Interact as You Intend: Intention-Driven Human-Object Interaction Detection

Bingjie Xu, Junnan Li, Yongkang Wong, *Member, IEEE,* Qi Zhao, *Member, IEEE,* and Mohan S. Kankanhalli, *Fellow, IEEE*

***Abstract*—The recent advances in instance-level detection tasks lay strong foundation for genuine comprehension of the visual scenes. However, the ability to fully comprehend a social scene is still in its preliminary stage. In this work, we focus on detecting human-object interactions (HOIs) in social scene images, which is demanding in terms of research and increasingly useful for practical applications. To undertake social tasks interacting with objects, humans direct their attention and move their body based on their intention. Based on this observation, we provide an unique computational perspective to explore human intention in HOI detection. Specifically, the proposed human intention-driven HOI detection (iHOI) framework models human pose with the relative distances from body joints to the object instances. It also utilizes human gaze to guide the attended contextual regions in a weakly-supervised setting. In addition, we propose a hard negative sampling strategy to address the problem of mis-grouping. We perform extensive experiments on two benchmark datasets, namely V-COCO and HICO-DET, and show that iHOI outperforms the existing approaches. The efficacy of each proposed component has also been validated.**

***Index Terms*—Human-Object Interactions (HOIs), Intention-Driven Analysis, Visual Relationships**

I. Introduction

In recent years, computer vision models have made tremendous improvements, especially in the instance-level tasks such as image classification and object detection [1], [2], [3], [4]. The advances in these fundamental tasks bear great potential for many fields, including security, medical care and robotics [5], [6], [7]. Enabling such applications requires deeper understanding of the scene semantics beyond instance-level understanding. Existing efforts on the high-level semantic understanding include visual relationships inference [8], [9], scene graphs generation [10], and visual reasoning [11], [12]. In this work, we focus on an important task that is human-centric, namely human-object interaction (HOI) detection, stepping towards higher level scene understanding.

The task of HOI understanding [13], [14], [15], [16], [17] is formulated as identifying the $\langle human, action, object \rangle$ triplets. It is a facet of visual relationships critically driven by humans. In contrast to general visual relationships involving verbs, prepositional, spatial, and comparative phrases, HOI understanding focuses on direct interactions (actions) performed on objects (e.g. *a person is holding a cup* in Figure 1). Precise detection and inference of HOIs are increasingly needed in practical applications, such as development of collaborative robotics, activity mining in social networks, and event detection in surveillance [7], [18], [19]. Nevertheless, it still remains a challenging research problem due to the fine granularity of actions and objects in social scenes. Earlier approaches for HOI detection mainly focus on the representation of visual data, such as joint modeling of body poses, spatial configuration and functional compatibility in images [13], [20], [21]. In recent years, several large-scale datasets with diverse interaction classes have enabled fine-grained exploration of HOIs [16], [17], [22], [23]. Motivated by advances in deep learning, especially the success of Convolutional Neural Network in object detection and classification, recent works utilize those datasets to learn deep visual representations of human and object for HOI detection [15], [16], [17], [22]. However, those works do not take special consideration that a human often exhibits purposeful behaviors with intention in mind to complete tasks. For example, in Figure 1, *the person is lifting the kettle and holding the cup, gazing around the target cup – intended to pour water into the cup.*

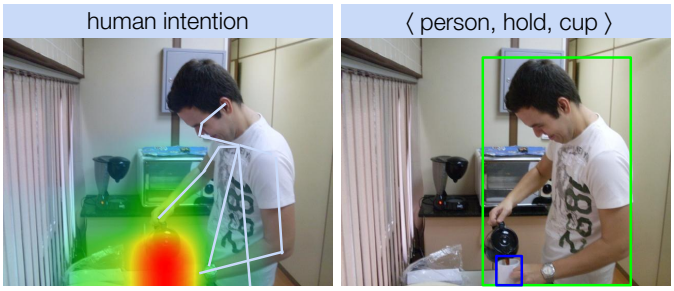

Fig. 1: An example where the actor's intention is informative of the HOI $\langle person, hold, cup \rangle$. The intention is represented using the attended regions and body pose. Specifically, he is fixating at the regions around the cup that he is interacting with, and his posture implicitly conveys his intention.

