Learning Human-Object Interaction Detection using Interaction Points

Tiancai Wang[1]* Tong Yang[1] Martin Danelljan[2] Fahad Shahbaz Khan[3,4]
Xiangyu Zhang[1] Jian Sun[1]
[1]MEGVII Technology [2]ETH Zurich, Switzerland [3]IIAI, UAE [4]Linköping University, Sweden

Abstract

Understanding interactions between humans and objects is one of the fundamental problems in visual classification and an essential step towards detailed scene understanding. Human-object interaction (HOI) detection strives to localize both the human and an object as well as the identification of complex interactions between them. Most existing HOI detection approaches are instance-centric where interactions between all possible human-object pairs are predicted based on appearance features and coarse spatial information. We argue that appearance features alone are insufficient to capture complex human-object interactions. In this paper, we therefore propose a novel fully-convolutional approach that directly detects the interactions between human-object pairs. Our network predicts interaction points*, which directly localize and classify the interaction. Paired with the densely predicted* interaction vectors*, the interactions are associated with human and object detections to obtain final predictions. To the best of our knowledge, we are the first to propose an approach where HOI detection is posed as a keypoint detection and grouping problem. Experiments are performed on two popular benchmarks: V-COCO and HICO-DET. Our approach sets a new state-of-the-art on both datasets. Code is available at* https://github.com/vaesl/IP-Net.

1. Introduction

Detailed semantic understanding of image contents, beyond instance-level recognition, is one of the fundamental problems in computer vision. Detecting human-object interaction (HOI) is a class of visual relationship detection where the task is to not only localize both a human and an object but also infer the relationship between them, such as "eating an apple" or "driving a car". The problem is challenging since an image may contain multiple humans performing the same interaction, same human simultaneously interacting with multiple objects ("sit on a couch and type on laptop"), multiple humans sharing the same interaction and object ("throw and catch ball"), or fine-grained interactions ("walk horse", "feed horse" and "jump horse"). These complex and diverse interaction scenarios impose significant challenges when designing an HOI detection solution.

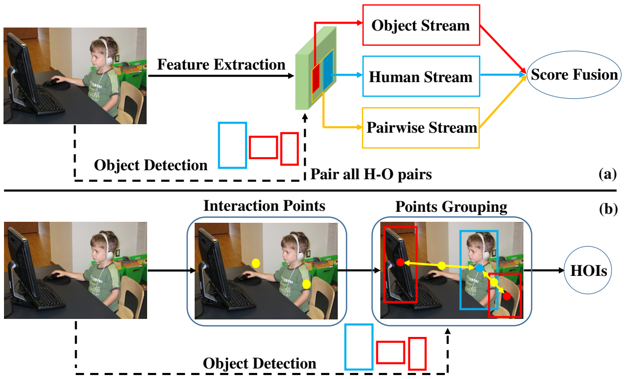


Figure 1. (a) Most existing approaches address the HOI detection problem where the detected bounding-boxes (human and object) from a pre-trained detector are first used to extract region-of-interest (RoI) features from the backbone. Then, a multi-stream architecture is employed where the individual scores from three parallel streams: human, object and pairwise are fused to obtain final interaction predictions for all human-object pairs. (b) Different from previous methods, our proposed approach poses HOI as a keypoint detection and grouping problem by learning to generate interaction points and vectors which are directly grouped along with human and object instances from the object detection branch.

Most existing approaches [3, 7, 14, 39] detect human-object interactions in the form of triplets $\langle$*human, action, object*$\rangle$ by decomposing the problem into two parts: object detection and interaction recognition. For object detection, a pre-trained object detector is typically employed to detect both humans and objects. For interaction recognition, several strategies exist in literature [23, 25, 34]. Most of the recent HOI detection approaches [3, 7, 14, 34] utilize a multi-stream architecture (see Fig. 1(a)) for interaction recognition. The multi-stream architecture typically contains three individual streams: a human, an object, and a pairwise. Both human and object streams encode appear-

*This work was performed during an internship at MEGVII Technology. This work was supported by The National Key Research and Development Program of China (No. 2017YFA0700800) and Beijing Academy of Artificial Intelligence (BAAI). Contact email: wangtc@tju.edu.cn

ance features of human and objects, respectively whereas the pairwise stream aims at encoding the spatial relationship between the human and object. Individual scores from the three streams are then fused in a late fusion fashion for interaction recognition.

While improving the HOI detection performance, state-of-the-art approaches based on the above mentioned multi-stream architecture are computationally expensive. During training, these instance-centric approaches require pairing all humans with all objects in order to learn both positive and negative human-object pairs. This implies that the inference time scales *quadratically* with the number of instances in the scene, since all human-object pairs are required to be passed through the network in order to obtain the final interaction scores. In addition to being computationally expensive, these approaches predominantly rely on appearance features and a simple pairwise stream that takes the union of the two boxes (human and object) to construct a binary image representation. We argue that this reliance on appearance features alone and coarse spatial information is insufficient to capture complex interactions, leading to inaccurate predictions. In this work, we look into an alternative approach that addresses these shortcomings by directly detecting the interactions between human-object pairs as a set of interaction points.

Contributions: In this work, we propose a novel approach for HOI detection. Motivated by the recent success of anchor-free object detection methods, we pose HOI detection as a keypoint detection and grouping problem (see Fig. 1(b)). The proposed approach directly detects interactions between human-object pairs as a set of interaction points. Based on the interaction point, our method learns to generate an interaction vector with respect to the human and object center points. We further introduce an interaction grouping scheme that pairs the interaction point and vector with the corresponding human and object bounding-box predictions, from the detection branch, to produce final interaction predictions. Extensive experiments are conducted on two HOI detection benchmarks: V-COCO [9] and HICO-DET [4] datasets. Our proposed architecture achieves state-of-the-art results on both two datasets, outperforming existing instance-centric methods by a significant margin. Additionally, we perform a thorough ablation study to demonstrate the effectiveness of our approach.

