related to PMFNet [2] exploiting human keypoints to extract detailed part features to predict actions. However, in contrast to PMFNet [2], we additionally encode the spatial relationship from object parts to the human center to better recognize the fine-grained interactions and suppress the non-interactive pairs. TIN [6] shares the same the goal as ours, which is to suppress non-interactive human-object pairs. However, TIN [6] requires the availability of multiple datasets to train the interactiveness network, and has a transfer stage to fine-tune on the target dataset. In contrast, our approach is simple and effective, which only learns interactiveness knowledge on the target dataset, and the interactiveness network is jointly trained with action classification network.

3. Method

We now introduce our relation reasoning method for human-object interaction detection. The task is to detect humans and objects in the input image, as well as to recognize the actions for each $< \text{human}, \text{object} >$ pair. We develop a multi-branch deep neural network (Fig. 2) leveraging multi-level features with holistic instance-level

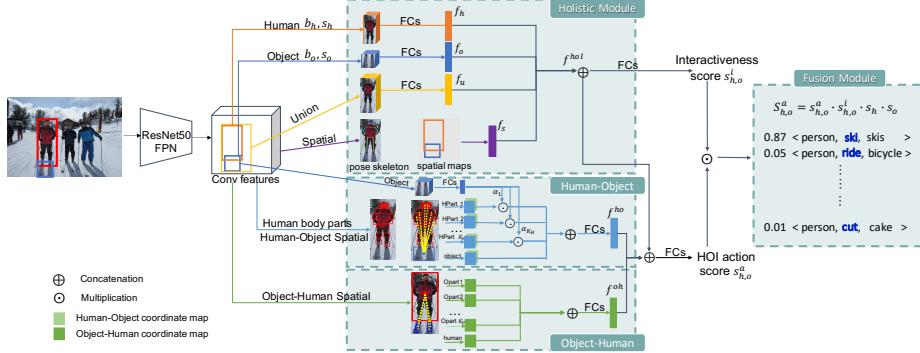


Figure 2: Overview of our framework. Given a pair of human-object proposal, holistic module employs human, object, union, and spatial-pose branches to extract appearance features from human, object and union boxes, and spatial features from pairwise location masks and human pose information. We employ two-direction spatial constraints symmetrically from human parts to the object center, and from object parts to the human center. We jointly predict interactiveness score and HOI action score. Finally, we combine the interactiveness score and action score to predict the final score for HOI actions. Better viewed in color.

appearance features, human part-level appearance features, and spatial features in both instance and part levels. We present the overview of our method in Sec. 3.1.

To address the mis-grouping problem, we propose a novel symmetric two-direction spatial enhancement module to encode the fine-level spatial configurations between human parts and object parts. We explain our spatial encoding method in details in Sec. 3.3. Furthermore, in the inference stage, we exploit the exclusive object prior knowledge to regroup human and object instances to further suppress the non-interactive human-object pair. We describe our regrouping strategy in Sec. 3.5.

3.1. Overview

We develop a bottom-up framework in which we first detect all the object proposals, and then predict the interactions between each $\langle \text{human}, \text{object} \rangle$ pair. Given an input image, we first employ Mask R-CNN from Detectron [17] to detect all the human/object bounding boxes b_h, b_o , human keypoints K_{ph} , and object instance segmentation masks m_o . We use s_h and s_o to denote the confidence scores for a detected human box b_h and object box b_o , respectively. All the human and object proposals will be paired up to form $\langle \text{human}, \text{object} \rangle$ pairs as HOI proposals. In the interaction

classification stage, for each human-object bounding box pair (b_h, b_o) , we predict the action score $s_{h,o}^a$ for each HOI action category a . Besides, we jointly predict a binary interactiveness score $s_{h,o}^i$ indicating whether the human is interacting with the object or not. The final HOI score $S_{h,o}^a$ is the combination of the confidence scores of the proposals s_h, s_o , interaction score $s_{h,o}^a$, and the interactiveness score $s_{h,o}^i$:

$$S_{h,o}^a = s_{h,o}^a \cdot s_{h,o}^i \cdot s_h \cdot s_o \quad (1)$$

3.2. Holistic Visual Representations

We adopt ResNet-50-FPN [15] as our convolutional backbone network to generate convolution feature maps Φ with dimension of $(H \times W \times D)$. We use Mask R-CNN [16] as the object detector to generate human-object pairs (b_h, b_o) . Given the detected human/object bounding boxes b_h, b_o , the holistic module consists of four branches: human branch, object branch, union branch, and spatial branch, as illustrated in Fig. 2. We extract the appearance features of human proposal b_h , object proposal b_o , and union proposal b_u from the shared feature map of ResNet-50 Conv5 by the ROI-Align operation [16] with resolution of $D_h \times D_h$. The union proposal b_u is the minimum box in the spatial region that contains b_h and b_o , which can the extract contextual information around humans and objects in the scene.