In cognitive studies, human intention is reported to commonly unveil complementary cues to explain the behaviors of individuals attempting to accomplish certain tasks [24], driving the coordination of eye and body movements [25]. For example, when interacting with a specific object, human tends to exhibit corresponding intention by adjusting position, pose

Bingjie Xu and Junnan Li are with the Graduate School for Integrative Sciences and Engineering, National University of Singapore (email: {bingjiexu, lijunnan}@u.nus.edu). Yongkang Wong is with the Smart Systems Institute, National University of Singapore (email: yongkang.wong@nus.edu.sg). Qi Zhao is with the Department of Computer Science and Engineering, University of Minnesota (email: qzhao@cs.umn.edu). Mohan S. Kankanhalli is with the School of Computing, National University of Singapore (email: mohan@comp.nus.edu.sg).

Manuscript received XXX, 201X; revised XXX, 201X.

and shifting attention (see Figure 1). By perceiving the latent goal, we can facilitate the inference of the interactions. In this work, we provide a novel computational perspective to exploit two forms of human intention that is visually observable: 1) human gaze, which explicitly conveys intention; 2) human body posture, which implicitly conveys the intention. The work most related to ours is the one characterizing human intention with attention and body skeleton [26]. Nevertheless, human intention has yet been investigated in the context of HOI detection in an integrative manner. Also, we offer more robustness to inaccurate gaze localization by exploring multiple contextual regions driven by gaze.

We utilize gaze to guide the model in exploring multiple object instances in a scene. The scene information has exhibited positive influence in various recognition tasks [9], [27], [28]. One approach to utilize this information is directly extracting the visual representations from Scene-CNN [27]. However, it is inefficient in some tasks due to the lack of instance-level information. Another approach is to leverage the visual cues from surrounding objects [9], [28]. Such cues could be informative as a semantic scene constraint. For example, when a $\langle human, spoon \rangle$ pair is surrounded by dining table, cups, bowls and microwave, the interaction class is recognized as *eating* in a dining scene rather than *selling* at a market. To leverage the informative scene instances, the existing approaches learn from corresponding tasks using large scale visual samples [9], [28], [29]. In contrast to their approaches, we propose to infer the informative regions from intention of human by utilizing the actor's gaze cue.

In this work, we aim to tackle the challenge of accurately detecting and recognizing HOIs in social scene images. We propose a human intention-driven HOI detection (iHOI) framework, consisting of an object detection module and two branches. The first branch models differential human-object feature embeddings, and the second leverages multiple gazed context regions in a weakly-supervised setting. Human pose information has been incorporated into the feature spaces using the relative distances from body joints to the instances. The contributions of this work are summarized as follows:

1) We have explored how to detect and recognize *what humans are doing in social scenes* by inferring *what they intended to do*. Specifically, we provide an unique computational perspective to exploit human intention, commonly explored in cognitive studies, and propose a joint framework to effectively model gaze and body pose information to assist HOI detection.
2) We propose an effective hard negative sample mining strategy to address commonly observed mis-grouping problem in HOI detection.
3) We perform extensive experiments on two benchmark datasets with ablation studies, and show that iHOI outperforms the existing approaches.

The rest of the paper is organized as follows. Section II reviews the related work. Section III delineates the details of the proposed method. Section IV elaborates on the experiments and discusses the results. Section V concludes the paper.

II. Related Work

This section reviews prior works related to visual relationships, HOI detection, and gaze in HOIs.

Visual Relationship Detection. The inference of general visual relationships [8], [10], [30] has attracted increasing research interests. The types of visual relationships include verbs, spatial, preposition or comparative phrase. Recently, Lu *et al.* [8] learned a language prior to refine visual relationships from vocabulary. Zhang *et al.* [30] embedded object class probabilities to highlight semantics constraint in visual relationships. Our focus is related, but different. We aim to explore direct interactions (actions) performed on objects, where human is the crucial indicator of the interactions.

HOI Detection. Different from visual relationships detection task, which focuses on two arbitrary objects in the images, HOI recognition is a human-centric problem with fine-grained action categories. Earlier studies [13], [20], [21], [31] mainly focus on recognizing the interactions, by joint modeling of body poses, spatial configuration, and functional compatibility in the images.