2. Related Work

Object Detection: In recent years, significant progress has been made in the field of object detection [15, 17, 19, 21, 28, 29, 35, 36], mainly due to the advances in deep convolutional neural networks (CNNs). Generally, modern object detection approaches can be divided into single-stage [17, 20, 26, 27, 33] and two-stage methods [1, 15, 28]. Two-stage object detection methods typically generate candidate object proposals and then perform classification and regression of these proposals in the second stage. On the other hand, single-stage object detection approaches work by directly classifying and regressing the default anchor box in each position. Two-stage object detectors are generally known to be more accurate whereas the main advantage of single-stage methods is their speed.

Within object detection, recent anchor-free single-stage detectors [13, 31, 40, 41] aim at eliminating the requirement of anchor boxes and treat object detection as keypoint estimation. CornerNet [13] detects the bounding-box of an object as a pair of keypoints, the top-left corner and the bottom-right corner. ExtremeNet [41] further detects four extreme points and one center point of objects and groups the five keypoints into a bounding-box. CenterNet [40] models an object as a single point — the center point of its bounding-box and is also extended to Human pose estimation [6] and 3D detection task [24].

Human-Object Interaction Detection Among existing human-object interaction (HOI) detection methods, the work of [9] is the first to explore the problem of visual semantic role labeling. The objective of this problem is to localize the agent (human) and object along with detecting the interaction between them. The work of [8] introduces a human-centric approach, called InteractNet, which extends the Faster R-CNN framework with an additional branch to learn the interaction-specific density map over target locations. Qi *et al.*, [25] proposes to utilize graph convolution neural network and regards the HOI task as a graph structure optimization problem. Chao *et al.*, [3] builds a multi-stream network that is based on the human-object region-of-interest and the pairwise interaction branch. The inputs to this multi-stream architecture are the predicted bounding-boxes from the pre-trained detector (*e.g.*, FPN [15]) and the original image. Human and object streams in such a multi-stream architecture are based on appearance features, extracted from the backbone network, to generate confidence predictions on the detected human and object bounding-boxes. The pairwise stream, on the other hand, simply encodes the spatial relationship between the human and object by taking the union of the two boxes (human and object). Later works have extended the above mentioned multi-stream architecture by, *e.g.*, introducing instance-centric attention [7], pose information [14] and deep contextual attention based on context-aware appearance features [34].

3. Method

Here, we present our approach based on interaction point generation (Sec. 3.3) and grouping (Sec. 3.4).

3.1. Motivation

As discussed earlier, most existing HOI detection approaches [3, 7, 34] adopt a multi-stream architecture where

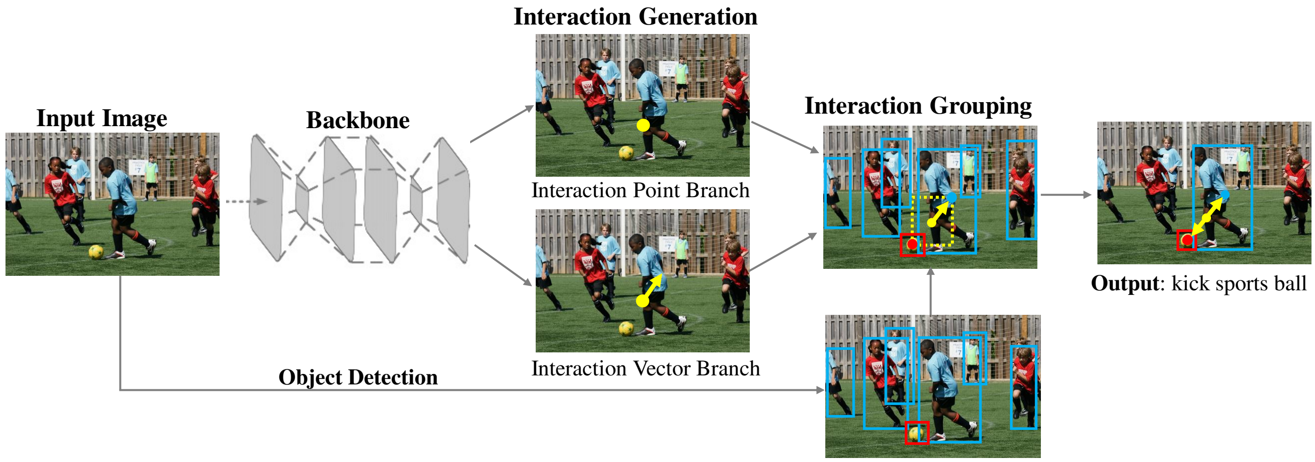


Figure 2. Overall architecture of the proposed HOI detection framework having a localization and an interaction prediction stage. As in several previous works [7, 14, 34], we adopt a standard object detector (FPN [15]) to obtain human and object bounding-box predictions. Our interaction prediction stage consists of three steps: feature extraction, interaction generation (Sec. 3.3) and interaction grouping (Sec. 3.4). The interaction generation contains two independent branches to produce interaction point and interaction vector, respectively. Interaction point and vector together with the detected human and object bounding-box predictions are then input to the interaction grouping for final HOI predictions: ⟨*human, action, object*⟩.

individual scores from a human, an object, and a pairwise stream are fused in a late fusion manner for interaction recognition. We argue that such a late fusion strategy is sub-optimal since appearance features alone are insufficient to capture complex human-object interactions. Further, the pairwise stream simply takes the union of the two boxes (human and object) as the reference box to construct a binary image representation which may lead to inaccurate predictions due to the coarse spatial information.