In addition to appearance features, spatial features can also provide important cues for recognizing interactions, such as "ride bike" vs. "walk bike". As for non-interactive pairs, if a human is far from a bike, he can hardly ride or walk the bike. And, if a human is riding a bike, the bike should be just below the human body, and not in front or behind the human body. We apply the two-channel binary masks human and object proposals in their union space to capture the spatial configurations of each human-object instance as in [7, 1, 2]. In addition, the pose structure of the human can also help differentiate interactions, such as "hit with the baseball bat" vs. "look at the tennis bat". As for non-interactive pairs, if a human is squatting down on the ground, he can hardly hitting with the bat. Following [6, 2], we represent the estimated human poses with $K_H = 17$ keypoints as a skeleton line-graph. The intensity values range from 0.05 to 0.95 in a uniform interval indicating different human parts. Finally, the binary masks and pose

skeleton map are rescaled to 64x64 and concatenated channel-wise to generate holistic spatial features.

For each branch in the holistic module, we employ two fully connected layers (FC) to embed the features into a low-dimensional space. We denote the output features of human, object, union, spatial branch as f_h, f_o, f_u, f_s . All the features are concatenated (\oplus) to form the final holistic feature $f^{hol} = f_h \oplus f_o \oplus f_u \oplus f_s$.

3.3. Two-Direction Spatial Enhancement

Our goal is to suppress non-interactive human-object pairs. To this end, our idea is not only to encode the rough spatial configurations between the human box and object box, but also to model the detailed interaction patterns between the human parts and object parts. Recent methods [2, 8] extract spatial features from human body parts to the object center. However, we argue that purely encoding the center point is not distinguishable enough when the object is large and surrounding the human, such as "lie in bed" and "drive car". Also for objects that have several separate components, such as "a pair of skis", the center point cannot describe the shape of the object instance. For example, "human ski" shown in Fig. 3, the spatial relationships between human parts to the object center is not distinct enough to differentiate between the interactive pair (in Fig. 3 (1-a)) and non-interactive pair (in Fig. 3 (2-a)), because the object center in both cases are close to the human feet.

Our method enhances the spatial interaction patterns in two directions: from human keypoints to the object center, and from the object parts to the human center. By comparing Fig. 3 (1-b) and (2-b), we can see that the spatial distributions from object parts to the human center can explicitly tell the network that when the two ski boards lie on one side of the standing human, the human is probably not skiing with them.

We first generate object parts from the segmentation mask of the object b_o . The segmentation is a list of polygon vertices around the object as in the COCO format [34]. Given a b_o and its segmentation polygons, we first convert the segmentation from the polygon format to a binary segmentation mask m_o with the same size as b_o . If a pixel i is within the object region, $m_o(i) = 1$, else $m_o(i) = 0$. Then, we use a regular 2D grid size of $R \times R$ to separate the mask m_o into $K_O = R^2$ bins. For each bin, we use

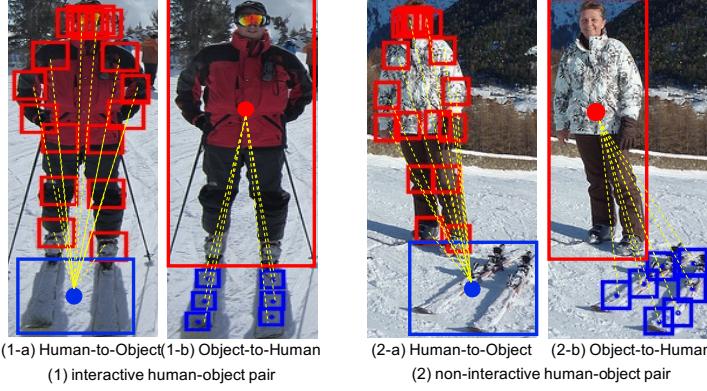


Figure 3: Our proposed symmetric two-direction spatial enhancement can describe the detailed spatial distributions between human body parts and object parts. We show an interactive human-object example in (1). The pair of skis are close to the human feet, and the two ski boards are located on the two sides of the human body. However, in the non-interactive example (2), only using human part to the object center cannot tell whether the human is standing on the ski board or not, because the object center in (2-a) is as close as in (1-a) to the human feet. We additionally encode the relative spatial locations from object parts to the human center shown in (2-b), which can tell that the two ski boards only lie on one side of the human. Hence, (2) is likely to be no-interaction.

the mean point of the mask inside the bin as our object part points o_k . Along with the part points, we generate a flag (length of R^2) for each part indicating the validness of the part. There are two requirements for a valid object part o_k where $flag[o_k] = 1$: (1) the mean point is inside the object region, i.e. $m_o(o_k) = 1$, (2) and the number of object pixels inside the bin is larger than a minimum percentage r of the bin size $\sum_{i \in bin} m_o(i) > rR^2$. We further define a local box region b_{o_k} for each valid part point o_k , which is a box centered at o_k with a size γ proportional to the size of the object proposal b_o . An illustration of object part points and local parts is shown in Fig. 4. Similarly, we generate human local part boxes by defining local boxes $b_{h_1}, \dots, b_{h_{K_H}}$ around all the human keypoints, as shown in Fig. 3 (1-a) and (2-a).