In recent years, several human-centric image datasets have been developed to enable fine-grained exploration of HOI detection, including V-COCO (*Verbs-COCO*) [16], HICO-DET (*Humans Interacting with Common Objects-DET*) [17], and HCVRD (*Human-Centered Visual Relationship Detection*) [23]. In these datasets, the bounding boxes of each human actor and the interacting object are annotated, together with the corresponding interactions. Motivated by the success of deep learning, especially Convolutional Neural Network for object detection and recognition, several recent works have taken advantage of the detailed annotated datasets to improve HOI detection. Gkioxari *et al.* [15] leveraged the human action-specific density to constrain potential objects, and significantly improved the precision of localizing interacting object. Chao *et al.* [17] set the benchmark in HICO-DET based on a three-stream detection framework, exploiting the visual and spatial representations of human, object and the pairwise bounding box. Shen *et al.* [32] analyzed the zero-shot problem with separate verb and object detection losses. Zhuang *et al.* [23] addressed the long-tail issue with supervision from web data.

In contrast to previous works treating humans and objects similarly, with no consideration that human behaviors are purposeful, we argue that human intention drives interactions. Therefore, in this work, we exploit the cues in an image that reflect an actor's intention, and leverage such information for more effective HOI detection.

Gaze in HOIs. Humans are the core element in HOI. Generally, an actor intends to leverage essential information in the scene to help performing the interaction. One important facet of intention is reflected by the gaze, which explicitly shows the task-driven attention [33]. Cognitive studies have reported that human often attends to the region that provides significant information during an interaction [34]. Though this might not be true in some cases such as lifting a familiar object without fixating at it, in general the fixated region provides informative cues.