Motivated by the advances in anchor-free object detection [13, 40, 41], we regard HOI detection as interaction point estimation problem by defining the interaction between the human and an object as an interaction point. Based on the interaction point, our method also learns to produce an interaction vector with respect to the human and object center points. It then pairs the interaction points with the corresponding human and object bounding-box predictions. Different from object detection where object instances are generally independent to each other in an image, interaction point estimation in HOI is more challenging due to diverse and complex real-world interaction scenarios, *e.g.*, multiple humans performing the same interaction or same human simultaneously interacting with multiple objects. To the best of our knowledge, we are the first to propose a HOI detection approach where interaction between the human and an object is defined as a keypoint.

3.2. Overall Architecture

Our overall architecture is shown in Fig. 2. It consists of object detection and interaction prediction. For object detection, we follow previous HOI detection works [7, 34] and employ a standard object detector, FPN [15], for generating bounding-boxes for all possible human and object instances in an image. The main focus of our design is a new representation for *interaction prediction*. It comprises three steps: feature extraction, interaction generation (Sec. 3.3) and interaction grouping (Sec. 3.4). For feature extraction, we employ the Hourglass [18] as the network backbone typically used in anchor-free single stage methods [13, 40, 41]. Given an input RGB image with size $H \times W \times 3$, the output of the Hourglass network is a feature map with size $\frac{H}{S} \times \frac{W}{S} \times D$, where H, W are the height, width of the input image and D, S are the output channels and stride, respectively. As in [2, 22], we adopt a stride of $S = 4$ to achieve a trade-off between accurate localization and computational efficiency. The resulting features from the backbone are input to the interaction generation module to produce interaction point and interaction vector. Interaction point is defined as the center point of the action between a human-object pair and is the starting point of the interaction vector. Consequently, the interaction point and vector together with the detected human and object bounding-boxes are input to the interaction grouping step for the final HOI triplet ⟨*human, action, object*⟩ prediction.

3.3. Interaction Generation

The interaction generation module contains two parallel branches: interaction point and interaction vector prediction. Both branches take the features extracted from the backbone as an input.

Interaction Point Branch: Given the feature maps generated from the backbone network, a single 3×3 convolution layer is employed to produce the interaction point heatmaps of size $\frac{H}{S} \times \frac{W}{S} \times C$, where C denotes the number of in-

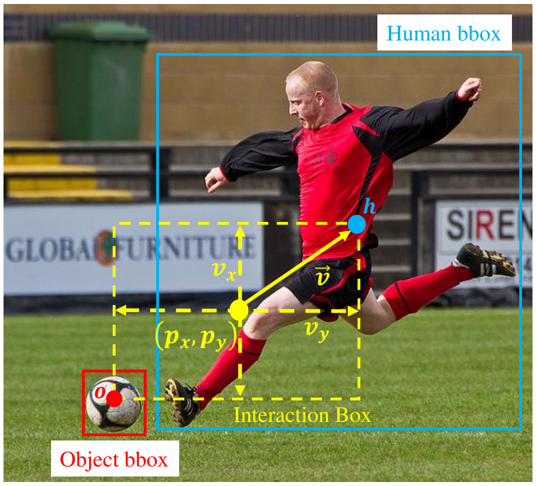


Figure 3. Illustration of interaction point and interaction vector on an example image. The interaction point p (yellow circle) is defined as the center point of the action between human-object pair and itself serves as the starting point of the interaction vector v (yellow arrow). During training, interaction point heatmaps are supervised by ground-truth Gaussian heatmaps generated from human and object center points (cyan and red circles). The raw-pixel coordinates of interaction points on the scale maps are supervised by horizontal and vertical lengths of the interaction vector using L1 loss. At inference, the generated interaction point heatmaps are used to extract top-k peak interaction points by employing a post-processing strategy, as in [13]. Based on the location of top-k interaction points, horizontal and vertical length of the interaction vector is obtained at corresponding coordinates of the scale maps.

teraction categories. During training, the interaction point heatmaps are supervised by the ground-truth heatmaps with multiple peaks, where each interaction point is defined with the same Gaussian Kernel, as in [13]. We empirically fix the standard deviation in the Gaussian Kernel throughout our experiments. Note that a single keypoint location can only represent one object class in single-stage object detection [40]. Different from object detection, a single keypoint location may refer to multiple interaction categories in HOI detection since the human can have multiple interactions with a given object. For instance, the human may hold and hit with tennis racket at the same time. In such a case, both the 'hit' and 'hold' interactions are located at the same position on the heatmap but are represented by different channels. Fig. 3 shows an example interaction point (yellow circle), defined as $p_x = \frac{h_x+o_x}{2}, p_y = \frac{h_y+o_y}{2}$, for a given human-object (HO) pair having center points $h = (h_x, h_y)$ and $o = (o_x, o_y)$, respectively. Note that interaction points are generated for categories involving both the human and an object. For interaction categories without any associated object (*e.g.*, walk and run), the interaction point generalizes to the center point of the corresponding human. Most categories in standard HOI detection datasets [4, 9] involve both the human and an object.

Interaction Vector Branch: As shown in Fig. 3, based on the interaction point (p_x, p_y), the interaction vector branch aims to predict the interaction vector towards the corresponding human center point. Given the paired human and object bounding-boxes, human center point h, and object center point o, the interaction point $p = (p_x, p_y)$ is calculated. Then, the interaction vector $v = (v_x, v_y)$ is defined such that $p + v = h$ and $p - v = o$.