With the generated human and object parts, we can now encode the spatial relationships between them. We first generate an object-center-normalized coordinate map for the human-object branch, and a human-center-normalized coordinate map for the object-human branch. Then, we obtain the spatial offset maps for the human parts and

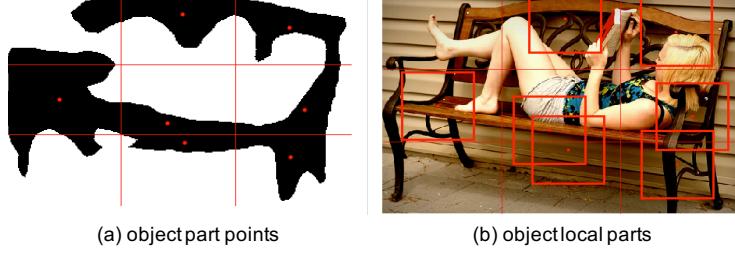


Figure 4: An illustration of our object parts on the bench instance using 3×3 grid ($R = 3$). The red points are valid object part points, and red boxes in (b) are the valid object parts with 0.1 object box size. Better viewed in color.

object parts by cropping part boxes from the two coordinate maps, respectively.

In particular, for each input human-object box pair (b_h, b_o) , supposing the center point location of b_h and b_o is (x_h, y_h) and (x_o, y_o) , respectively, we generate a two-channel coordinate map $\text{Coord} = (\text{Coord}_x, \text{Coord}_y)$ representing the x, y coordinates for each pixel in Φ . Each map has the same spatial dimension $H \times W$ as the convolutional feature map Φ . Then, we compute the relative spatial offset maps in two directions: from the human local parts to the object center $\text{Coord}^{ho} = (\text{Coord}_x^{ho}, \text{Coord}_y^{ho})$, and from the object local parts to the human center $\text{Coord}^{oh} = (\text{Coord}_x^{oh}, \text{Coord}_y^{oh})$.

$$\begin{aligned} \text{Coord}_x^{ho} &= \text{Coord}_x - x_o, & \text{Coord}_y^{ho} &= \text{Coord}_y - y_o, \\ \text{Coord}_x^{oh} &= \text{Coord}_x - x_h, & \text{Coord}_y^{oh} &= \text{Coord}_y - y_h. \end{aligned} \quad (2)$$

Then, in human to object direction, we obtain the spatial offset map $\text{Coord}_{h_k}^{ho}$ for human part h_k and Coord_o^{ho} for object by cropping (ROI-Align [16]) each human part b_{h_k} and object box b_o from the object-center-normalized coordinate map $(\text{Coord}_x^{ho}, \text{Coord}_y^{ho})$, with the same spatial dimension $D_p \times D_p$. In parallel, in object to human direction, we generate spatial offset map $\text{Coord}_{o_k}^{oh}$ for object part o_k and Coord_h^{oh} for human by cropping each object part region b_{o_k} and human box b_h from the human-center-normalized map $(\text{Coord}_x^{oh}, \text{Coord}_y^{oh})$.

We distribute the two direction spatial configurations into two branches. In human-object branch, we adopt ROI-Align for the created human part boxes $b_{h_1}, \dots, b_{h_{K_H}}$

together with object box b_o on the convolution feature map Φ to obtain appearance features $\{f_{h_1}^{app}, \dots, f_{h_{K_H}}^{app}\}$ and $f_{h_o}^{app}$ where each feature is of size $D_p \times D_p \times D$. Then, for each human part h_k , we obtain human part features f_{h_k} by concatenating the part appearance feature $f_{h_k}^{app}$ with human to object spatial feature $Coord_{h_k}^{ho}$. Instead of treating all the human body parts equally, we predict the human body part attention guided by the object appearance. Our idea is that different object usually interact with different human parts. For instance, given a "book", if we want to predict "read book", we need to look at human hands and eyes to decide. Hence, object can infer the importance of different human parts. We predict human part attention using a small network consists of two FC layers with sigmoid function to normalize the attention values to $[0, 1]$. The predicted attention value for part h_k is denoted as α_k . The final feature f_{h_k} for each human part h_k contains both part appearance feature and spatial offsets, as in Eq. 3. Finally, we concatenate all human part features with object feature, and feed it into several FC layers to extract final local feature f^{ho} .