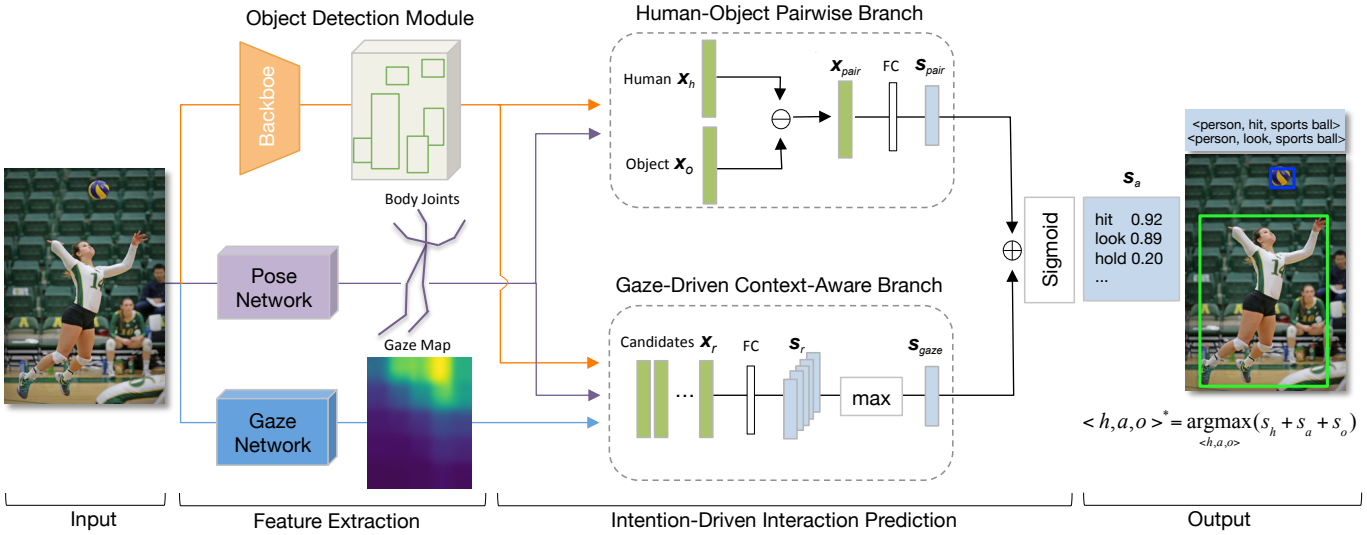

Fig. 2: The proposed iHOI takes an image as the input for feature extraction and human intention-driven interaction prediction, and outputs the detected triplets. Human intention is modeled as follows: 1) the pose information is incorporated with the distances from body joints to the instance center; 2) human gaze guides the attended context regions in a weakly-supervised setting. The Feature spaces of $\boldsymbol{x}_h$, $\boldsymbol{x}_o$ and $\boldsymbol{x}_r$ are represented with class probabilities $\boldsymbol{v}_c$, visual appearance $\boldsymbol{v}_a$, relative locations $\boldsymbol{v}_l$, and human pose information $\boldsymbol{v}_p$. Scores $\boldsymbol{s}_{pair}$, $\boldsymbol{s}_{gaze}$ and $\boldsymbol{s}_a$ are of the same dimension as the action amount. Operations $\ominus$ and $\oplus$ denote element-wise subtraction and summation.

Inspired by the cognitive study findings, some computational tasks have explored human gaze [35], [36], [37]. For example, Fathi *et al.* [35] built a probabilistic generative model to simultaneously predict the sequence of gaze locations and the respective action label from first person view videos. Mukherjee and Robertson [38] estimated the gaze direction based on the head pose in multimodal videos, and managed to recover human-human/scene interactions. In the image domain, Recasens *et al.* [36] proposed a method to detect the object regions being fixated at by human in the scene. In another task, Gorji and Clark [37] augmented saliency prediction in images by incorporating the actor's gaze location. Despite the efforts made in various tasks, the existing methods have yet to explore actor's gaze in the context of HOI detection. In this study, we explore the role of actor's gaze in guiding informative scene regions for HOIs.

III. Proposed Method

This section presents the human intention-driven HOI detection (iHOI) framework, as shown in Figure 2. The task is formulated as follows: given a 2D image $\boldsymbol{I}$ as the input, it aims to detect and recognize triplets of the form $\langle human, action, object \rangle$. We will first describe our model architecture, followed by the details of training and inference.

A. Model Architecture

An overview of the proposed model is shown in Figure 2. The iHOI framework consists of three modules: 1) object detection module, 2) human-object pairwise branch, and 3) gaze-driven context-aware branch. Human body joints locations and gaze direction are obtained through transfer learning from other social-activity datasets [39], [36], since our aim is to effectively model intention rather than extract features, and both experimental datasets lack the ground-truth. Object proposals are generated by Faster R-CNN [2]. The first branch focuses on a specific human-object pair, and learns a differential feature embedding $\boldsymbol{x}_{pair}$ to produce a score vector $\boldsymbol{s}_{pair}$ over possible action classes. The second branch leverages the gaze of the actor to exploit the contextual regions that the actor is attending to, and produces a score $\boldsymbol{s}_{gaze}$. The final action score is defined as

$$\boldsymbol{s_a} = Sigmoid\big(\boldsymbol{s}_{pair} \oplus \boldsymbol{s}_{gaze}\big) \tag{1}$$

where the operation $\oplus$ refers to element-wise addition. We use sigmoid function because we need to classify multiple actions independently. For example, *a person can be standing and looking at a skateboard simultaneously*. The training objective is to minimize the binary cross entropy loss between the ground-truth labels and the predicted scores $\boldsymbol{s_a}$.

Next, We describe each component of the architecture.

1) Object Detection Module: We adopt Faster R-CNN [2] for object detection. First, a Region Proposal Network (RPN) is used to generate object proposals (ROIs). Then, ROI pooling is applied on each object proposal to extract a fixed-length feature vector of each ROI. Object classification and bounding box regression are performed, generating a set of bounding boxes, each associated with object classification probabilities. These are the candidate bounding boxes $\boldsymbol{b} = (b^1, ..., b^m)$ used for HOI detection.

2) Human-Object Pairwise Branch: Given a human bounding box b_h and an object bounding box b_o generated from the object detection module, we aim to learn a pairwise feature embedding that can preserve their semantic interactions. For example, the interaction of *a person riding a bike* can almost

be described by the visual appearance of the pair, person on top of the object, and the estimated *bike* label.