The interaction vector branch is trained to predict the value of the unsigned interaction vector $v' = (|v_x|, |v_y|)$, which is used as the ground-truth in our training. As in the interaction point branch, we employ a single 3×3 convolution layer to produce the unsigned interaction vector map V of size $\frac{H}{S} \times \frac{W}{S} \times 2$, where one is for the length of interaction vector in horizontal direction and the other is for the length of interaction vector in vertical direction. At inference, we extract four possible locations of the human center based on the interaction point and the unsigned interaction vector as,

$$(x_h^i, y_h^i) = (p_x \pm |v_x|, p_y \pm |v_y|)\,,\ i = 1, 2, 3, 4\,. \qquad (1)$$

We further define the *interaction box* as the rectangle with corners given by (1). Next, we describe the interaction grouping scheme.

3.4. Interaction Grouping

During training, the interaction point and its corresponding human and object center points have a fixed geometric structure. During the inference stage, the generated interaction points need to be grouped with the object detection results (human and object bounding-boxes). This implies that the generated interaction point p is paired with the human having center h and object having center o, if the following condition is satisfied: $h \approx p + v$ and $o \approx p - v$.

For efficient and accurate grouping of interaction points with human and object bounding-boxes, we further propose an interaction grouping scheme which utilizes soft constraints to filter out bulk of the negative HOI pairs. Fig. 4 shows an illustration of our interaction grouping scheme. It has three inputs: human/object bounding-box (cyan and red), interaction point (orange point) extracted from interaction heatmaps, and the interaction vector (orange arrow) at the location of interaction point. The four corners (in green) of the interaction box (in orange) are calculated by the given interaction point and the unsigned interaction vector, using Eq. (1). The four corners of the *reference box* r_{box} (in purple) can be determined by the center points of the detected human and object bounding-boxes. Then, based on the generated interaction and reference boxes, we compute the vector lengths, $d_{\text{tl}}, d_{\text{tr}}, d_{\text{bl}}, d_{\text{br}}$, for four corners of these two boxes. In case the interaction box and the four vector lengths satisfy the constraints in (2) below, then the current human and object bounding-boxes and the interaction point are regarded as the true positive HOI pair.

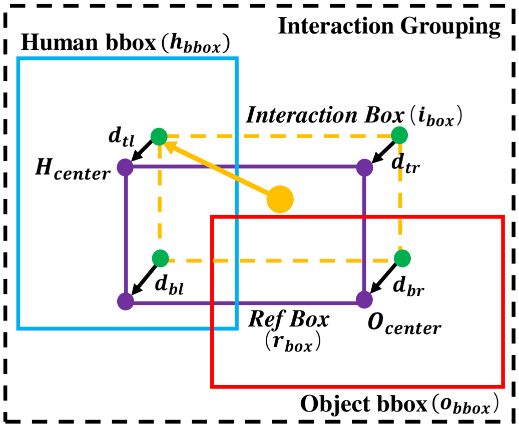


Figure 4. The procedure of interaction grouping scheme. It has three inputs: the human/object bounding-boxes from object detection branch, the interaction points from the interaction point branch and the interaction vector predicted by the interaction vector branch. The interaction box (orange box) is determined by the given interaction point and the lengths (horizontal and vertical) of the interaction vector. The reference box (purple) can also determined by the detected human/object boxes. In case the reference box, the interaction box and human/object bounding-boxes satisfy the conditions (2), then the current human/object bounding-boxes and the interaction point are regarded as a true positive HO pair.

$$\begin{cases} \mathrm{IoU}(h_{\text{bbox}}, i_{\text{box}}) > 0, \\ \mathrm{IoU}(o_{\text{bbox}}, i_{\text{box}}) > 0, \\ d_{\text{tl}}, d_{\text{tr}}, d_{\text{bl}}, d_{\text{br}} < d_\tau \end{cases} \tag{2}$$

Here, h_{bbox} and o_{bbox} are the given human and object bounding-boxes from object detection branch. i_{box} is the interaction box generated by the interaction point and the interaction vector. d_{tl}, d_{br}, d_{tr} and d_{bl} are four vector lengths of the four corners between interaction box i_{box} and the reference box r_{box}. d_τ is the vector length threshold set for filtering the negative HOI pairs. Interaction grouping scheme is presented in Algorithm 1.

Algorithm 1 Interaction Grouping

```
Input:
  Human/object bboxes from object detector: $H_{\text{bbox}}$, $O_{\text{bbox}}$
  Interaction point and vector heatmaps: $P$, $V$
  Human, object and action score thresholds:$h_\tau, o_\tau, a_\tau$
  Vector length threshold for corners: $d_\tau$
Output: HOI triplets with final score.
  // Convert heatmaps $P$ into an interaction point set $A$.
  // Extract the interaction vectors from $V$.
  for $h_{\text{box}} \in H_{\text{bbox}}, o_{\text{box}} \in O_{\text{bbox}}, a \in A$ do
    if $h_{\text{score}} > h_\tau, o_{\text{score}} > o_\tau, a_{\text{score}} > a_\tau$ then
      // Obtain interaction box $i_{\text{box}}$ using Ed. 1.
      // Calculate reference box $r_{\text{box}}$ by $h_{\text{box}}$ and $o_{\text{box}}$.
      if $h_{\text{bbox}}, o_{\text{bbox}}, i_{\text{box}}, r_{\text{box}}$ satisfy Condition (2) then
        $s_{\text{f}} \leftarrow h_{\text{score}} \cdot o_{\text{score}} \cdot p_{\text{score}}$
        // Output the current HOI pair with final score $s_{\text{f}}$ .
      end if
    end if
  end for
```