In object-human branch, to enhance spatial feature from object parts to human, we concatenate the spatial offsets for all object parts and feed into several FC layers to extract final local feature f^{oh} . Note, for invalid object parts, we remain both appearance and spatial features to be zero, so that the network parameters will not be trained using those invalid parts.

$$\begin{aligned} f_{h_k} &= \alpha_k \cdot (f_{h_k}^{app} \oplus Coord_{h_k}^{ho}) \\ f_{h_o} &= f_{h_o}^{app} \oplus Coord_o^{ho} \\ f^{ho} &= FC(f_{h_1} \oplus f_{h_2} \dots \oplus f_{h_{K_H}} \oplus f_{h_o}) \\ f^{oh} &= FC(f_{o_1} \oplus f_{o_2} \dots \oplus f_{o_{K_O}} \oplus Coord_h^{oh}) \end{aligned} \tag{3}$$

Finally, we take the holistic features f_{hol} and feed it into the interactiveness classification head (consists of two FC layers followed by sigmoid function σ) to generate the interactiveness score $s_{h,o}^i$. We predict HOI action score $s_{h,o}^a$ from all relation features, including holistic f_{hol} and part-level features f^{ho} and f^{oh} .

$$\begin{aligned} s_{h,o}^i &= \sigma(FC(f^{hol})) \\ s_{h,o}^a &= \sigma(FC(f^{hol} \oplus f^{ho} \oplus f^{oh})) \end{aligned} \tag{4}$$

3.4. Model Training

In the training stage, the training data consist of ground-truth HOI annotations, labeled as $\langle \text{human}, \text{action}, \text{object} \rangle$ triplets, where the humans and objects are localized as bounding boxes. Besides the ground-truth instance bounding boxes, we also augment the HOI training samples using the object proposals given by the detector. We freeze ResNet-50 backbone network, and only train our HOI classification network to predict HOI action categories. Note that the object detector head, pose estimation head, and instance segmentation head in Mask R-CNN [16] are external modules and are not involved in the training process.

Assume for a training sample with a form of human-object pair (b_h, b_o) , the ground-truth action label is an action set $a_{GT} = \{a_1, \dots, a_n\}$, where n is the number of ground-truth actions. We form the target HOI action label vector as $\mathbf{y} = [y^1, \dots, y^A]$, where $y^a = 1$ for $a \in a_{GT}$, and A is the total number of HOI action categories. We also generate the binary interactiveness label $z \in \{0, 1\}$. We define $z = 1$ if $a_{GT} \neq \emptyset$, else $z = 0$.

We adopt holistic features f_{hol} to infer the interactiveness score, and both holistic f_{hol} and part-level features f^{ho} and f^{oh} for HOI action score. Suppose the predicted HOI action score for HOI action category a is $s_{h,o}^a$, and interactiveness score is $s_{h,o}^i$. Our network jointly optimizes HOI action classification loss L_{cls} , and binary interactiveness loss L_{inter} . The HOI classification is a multi-label classification problem, since a human can perform multiple actions to one or multiple target objects simultaneously, e.g. a person can "hit with" and "hold" a baseball bat at the same time. We apply a binary sigmoid classifier for each action category, and then minimize the cross-entropy loss between action score and the ground-truth action label for each action category as:

$$L = \sum_{a=1}^A L_{cls}(y^a, s_{h,o}^a) + \lambda L_{inter}(z, s_{h,o}^i), \quad (5)$$

where both L_{cls} and L_{inter} are cross-entropy loss computed as $L_{CE}(a, b) = a \log(b) + (1 - a) \log(1 - b)$, and λ is the loss weight to balance the contribution of the multi-label interaction prediction and binary interactiveness prediction.

3.5. Exclusive Object Regrouping

Our goal is to suppress the non-interactive pairs to address the mis-grouping problem. Besides enhancing the spatial interaction patterns during training, we also propose a human object regrouping method in the inference stage by considering the exclusion property of the object conditioned on the interested action.

At inference, given an input image with N_h human proposals, N_o object proposals, we can form $N_h \times N_o$ human-object candidate pairs. For each pair (b_h, b_o) , the network predicts an action score $s_{h,o}^a$ for HOI action class a , corresponding to the $\langle \text{human}, \text{action}, \text{object} \rangle$ HOI triplets using Eq. 1. Thus, for each image, we obtain a score matrix S size of $N_h \times N_o \times A$, where each entry of S is $S_{h,o}^a$, and A is the total number of HOI action categories. Then, for each human-action pair (h, a) , we apply max-object selection [3] to select the object box that maximizes $S_{h,o}^a$:

$$b_{o^*} = \underset{b_o}{\operatorname{argmax}} S_{h,o}^a \quad (6)$$

Intuitively, for each human and each action, this max-object selection only remains the max-scoring object as the target object. For example, one human cannot ride more than one bike at the same time. It assumes that there is no cases like human is walking multiple dogs or herding multiple sheep. This max-object selection can reduce the potential triplet number from $N_h \cdot N_o \cdot A$ to $N_h \cdot A$.

We further introduce our novel exclusive object regrouping algorithm by considering the exclusion property of the object. As illustrated in Fig. 5, common sense tells us that one pair of skis cannot be ridden by more than one human at the same time. In order to correct the hard false positive pair between $H2$ and $O1$, our idea is to reselect the next scoring target object for the other non-max humans in the image.