Similar to the recent *VTransE* [30] for general visual relationships among objects, the feature space $\boldsymbol{x}$ for each b_h and b_o contains visual appearance, relative spatial layout, and object semantic likelihood, referred as $\boldsymbol{v}_a$, $\boldsymbol{v}_l$ and $\boldsymbol{v}_c$ respectively. $\boldsymbol{v}_a$ is a 2048-d vector, extracted from *fc7* layer in the object detector to capture the appearance of each b_o. $\boldsymbol{v}_l$ is a 4-d vector consisted of $\{l_x, l_y, l_w, l_h\}$. $\{l_x, l_y\}$ specifies the bounding box coordinates distances, and $\{l_w, l_h\}$ specifies the log-space height/width shift, all relative to a counterpart as parameterized in Faster R-CNN. $\boldsymbol{v}_c$ is a 81-d vector of object classification probabilities over MS-COCO object categories, generated by the object detectors.

In contrast to the general visual relationships, we extend the feature space with human pose information since our task is intrinsically human-centric. Human pose bridges the human body with the interacting object. For example, the up-stretching arms, jumping posture and the relative distances to the ball possibly reveal that the person is *hitting a sports ball*. Since body pose ground-truth is not available, we use the pose estimation network in [39] to extract body joints locations for each human. The output of the pose estimation network is the locations of 18 body joints. We consider eight representative body joints[1] that are more frequently detected, which cover the head, upper and lower body. For each joint $i \in 1, ..., 8$, we calculate its distance from the center of b_h and b_o to get two distance vectors $\{d^i_{xh}, d^i_{yh}\}$ and $\{d^i_{xo}, d^i_{yo}\}$, where d^j_{xh} denotes its distance from b_h center along x-axis, and d^j_{yo} denotes the distance from b_o center along y-axis. Since human-object pairs have different scales, we normalize the distances *w.r.t.* the width of b_h. We concatenate the normalized distance vectors for all eight joints to get two 16-d vectors $\boldsymbol{v}^h_p = \{d^i_{xh}, d^i_{yh} | i = 1, ..., 8\}$ and $\boldsymbol{v}^o_p = \{d^i_{xo}, d^i_{yo} | i = 1, ..., 8\}$ that encode the pose information. In cases where not all eight joints are detected, we set $\boldsymbol{v}_p$ to be zeros. An alternative way of implementing the pose information have been experimented, as shown in Section IV.

The above-mentioned features are concatenated to form the feature spaces for human $\boldsymbol{x}_h = \{\boldsymbol{v}^h_c, \boldsymbol{v}^h_a, \boldsymbol{v}^h_l, \boldsymbol{v}^h_p\}$ and object $\boldsymbol{x}_o = \{\boldsymbol{v}^o_c, \boldsymbol{v}^o_a, \boldsymbol{v}^o_l, \boldsymbol{v}^o_p\}$. Following [30], we calculate the pairwise feature embedding as

$$\boldsymbol{x}_{pair} = \boldsymbol{x}_h \ominus \boldsymbol{x}_o \tag{2}$$

The differential embedding is used since it represents the comparative information between the human and object feature spaces, increasing the discriminative ability. Pairwise feature summation has also been experimented but shown less effectiveness. The pairwise embedding is passed through a fully-connected layer to produce the pairwise action scores $\boldsymbol{s}_{pair}$.

3) Gaze-Driven Context-Aware Branch: We observe that the regions where an actor is fixating often contain useful information for the interactions. For example, when the person intends to pour water to a cup, he normally *fixates around the cup while holding it*. Therefore, we exploit the fixated contextual information to help recognizing the actor's action. In particular, we use human gaze as a guidance to leverage the fixated scene regions. The gazed location is predicted with a pretrained two-pathway model proposed in [36]. As there is no gaze ground-truth in the HOI datasets, we have manually checked the gaze prediction and observed that most predictions are reasonable. The gaze prediction model takes the image $\boldsymbol{I}$ and the central human eye position (calculated from the pose estimation network) as input, and outputs a probability density map $\boldsymbol{G}$ for the fixation location.