3.5. Model Learning

For the predicted interaction point heatmap P, the ground-truth heatmaps $\hat{P}$ are with all interaction points, and each of them is defined as the Gaussian Kernel. We follow the modified focal loss originally proposed in [13] to balance the positive and negative samples,

$$L_p = \frac{-1}{N_p} \begin{cases} (1-P_{xyc})^\alpha \log(P_{xyc}), & \text{if } \hat{P}_{xyc} = 1 \\ (1-\hat{P}_{xyc})^\beta (P_{xyc})^\alpha \log(1-P_{xyc}), & \text{o.w.} \end{cases} \tag{3}$$

where N_p is the number of interaction points in the image. α and β are the hyper-parameters to control the contribution of each point (we set α to 2 and β to 4, as in [13]). For the predicted interaction vector maps V, we use the value of the unsigned interaction vector $v'_k = (|v_x|_k, |v_y|_k)$ at the interaction point p_k as the ground-truth. Then, L1 loss is employed for all the interaction points,

$$L_v = \frac{1}{N} \sum_{k=1}^{N} |V_{p_k} - v'_k| \, . \tag{4}$$

Here V_{p_k} denotes the predicted interaction vectors at point p_k. The overall loss function is summarized as,

$$L_{tot} = L_p + \lambda_v L_v \, , \tag{5}$$

where λ_v is the weight for the vector loss term. Here we simply set $\lambda_v = 0.1$ for all our experiments.

4. Experiments

4.1. Datasets and Metrics

Datasets: We conduct comprehensive experiments on two challenging HOI datasets: V-COCO [9] and HICO-DET [4]. The V-COCO dataset contains 2533, 2867, and 4946 images for training, validation and testing, respectively. Typically, the combined training and validation sets (5400 images in total) are used for model training. Human instances in V-COCO dataset has 26 binary action labels and three action categories (cut, hit, eat) are annotated with two types of targets (i.e., instrument and direct object). Note that three classes (run, stand, walk) are annotated with no interaction object. The HICO-DET dataset contains 38,118 images for training and 9658 images for testing. In this dataset, each human instance is annotated with 600 classes of different interactions, corresponding to 80 object categories and 117 action verbs. Note that those 117 action verbs include the 'no interaction' class.

Metrics: We follow the standard evaluation protocols, as in [4, 9], to evaluate our proposed method. The results are reported in terms of role mean Average Precision (mAP_{role}). In mAP_{role}, one HOI triplet is regarded 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.

4.2. Implementation Details

For interaction prediction, we use Hourglass-104 [18] as feature extractor, pre-trained on MS COCO (train2017 set), as in [40]. The head network for the interaction point and interaction vector generation is randomly initialized. During training, we adopt an input resolution of 512×512. This yields an output resolution of 128×128 for the Hourglass backbone. We employ standard data augmentation techniques (random flip, random scaling (between 0.6 to 1.3), cropping, and color jittering) and use Adam optimizer [11] to optimize the loss function during training. During test, we use flip augmentation to obtain final predictions. Following [40], we use batch-size of 29 (on 5 GPUs, with master GPU batch-size 4) and learning rate $2.5 \cdot 10^{-4}$ for 50 epochs with $10\times$ learning rate drop at 40 epoch. For detection branch, we follow previous HOI detection methods [7, 14, 34] and utilize Faster-RCNN [28] with ResNet-50-FPN [15] pre-trained on COCO [16] train2017 split. To obtain bounding-boxes at inference, we set score threshold greater than 0.4 for humans and 0.1 for objects. These score thresholds are relatively lower than the thresholds set in [7, 34], since the interaction box generated by our interaction point and vector can filter out most negative pairs. For interaction generation, Hourglass-104 takes about 77ms. Our interaction grouping has a complexity of $\mathcal{O}(N_\text{h} N_\text{o} N_\text{ip})$, where N_h, N_o, N_ip is the number of humans, objects and interaction points, respectively. In practice, our grouping is efficient, taking less than 5 ms ($< 6.1\%$ of total time).

4.3. State-of-the-art Comparison:

We first compare our proposed approach with state-of-the-art methods in literature. Tab. 1 shows the comparison on the V-COCO dataset. Among existing approaches, BAR [12], iCAN [7] and DCA [34] utilize human and object appearance features in a multi-stream architecture. The DCA method [34] consisting of a deep contextual attention module that generates contextually-aware appearance features within a multi-stream architecture achieves a mAP_{role} of 47.3. The RPNN approach [38] based on attention graphs for parsing relations of object and human body-parts obtains a mAP_{role} of 47.3. The work of [14], denoted in Tab. 1 as TIK, introduces an interactiveness network to perform Non-interaction Suppression and reports a mAP_{role} of 47.8. Our approach achieves superior performance compared to existing methods with a mAP_{role} of 51.0. The results are further improved (mAP_{role} of 52.3) by first pre-training our network on HICO-DET and then fine-tuning the pre-trained HICO-DET model on the V-COCO dataset.