We first obtain the exclusive object prior knowledge from the statistics of the dataset. From the HOI ground-truth annotations, we compute for each action a , how many humans share the same object instance in the image. We compute two statistics in the whole dataset: (1) exclusive object number: the number of ground-truth human-object pairs where one object is only paired with one human, denoted as Q_e^a ; (2) sharable object number: the number of ground-truth pairs where one object is paired with more than one human, denoted as Q_s^a . If for action a , the majority of human-object pairs

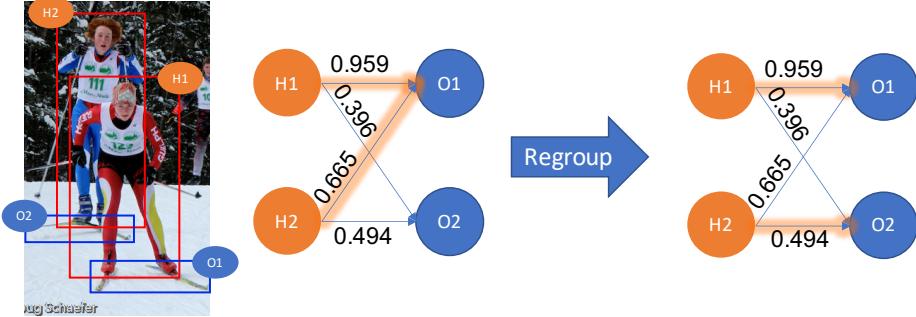


Figure 5: Our proposed exclusive object regrouping approach. The score on the edge is the predicted score of the human-object pair for action "ski". The highlighted edge is the predicted human-object pair for action "ski". Since "ski" is an object-exclusive action, i.e. one pair of skis cannot be shared by more than one human, our method regroup the mis-grouped non-interactive pair between $H2$ and $O1$, and correctly pair the background human $H2$ with her skis $O2$ in the background. Better viewed in color.

are exclusive, defined as $Q_e^a/(Q_e^a + Q_s^a) > \beta$, then the action a is regarded as an object-exclusive action.

With the acquired object-exclusive action list, when predicting an object-exclusive action (e.g. ski, skateboard), if there are multiple humans sharing the same object instance b_{o^*} ($O1$ in Fig. 5), we assign this shared object b_{o^*} to the max-scoring human b_{h^*} ($H1$ in Fig. 5), and reselect the next scoring target object b'_{o^*} ($O2$ in Fig. 5) for the other non-max humans ($H2$ in Fig. 5) in the image in a descending order. Thus, basically, we find the one-to-one matching between human and object instances for object-exclusive actions. To avoid wrong replacement for the non-max humans, we only apply the regrouping algorithm when the prediction score of the max-scoring human-object pair is confident enough, i.e. $S_{b_{h^*}, b_{o^*}}^a > S_{min}$.

4. Experiments

Dataset. We evaluate our method on V-COCO [9] and HICO-DET [10] datasets. **V-COCO dataset** (Verbs in COCO), which is a subset of COCO [34], includes 5,400 train+val images and 4,946 test images. It is annotated with 26 unique verb classes, and has bounding boxes for humans and interacting objects. In particular, three verb classes (i.e. cut, hit, eat) are annotated with two types of targets (i.e. instrument "instr" and

direct object "obj"). **HICO-DET dataset** [10] consists of 47,776 images with more than 150K human-object pairs (38,118 images in training set and 9,658 in test set). It is annotated 600 HOI categories over 80 object categories (same as in COCO [34]) and 117 unique verbs. The bounding boxes of humans and corresponding objects are also annotated.

Evaluation Metrics. We follow the standard evaluation metric as in [9, 3] and report mean Average Precision (mAP) to measure the HOI detection performance. The mAP is computed based on both recall and precision, which is appropriate for the detection task. A HOI triplet $\langle \text{human}, \text{action}, \text{object} \rangle$ is considered as a true positive only when both the human and object detected bounding boxes have IoUs (intersection-over-union) greater than 0.5 with the respective ground-truth and the associated interaction class is correctly classified.

4.1. Implementation Details

We use Mask R-CNN [16] from Detectron [17] with a feature backbone of ResNet-50-FPN [15] to detect object proposals, human keypoints, and the object instance segmentation mask. We implement our network based on ResNet-50-FPN [15] to be comparable to [3, 2]. We crop the ROI features from the highest resolution feature map in FPN with channel dimension $D = 256$. We set the ROI-Align resolution in the holistic module to be $D_h = 7$, and in human-object and object-human part-level modules, we set resolution to be $D_p = 5$. The dimension of features $f_h, f_o, f_u, f_s, f^{oh}$ is set to 256, and f^{ho} is set to 512. The size of the local part boxes for human/object is $\gamma = 0.1$ of human/object box length=max(height, width). We use 3×3 grid (i.e. $R = 3$) to generate object parts, and minimum ratio for a valid bin $r = 1/16$. During testing, the human and object bounding boxes with detection confidence scores above 0.5 and 0.4 respectively are kept.