For each human in the image, we select five regions from the candidates $\boldsymbol{b} = (b^1, ..., b^m)$ generated by the object detectors, which have higher probabilities of being fixated on. Specifically, for each candidate region $b \in \boldsymbol{b}$, we assign a gaze weight g_b to it, where g_b is obtained by summing up the values of $\boldsymbol{G}$ in b and then normalized by the area of b:

$$g_b = \frac{\sum_{x,y \in b} \boldsymbol{G}_{x,y}}{area_b}, b \in \boldsymbol{b} \tag{3}$$

Then we select the top-5 regions $\boldsymbol{r} = (r^1, ..., r^5)$ that have the largest g_b. We have experimented with different numbers of candidate regions from one to all of the detected objects. Using top-5 candidate regions guided by gaze achieves plateau performance, which suggests that five candidates are sufficient to capture informative cues. For each selected region r, we first get its corresponding feature vector $\boldsymbol{x}_r = \{\boldsymbol{v}^r_c, \boldsymbol{v}^r_a, \boldsymbol{v}^r_l, \boldsymbol{v}^r_p\}$, and pass it through a fully-connected layer to acquire the action scores $\boldsymbol{s}_r$ of each region $r \in \boldsymbol{r}$. Then we compute the prediction score for this branch as follows:

$$\boldsymbol{s}_{gaze} = \max(\boldsymbol{s}_r), r \in \boldsymbol{r} \tag{4}$$

$\max(\cdot)$ is used because generally there is only one region an actor can fixate on. The most informative region among the gazed candidates can be discovered in a weakly-supervised manner. Note that if the gaze of the actor cannot be predicted (i.e. the eyes are invisible in images, or the actor faces to the frontal direction to the camera), we set $\boldsymbol{x}_r = 0$.

In contrast to a recent work [26] directly leveraging the fixated patch, learning with multiple gazed regions in our framework makes it robust to the inaccurate gaze predictions, i.e. it can still find the most informative region among a reasonable amount of guided candidate regions.

B. Hard Negative Triplet Mining

We observe that mis-grouping is a common category of false positive HOI detection [16]. Mis-grouping refers to cases where the class of a HOI is correctly predicted, but a wrong object instance is assigned to the actor (e.g. a person is *cutting* another person's cake). We argue that such negative HOI triplets are more difficult for a model to reject in the tasks requiring pairing proposals, due to less discriminative patterns, compared with other negative triplets of inaccurate localization and false classification in [17].

We propose a simple yet effective method to mine for those hard negative triplets. For each image, we deliberately mis-group a non-interacting human-object pair from the annotated pairs, and label their action labels as negative, i.e. all zeros. These human-object pairs together with negative labels form

[1]nose, neck, left and right shoulder, left and right elbow, left and right hip

the negative triplets. We adopt the image-centric training strategy [40], where each mini-batch of HOI triplets arises from a single image. We empirically keep the number of positive and negative triplets in each mini-batch to be 1:2.

C. Inference

During inference, we aim to calculate the HOI score $s_{h,o,a}$ for a triplet $\langle human, action, object \rangle$. Given the detection score for human s_h, object s_o, and the largest value s_a among the scores for all actions $\boldsymbol{s}_a$, we decompose $s_{h,o,a}$ as follows:

$$s_{h,o,a} = s_h + s_o + s_a \tag{5}$$

To predict HOIs in an image, we must compute the scores for all detected triplets. However, scoring every potential triplet is almost intractable in practice, calling for high-recall human-object proposals. To solve this, we leverage the predefined relevant object categories $c \in C$ for each action [16] as a prior knowledge, which is extracted from the HOI ground-truth in the corresponding dataset. For instance, *sports ball* is relevant to the action *kick* but *book* is not relevant. Unlike pairing human and objects according to the ground-truth during training, we filter out the detected objects irrelevant to the action for each human-action pair during inference/testing. We then select the object that maximizes the triplet score $s_{h,o,a}$ within each relevant category to form the triplet. Note that for HICO-DET, there exist many samples of human interacting with multiple objects of the same category (e.g. *a person is herding multiple cows*), therefore we retain at most 10 objects sorting by $s_{h,o,a}$ for each human-action pair.

With objects selected for each human and action, we have triplets of $\langle human, action, object \rangle$. The bounding boxes of the human-object pairs, along with their respective HOI triplet score $s_{h,o,a}$, are the final outputs of our model.

IV. Experiments

In this section, we first describe the benchmark datasets (i.e. V-COCO [16] and HICO-DET [17]), evaluation metric, and implementation details. Then, we compare our proposed iHOI framework to the existing approaches, and show that it outperforms the others. Ablation studies are also conducted to examine the effect of each proposed component. Finally, we show some qualitative examples, as well as discuss on some failure cases.