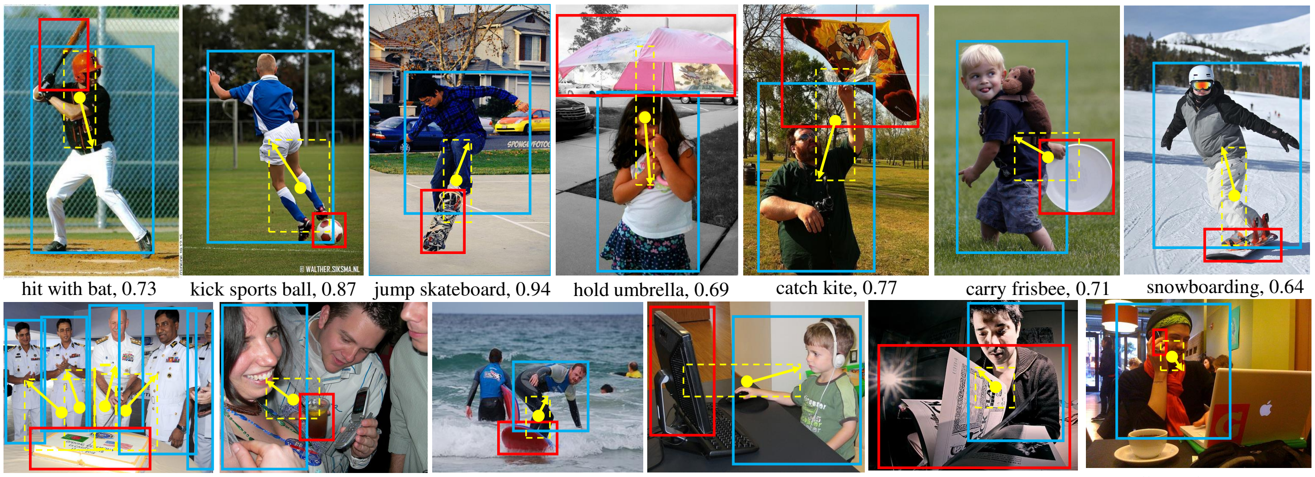
Methods	mAP_{role}
VSRL[9]*	31.8
InteractNet [8]	40.0
BAR [12]	41.1
GPNN [25]	44.0
iCAN [7]	45.3
HOI w knowledge [37]	45.9
DCA [34]	47.3
RPNN [38]	47.5
TIK [14]	47.8
PMFNet [32]	52.0
Ours	**51.0**
Ours + HICO	**52.3**

Table 1. State-of-the-art comparison (in terms of mAP_{role}) on the V-COCO dataset. * refers to implementation of [9] by [8]. Our approach sets a new state-of-the-art with mAP_{role} of 51.0 and achieves an absolute gain of 3.2% over TIK [14]. The results are further improved (mAP_{role} of 52.3) when utilizing pre-training on HICO-DET and then fine-tuning on V-COCO dataset.

	Default			Known Object		
Methods	full	rare	non-rare	full	rare	non-rare
Shen *et al.*, [30]	6.46	4.24	7.12	-	-	-
Chao *et al.*, [3]	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [8]	9.94	7.16	10.77	-	-	-
GPNN [25]	13.11	9.34	14.23	-	-	-
Xu et.al [37]	14.70	13.26	15.13	-	-	-
iCAN [7]	14.84	10.45	16.15	16.43	12.01	17.75
DCA [34]	16.24	11.16	17.75	17.73	12.78	19.21
TIK [14] [14]	17.03	13.42	18.11	19.17	15.51	20.26
Gupta et.al [10]	17.18	12.17	18.68	-	-	-
RPNN [38]	17.35	12.78	18.71	-	-	-
PMFNet [32]	17.46	**15.65**	18.00	20.34	**17.47**	21.20
Peyre et.al [23]	19.40	14.60	20.90	-	-	-
Ours	**19.56**	12.79	**21.58**	**22.05**	15.77	**23.92**

Table 2. State-of-the-art comparison (in terms of mAP_{role}) on the HICO-DET using two different settings: Default and Known Object on all three sets (full, rare, non-rare). Note that Shen *et al.* [30], InteractNet [8] and GPNN [25] only report results on the Default settings. For both settings, our approach provides superior performance compared to existing methods. In case of default settings, our approach achieves mAP_{role} of 19.56 on the full set. Further, our approach obtains an absolute gain of 2.9% over TIK [14] on the full set of Known Object setting.

Tab. 2 shows the comparison on HICO-DET. As in [4], we report results on three different HOI category sets: full, rare, and non-rare with two different settings of Default and Known Objects. Our approach achieves superior performance compared to the state-of-the-art on both settings. For the Default settings, our approach obtains mAP_{role} of 19.56, 12.79 and 21.58 on the full, rare and non-rare sets, respectively. In case of Known Object setting, our approach achieves an absolute gain of 2.9% over [14] on the full set.



hit with bat, 0.73 kick sports ball, 0.87 jump skateboard, 0.94 hold umbrella, 0.69 catch kite, 0.77 carry frisbee, 0.71 snowboarding, 0.64

cut cake, 0.82 drink cup, 0.65 surf surfboard, 0.98 work on computer, 0.94 read book, 0.67 talk on phone, 0.78

Figure 5. Example detections on V-COCO. Each example involves a human-object interaction, such as *skateboarding* or multiple humans sharing the same interaction and object - *cut cake*. We also show our interaction point, vector and box (in yellow).

4.4. Ablation Study

We first perform an ablation study on V-COCO using the Hourglass backbone to show the impact of different components in our approach. Tab. 3 shows the impact of interaction points, angle-filter, dist-ratio-filter, interaction boxes, corner distance and center-pool components on the V-COCO dataset. To validate the effectiveness of our interaction grouping scheme, we first compare the proposed interaction grouping with the so-called 'angle-filter' and 'dist-ratio-filter'. Tab. 3 shows these comparisons on V-COCO.