In VCOCO, out of 26 HOI action categories, we obtain 13 object-exclusive actions from annotation statistics, including "drink-instr", "eat-instr", "hit-instr", "hit-obj", "hold-obj", "jump-instr", "read-obj", "skateboard-instr", "ski-instr", "snowboard-instr", "talk on phone-instr", "throw-obj", "work on computer-instr". In HICO-DET, out of 600 HOI categories, we obtain 372 object-exclusive actions, such as "read/open book",

”hold keyboard/wine glass”, and ”drive car/bus”. We set the minimum confidence score between max human-object pair as $S_{min} = 0.5$, and the statistic ratio of an object-exclusive action as $\beta = 95\%$.

We freeze ResNet-50-FPN backbone, and train the parameters of the HOI classification network. In VCOCO, we use SGD optimizer for training with initial learning rate 4e-2, weight decay 1e-4, and momentum 0.9. We reduce the learning rate to 4e-3 at iteration 24k, and stop training at iteration 48k. In HICO-DET, we use SGD optimizer for training with initial learning rate 5e-2, weight decay 1e-4, and momentum 0.9. We reduce the learning rate to 5e-3 at iteration 30k, and stop training at iteration 90k. The ratio of positive and negative (i.e. non-interactive) samples is 1:3. The weight in the loss function is set as $\lambda = 0.1$.

4.2. Quantitative Results

We first compare our proposed approach with state-of-the-art methods in literature. Tab. 1 shows the comparison on the V-COCO and HICO-DET datasets. On V-COCO, among existing approaches, VRSL [9] uses spatial constraints for the interacting objects. BAR [35] applies box attention based on Faster R-CNN framework. iCAN [1] utilizes human, object appearance and pairwise spatial configuration with contextual attention. The RPNN approach [8] using attention graphs for parsing relations of object and human body-parts obtains a mAP_{role} of 47.5. TIN [6] transfers the interactiveness network trained from multiple datasets to the target dataset to perform non-interaction suppression, and reports a mAP_{role} of 47.8. The recent PMFNet [2] utilizes human poses to extract body part appearance and keypoints to object spatial feature, and achieves mAP_{role} of 52.0. Our method achieves superior performance compared to PMFNet [2] with a mAP_{role} of 52.28, which is the state-of-the-art result.

On HICO-DET, we report the mAP results for the full HOI category with two different settings of Default and Known Objects. We observe that our proposed model achieves competitive results over the state-of-the-art approaches [8, 2, 36]. Our approach obtains mAP of 17.55, and 20.61 in Default and Known Object settings with absolute gains of 0.09 and 0.27 over PMFNet [2], respectively.

Table 1: Results comparisons on V-COCO [9] and HICO-DET [10] datasets. We report mAP_{role} on V-COCO, and mAP on HICO-DET for the full HOI categories in both default setting (object unknown) and Know Object setting. Higher values indicates better performance.

Methods	mAP_{role}	HICO-DET	
		Default	Known Object
VRSL [9] (impl. by [3])	31.8	–	–
Shen et al. [31]	–	6.46	–
HO-RCNN [7]	–	7.81	10.41
InteractNet [3]	40.0	9.94	–
BAR [35]	41.1	–	–
GPNN [30]	44.0	13.11	–
iCAN [1]	45.3	14.84	16.26
Xu et al. [4]	45.9	14.70	–
PMFNet-Base [2]	48.6	14.92	18.83
DCA [5]	47.3	16.24	17.73
RPNN [8]	47.5	17.35	–
No-Frills [36]	–	17.18	–
TIN ($RP_D C_D$) [6]	47.8	17.03	19.17
PMFNet [2]	52.0	17.46	20.34
Ours	52.28	17.55	20.61

4.3. Ablation Study

We perform several experiments to evaluate the effectiveness of our model components on V-COCO dataset (Tab. 2). The baseline model consists of holistic module and human-object appearance feature, and achieves 51.67 mAP. Adding human-object spatial and object-human spatial improves absolute gain of 0.21 and 0.29, respectively. We further add both two-direction spatial branch, and achieves 52.18 mAP with 0.22 gain over baseline. Adding exclusive-prior regrouping, we obtain 52.16 mAP with 0.2 absolute gain compared to without regrouping (51.96). Using both our proposed spatial

Table 2: Ablation study of each component in our network on V-COCO dataset.