A. Datasets and Evaluation Metric

There exist a number of HOI datasets [16], [17], [22], [23]. In this work, we focus on the V-COCO dataset [16] and HICO-DET dataset [17], which is more relevant for HOI detection task. The other datasets either is not in the context of detection task or contains general human-object predicates that are out of our exploration range.

V-COCO Dataset [16]. This dataset is a subset of MS-COCO [41], with 5400 images in the trainval (training plus validation) set and 4946 images in the test set. It is annotated with 26 common action classes, and the bounding boxes for human and interacting objects. In V-COCO, a person can perform multiple actions on the same object (e.g. *skiing* the skis while *holding* it), and perform the same action on different types of objects. In particular, three actions (i.e. cut, hit, eat) are annotated with two types of targets (i.e. instrument and direct object). For example, a person can be hitting *racket* (instrument) and *sports ball* (direct object) simultaneously.

HICO-DET Dataset [17]. HICO-DET contains 38118 images in the training set and 9658 test images. It is annotated with 600 types of interactions: 80 object categories as in MS-COCO and 117 verbs. The bounding boxes of human and targeting objects are also annotated. Similar to V-COCO, HICO-DET allows a person to perform multiple actions on the same object, or perform the same action on multiple objects (e.g. herding *multiple cows*).

Evaluation Metric. We follow the standard evaluation metric and report mean Average Precision (mAP). AP is computed based on both recall and precision, which is appropriate for detection task. Our aim is to detect interactions between human and objects, thus, actions without any interacting object (i.e. run, smile, stand, walk, and point [15]) are out of the exploration range. Formally, a triplet of $\langle human, action, object \rangle$ is considered as a true positive if: 1) the predicted human box has $IoU \geq 0.5$ with the ground-truth human box, 2) the predicted object box has $IoU \geq 0.5$ with the ground-truth object in interaction, and 3) the predicted and ground-truth actions match. The definition of true positive is identical except that HICO-DET considers the specific object categories, while V-COCO considers rough object types, namely *instrument* and *direct object*, as in the standard evaluation metric. Our method can predict the object categories for both datasets.

B. Implementation Details

Our implementation is based on Faster R-CNN [2] with a Feature Pyramid Network (FPN) [3] backbone built on ResNet-50 [4]. The weights are pretrained on MS-COCO dataset. We use the approximate FPN baseline implementation [42] that resizes the image with 800 pixels, adds 32×32 anchors, and keeps 1000 proposals. The Faster R-CNN object detector is reported to have mAP of 34.2% on the MS-COCO minival split [42], which shares the same object categories as in V-COCO and HICO-DET. For object detection, we apply non-maximum suppression with IoU threshold of 0.2 on candidate boxes, and set a threshold of 0.15 on the object score. The thresholds are set conservatively to keep most objects.

The model is trained for 5000 iterations with a learning rate of 0.001 and another 5000 iterations with rate of 0.0001 to converge. The object detection backbone is kept frozen during training. We follow the image-centric training strategy with mini-batch size set to 32 for both datasets. We use a weight decay of 0.0005 and a momentum of 0.9. Stochastic gradient descent (SGD) is used for optimization.

The aim of this work is to effectively model human intention into HOI detection framework rather than extract features. Therefore, human gaze and pose information are transferred from other social-activity datasets [39], [36]. Our framework could be further trained in an end-to-end manner if the human gaze and body pose annotations are available.