Angle Filter: During training, the interaction point P and its corresponding human center point H, object center point O have a fixed geometric structure, *i.e.*, the angle between the vector $\overrightarrow{PH}$ and vector $\overrightarrow{PO}$ is equal to π. During inference, angle filter aims to reduce those HOI pairs for which the angle between $\overrightarrow{PH}$ and vector $\overrightarrow{PO}$ is lower than a given angle threshold. Tab. 3 shows that this baseline of interaction points with angle filter achieves a mAP$_{\text{role}}$ of 39.6.

Dist-ratio Filter: Similar to the angle-filter, the ratio between $|\overrightarrow{PH}|$ and $|\overrightarrow{PO}|$ is equal to 1 during training. Therefore the dist-ratio filter can also be employed to filter HOI pairs where the ratio between $\max(|\overrightarrow{PH}|, |\overrightarrow{PO}|)$ and $\min(|\overrightarrow{PH}|, |\overrightarrow{PO}|)$ is greater than the given distance ratio threshold. Tab. 3 shows this constraint improves over interaction points with angle filter by 1.7% in terms of mAP$_{\text{role}}$.

Interaction Grouping: To explore the effectiveness of our proposed interaction grouping scheme, we divide this scheme into two parts: *interaction box* and *corner-dist*, to verify the power of three soft constraints in (2). During training, the IoU between the human/object bbox and interaction box is greater than zero. Therefore, to satisfy the first two IoU conditions in (2), we first integrate the interaction box generated by the interaction vectors to filter out the negative HOI pairs. Tab. 3 shows that it significantly improves the HOI detection performance to 46.2 mAP$_{\text{role}}$. We also found that when only adding interaction box on the interaction points, it further improves the performance by 2%, from 46.2 to 48.2. Note that, in our approach the four corners of the interaction box are considered as the four corners of the reference box during training. At inference, with the corner distance constraint ($|\text{dist}| < d_\tau$) in (2), some negative pairs are further filtered out resulting in an improved overall performance of 50.5 mAP$_{\text{role}}$.

Add-on	Baseline					
interaction points	✓	✓	✓	✓	✓	✓
angle-filter	✓	✓	✓			
dist-ratio-filter		✓	✓			
interaction box			✓	✓	✓	✓
corner-dist					✓	✓
center-pool						✓
mAP$_{\text{role}}$	39.6	41.3	46.2	48.2	50.5	**51.0**

Table 3. Impact of integrating our contributions into the baseline on V-COCO. Results are reported in terms of role mean average precision (mAP$_{\text{role}}$). For fair comparison, we use the same backbone (Hourglass-104) for all the ablation experiments. Our overall architecture achieves a absolute gain of 11.4% over the baseline.

Score thres	0.01	0.02	0.05	0.08	0.10	Dynamic
Default:						
Full	19.26	19.32	19.08	18.66	18.25	**19.56**
Rare	12.53	12.00	10.32	9.13	8.12	**12.79**
Non-rare	21.27	21.51	**21.70**	21.50	21.27	21.58
Known-Obj:						
Full	21.80	21.81	21.57	21.08	20.65	**22.05**
Rare	15.74	15.06	13.39	12.00	10.79	**15.77**
Non-rare	23.61	23.83	**24.01**	23.80	23.60	23.92

Table 4. Performance comparison (in terms of mAP$_{\text{role}}$) regarding the classification capabilities of our approach for the rare and non-rare classes on the HICO-DET. We show the results with different score thresholds, used during the evaluation. Our proposed dynamic threshold inference achieves a good performance trade-off between the rare and non-rare classes.

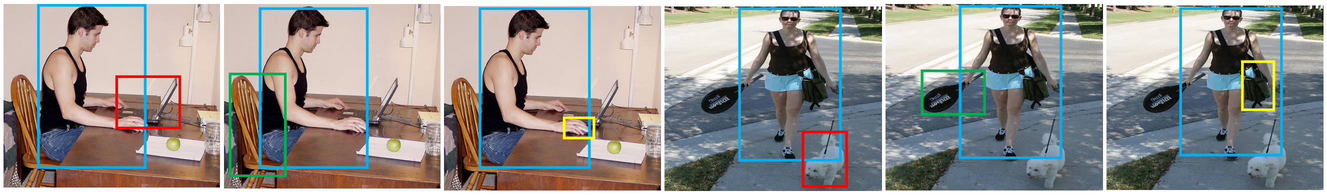

work on computer, 0.84 sit on chair, 0.77 hold mouse, 0.58 walk dog, 0.69 hold tennis racket, 0.85 carry bag, 0.74

Figure 6. Multiple interaction detection on V-COCO. Our approach detects human instance doing multiple (different) actions and interacting with various objects (represented with different colors). In all cases, the detected agent is represented with the same color.

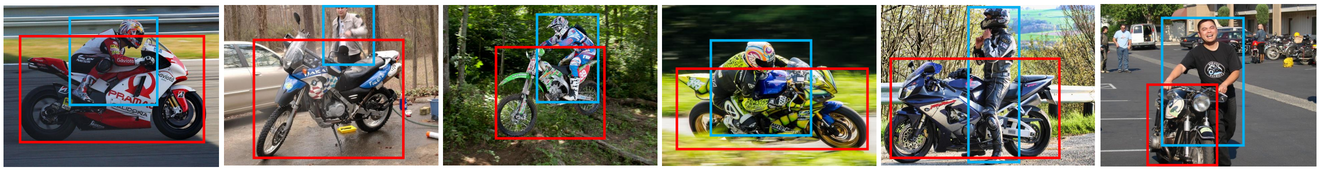

race motorcycle wash motorcycle jump motorcycle turn motorcycle park motorcycle walk motorcycle

Figure 7. Results on HICO-DET showing one detected triplet. Blue boxes represent a detected human instance, while the red boxes show the detected object of interaction. Our approach detects various fine-grained interactions.