Add-on	baseline					
holistic	✓	✓	✓	✓	✓	✓
human-object appearance	✓	✓	✓	✓	✓	✓
human-object spatial			✓	✓	✓	✓
object-human spatial		✓		✓		✓
exclusive-prior					✓	✓
mAP _{role}	51.67	51.88	51.96	52.18	52.16	52.28

enhancement and exclusive prior regrouping, our full model achieves 52.28 mAP with 0.61 gain over baseline. These results demonstrate that our enhanced two-direction spatial configurations enables the network to suppress non-interactive human-object instances more effectively. By further considering the exclusive property of the object, our method can correct the mis-grouped pairs and associate the human to its associate target object.

4.4. Qualitative Visualizations

We show our human-object interaction detection results on VCOCO in Fig. 6. Each subplot shows one detected $\langle \text{human}, \text{action}, \text{object} \rangle$ triplet. The first row in Fig. 6 shows our correctly detected triplets of one human taking multiple actions on multiple objects. When multiple humans appear in the scene, our method can correctly pair the interactive human-object pairs and recognize the action between them (Fig. 6 row 3).

Fig. 7 shows our detection results on HICO-DET. It shows that our model can predict HOIs in a wide variety of different situations in daily life.

Fig. 8 shows the effectiveness of our proposed exclusive object regrouping method on VCOCO. Due to occlusion and spatial affinity between humans and objects, purely relying on the network prediction can mis-group non-interactive human and object with a high confidence, resulting in hard False Positive (FP) detections (see Fig. 8 (b, e)). Our method can correctly associate the human with its true target object at inference



Figure 6: HOI detection results of our method on VCOCO. A human (red box) can perform multiple actions to multiple objects (blue box). Our method can correctly pair the human with the corresponding target object when multiple humans in the scene. Better viewed in color.



Figure 7: HOI detections results of our method on the HICO-DET. Our model can predict HOIs in a wide variety of different situations in daily life. Better viewed in color.



Figure 8: HOI detection results of our full model with and without exclusive object regrouping at inference on VCOCO. Our method can remove the mis-grouped false positive human-object pairs shown in column (b, e). In particular, on the left (1), we show three examples that multiple humans are performing the same actions in the scene. Our method can correctly associate the human (i.e. H2) with his true target object (see column (c)). One the right (2), we show three examples that multiple humans are performing different actions in the scene. In column (e), the human in red box is not performing the interested action, but can be mis-grouped with an object. For instance, in row 2, column (f), the child is not hitting anything. Our method successfully suppresses the non-interactive pairs with low confidence scores when the human is not performing the interested action (see column (f)). We can see that our proposed method can alleviate the mis-grouping issue in HOI detection. Better viewed in color.

(Fig. 8 (c)), and suppress non-interactive pairs when the human is not performing the action interested (Fig. 8 (f)).

We further show more detection results under the presence of multiple humans in the scene on HICO-DET in Fig. 9. We can see that our proposed method can correctly pair the interactive human-object pairs under the presence of multiple humans in the scene, even when they are performing similar actions with occlusions.

5. Conclusion

In this paper, we address the problem of mis-grouping in human-object interaction detection task, where non-interactive pairs are wrongly paired and classified as an action. Our goal is to suppress non-interactive pairs, and hence can reduce false positive HOI detections. We propose two-direction spatial constraints between human body

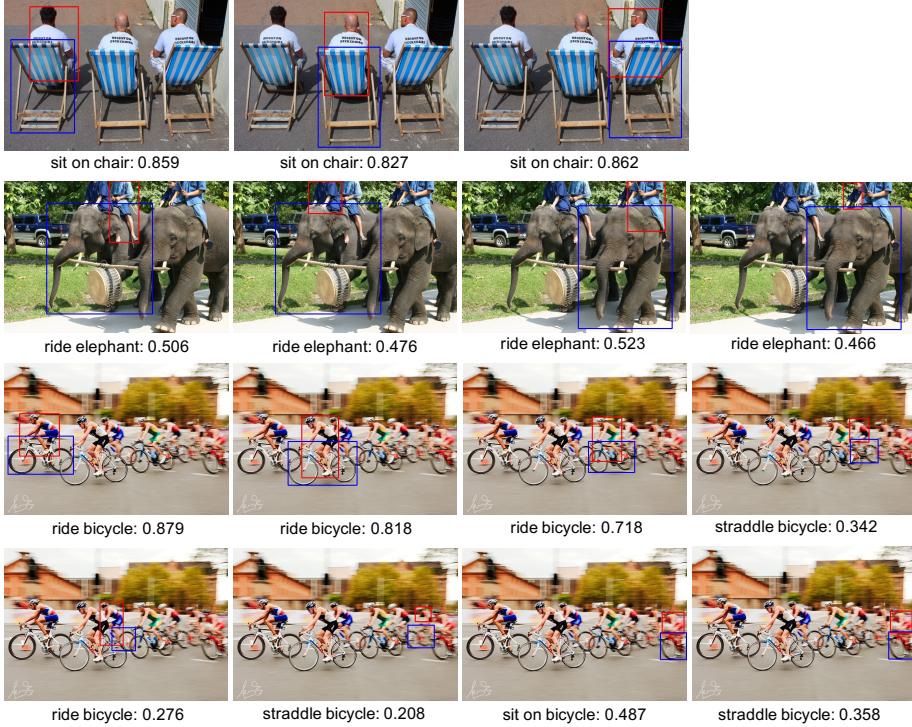


Figure 9: HOI detections results of our method on HICO-DET under the presence of multiple humans in the scene. We can see that our proposed method can correctly pair the interactive human-object pairs, even when they are performing similar actions with occlusions. Better viewed in color.