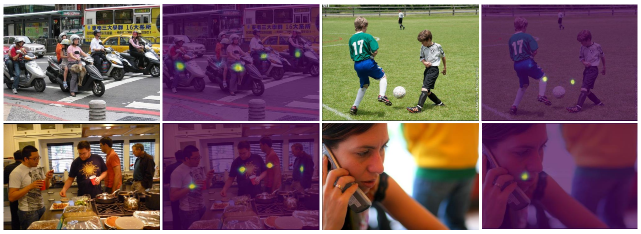


Figure 8. Visualization of interaction point heat-maps. Our method is able to cope with challenging scenarios, such as multiple HOI pairs and multiple humans sharing the same object.

Center-pool: For improved detection of interaction points, we introduce the center-pool operation, as in in [5], aiding to obtain more distinct visual patterns between the human and object instances. This operation is achieved by getting out the max summed response in both horizontal and vertical directions of the interaction point on the feature map. In our method, this is employed before the interaction point and vector branch. This operation results in a slight improvement in performance (0.5 points), as shown in Tab. 3.

We also conduct experiments to evaluate the impact of interaction score thresold on rare and non-rare action classes on HICO-DET. We select different interaction thresholds in the range $[0.01, 0.1]$, used in the test evaluation of interaction recognition performance. The results are presented in Tab. 4. These results suggest that the suitable score thresholds are different for the rare and non-rare interaction classes. This is likely due to the fact that the rare classes include less training samples, which results in relatively lower prediction scores for those classes. In contrast, the prediction scores for the non-rare classes tend to be relatively higher. Therefore, it becomes a trade-off problem for the point-based HOI methods to obtain a good performance. We further develop a dynamic threshold inference, which sets different score thresholds for different interaction classes based on their training samples. As show in Tab. 4, our dynamic threshold inference leads to a good performance trade-off between rare and non-rare classes.

4.5. Qualitative Visualization Results

Fig. 5 shows examples of both single human-object interactions, such as *hold a umbrella* and *work on computer*, and multiple humans sharing same interaction and object (*cut cake*) along with corresponding interaction scores on V-COCO. The interaction boxes (yellow dash line) generated by the interaction vectors (yellow solid line) are also drawn. These interaction boxes are paired with the positive human and object bounding-boxes using interaction grouping. Fig. 6 shows examples of a human performing multiple interactions. Different interaction objects are annotated with bounding-boxes of different colors. Fig. 7 shows fine-grained human-object interaction results on HICO-DET. The heatmap visualization for the interaction point map is shown in Fig. 8. Similar to previous works, we observe long-tailed classes to be particularly challenging for HOI detection. Further, a minor limitation of our approach is that multiple HOI pairs cannot share the same interaction point. However, such cases are rare in practice.

5. Conclusion

We propose a point-based framework for HOI detection. Our approach regards the HOI detection as a keypoint detection and grouping problem. The interaction point and its corresponding interaction vector are first generated by the keypoint detection network. Then, we directly pair those interaction points with the human and object bounding boxes from object detection branch using the proposed interaction grouping scheme. Experiments are performed on two HOI detection benchmarks. Our points-based approach outperforms state-of-the-art methods on both datasets.

References

[1] Jiale Cao, Yanwei Pang, and Xuelong Li. Triply supervised decoder networks for joint detection and segmentation. In *CVPR*, 2019. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

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

[4] Yuwei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 2, 4, 5, 6

[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019. 8

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 2

[7] 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

[8] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object intaractions. In *CVPR*, 2018. 2, 6

[9] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 4, 5, 6

[10] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019. 6

[11] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 6

[12] Alexander Kolesnikov, Christoph H. Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. In *arXiv preprint arXiv:1807.02136*, 2018. 6

[13] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 2, 3, 4, 5

[14] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Yan-Feng Wang Hao-Shu Fang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 1, 2, 3, 6

[15] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 3, 6

[16] 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. 6

[17] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 2

[18] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 3, 6

[19] Jing Nie, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Enriched feature guided refinement network for object detection. In *ICCV*, 2019. 2

[20] Yanwei Pang, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Efficient featurized image pyramid network for single shot detector. In *CVPR*, 2019. 2

[21] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-Guided attention network for occluded pedestrian detection. In *ICCV*, 2019. 2

[22] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 3

[23] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting unseen visual relations using analogies. In *ICCV*, 2019. 1, 6

[24] Charles Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018. 2

[25] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018. 1, 2, 6

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[27] Joseph Redmon and Ali Farhadi. Yolo9000:better, faster, stronger. In *CVPR*, 2017. 2

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 6

[29] Fahad Shahbaz Khan, Jiaolong Xu, Joost van de Weijer, Andrew Bagdanov, Rao Muhammad Anwer, and Antonio Lopez. Recognizing actions through action-specific person detection. *IEEE Transactions on Image Processing*, 24(11):4422–4432, 2015. 2

[30] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018. 6

[31] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 2

[32] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, , and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019. 6

[33] Tiancai Wang, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Learning rich features at high-speed for single-shot object detection. In *ICCV*, 2019. 2

[34] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019. 1, 2, 3, 6

[35] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018. 2

[36] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, 2019. 2

[37] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mo-

han Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019. 6

[38] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019. 6

[39] Tianfei Zhou, Wenguan Wang, Siyuan Qi, Jianbing Shen, and Haibin Ling. Cascaded human-object interaction recognition. In *CVPR*, 2020. 1

[40] Xingyi Zhou, Dequan Wang, and Philipp Krahenbuhl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2, 3, 4, 6

[41] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 2, 3