parts and object parts. To further suppress the non-interactive pairs, at inference, we introduce an exclusive object regrouping approach by considering the exclusion property of the object conditioned on the interested action. Our approach is able to correct mis-grouped hard false positive human-object pairs when multiple humans performing similar actions in the scene, and suppress the non-interactive human-object pairs even when the target the object is close to the human. Experiments on V-COCO and HICO-DET datasets demonstrate our approach is more robust compared to the existing methods under the presence of multiple humans and objects in the scene.

References

- [1] C. Gao, Y. Zou, J.-B. Huang, ican: Instance-centric attention network for human-object interaction detection, arXiv preprint arXiv:1808.10437.
- [2] B. Wan, D. Zhou, Y. Liu, R. Li, X. He, Pose-aware multi-level feature network for human object interaction detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9469–9478.
- [3] G. Gkioxari, R. Girshick, P. Dollár, K. He, Detecting and recognizing human-object interactions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359–8367.
- [4] B. Xu, Y. Wong, J. Li, Q. Zhao, M. S. Kankanhalli, Learning to detect human-object interactions with knowledge, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
- [5] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, J. Laaksonen, Deep contextual attention for human-object interaction detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5694–5702.
- [6] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, C. Lu, Transferable interactivity knowledge for human-object interaction detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3585–3594.
- [7] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng, Learning to detect human-object interactions, in: 2018 ieee winter conference on applications of computer vision (wacv), IEEE, 2018, pp. 381–389.
- [8] P. Zhou, M. Chi, Relation parsing neural network for human-object interaction detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 843–851.
- [9] S. Gupta, J. Malik, Visual semantic role labeling, arXiv preprint arXiv:1505.04474.

- [10] Y.-W. Chao, Z. Wang, Y. He, J. Wang, J. Deng, Hico: A benchmark for recognizing human-object interactions in images, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1017–1025.
- [11] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
- [12] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, 2015, pp. 91–99.
- [13] J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks, in: Advances in neural information processing systems, 2016, pp. 379–387.
- [14] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
- [15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
- [16] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
- [17] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, <https://github.com/facebookresearch/detectron2> (2019).
- [18] G. Gkioxari, R. Girshick, J. Malik, Actions and attributes from wholes and parts, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2470–2478.
- [19] Z. Zhao, H. Ma, S. You, Single image action recognition using semantic body part actions, in: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 3411–3419.

- [20] L. Liu, R. T. Tan, S. You, Loss guided activation for action recognition in still images, in: C. Jawahar, H. Li, G. Mori, K. Schindler (Eds.), Computer Vision – ACCV 2018, Springer International Publishing, Cham, 2019, pp. 152–167.
- [21] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 4489–4497.
- [22] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.
- [23] R. Girdhar, D. Ramanan, Attentional pooling for action recognition, in: NIPS, 2017.
- [24] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r* cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1080–1088.
- [25] C. Lu, R. Krishna, M. Bernstein, L. Fei-Fei, Visual relationship detection with language priors, in: European conference on computer vision, Springer, 2016, pp. 852–869.
- [26] H. Zhang, Z. Kyaw, S.-F. Chang, T.-S. Chua, Visual translation embedding network for visual relation detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5532–5540.
- [27] Y. Zhan, J. Yu, T. Yu, D. Tao, On exploring undetermined relationships for visual relationship detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5128–5137.
- [28] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, M. Elhoseiny, Large-scale visual relationship understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 9185–9194.

- [29] J. Peyre, I. Laptev, C. Schmid, J. Sivic, Detecting unseen visual relations using analogies, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1981–1990.
- [30] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, Learning human-object interactions by graph parsing neural networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 401–417.
- [31] L. Shen, S. Yeung, J. Hoffman, G. Mori, L. Fei-Fei, Scaling human-object interaction recognition through zero-shot learning, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1568–1576.
- [32] K. Kato, Y. Li, A. Gupta, Compositional learning for human object interaction, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 234–251.
- [33] H.-S. Fang, J. Cao, Y.-W. Tai, C. Lu, Pairwise body-part attention for recognizing human-object interactions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 51–67.
- [34] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
- [35] A. Kolesnikov, A. Kuznetsova, C. Lampert, V. Ferrari, Detecting visual relationships using box attention, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.
- [36] T. Gupta, A. Schwing, D. Hoiem, No-frills human-object interaction detection: Factorization, layout encodings, and training techniques, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9677–9685.