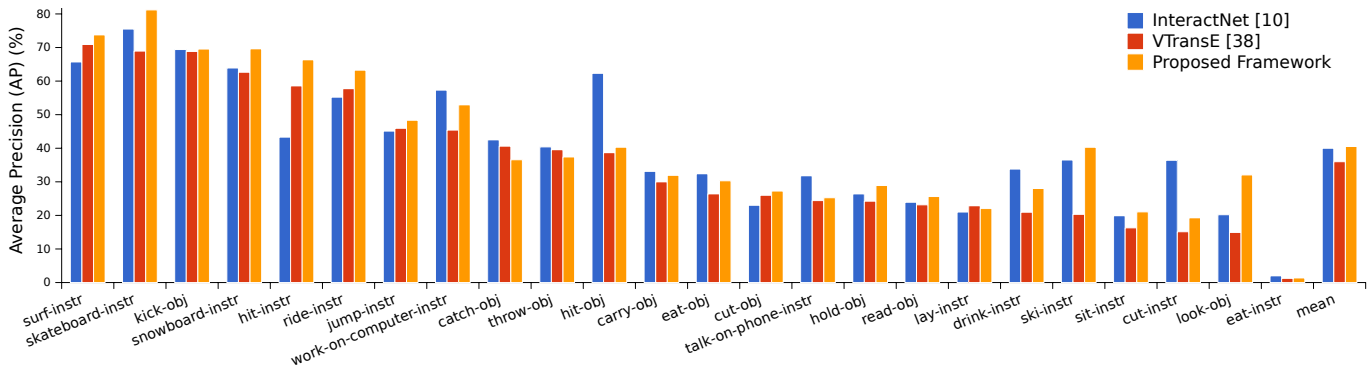

Fig. 3: Per-action mAP (%) of the triplets (mAP_{role}) on V-COCO test set. We show two main baselines and our framework for each of the actions with interacting objects. There are 26 actions defined in [16]. Five out of them (*point, run, smile, stand, walk*) are defined without objects, out of our exploration range. We list the detailed results for 21 actions together with all possible types of targeting objects (i.e. *instrument* and/or *direct object*).

TABLE I: Comparisons with the state-of-the-arts approaches and variants of iHOI on V-COCO dataset. mAP (%) equals to mAP_{role} as in the standard evaluation metric.

Methods	mAP (%)
VSRL [16]	31.80
InteractNet [15]	40.00
VTransE [30]	35.63
(a) w/ pose locations	35.65
(b) w/ P (pose distances)	35.83
(c) w/ sorted r	36.78
(d) w/ G (gazed r)	37.49
(e) w/ P+G	37.65
(f) w/ P+G+an alternative mining [17]	39.28
iHOI	**40.41**

TABLE II: Comparisons with the state-of-the-arts approaches and variants of iHOI on HICO-DET dataset. Results are reported with mean Average Precision (mAP) (%).

Methods	Full	Rare	Non-Rare
Shen *et al.* [32]	6.46	4.24	7.12
HO-RCNN [17]	7.81	5.37	8.54
InteractNet [15]	9.94	**7.16**	10.77
VTransE [30]	7.87	6.01	8.43
(a) w/ pose locations	7.89	6.01	8.45
(b) w/ P (pose distances)	7.95	6.02	8.52
(c) w/ sorted r	8.39	6.13	9.06
(d) w/ G (gazed r)	8.65	6.26	9.37
(e) w/ P+G	8.72	6.27	9.45
(f) w/ P+G+an alternative mining [17]	9.35	6.82	10.11
iHOI	**9.97**	7.11	**10.83**

C. Comparing with the State-Of-The-Art Approaches

In this work, we compare our proposed framework with the existing approaches to evaluate our model. Specifically, we compare with the following approaches in V-COCO

- **VSRL** [16] establishes V-COCO dataset, and proposes to regress to the target location. It is reimplemented by [15] with ResNet-50-FPN backbone for fair comparisons.
- **InteractNet** [15] jointly models object detection and interaction classification with target re-localization, achieving the existing best performance.
- **VTransE** [30] is a base framework of our method upon which the iHOI variants are implemented. It is originally proposed for visual relationships, and reimplemented by us with ResNet-50-FPN backbone for fair comparisons.

whereas the following approaches are compared in HICO-DET together with InteractNet and VTransE

- **Shen** *et al.* [32] focuses on the zero-shot problem with separate verb and object detection losses.
- **HO-RCNN** [17] establishes HICO-DET dataset, as well as forms the benchmark performance with human-object pairwise visual and spatial representations.

In general, Table I and Table II show that iHOI outperforms other approaches. HICO-DET is generally observed with lower mAP because it contains more fine-grained HOI categories with severe long-tail problem, and is evaluated with specific object categories rather than the two rough types of objects in V-COCO. Our iHOI outperforms the best performing method (i.e. InteractNet) with improvements of +0.41 on V-COCO and +0.03 on HICO-DET.

Existing approaches mainly rely on the pairwise human-object appearance and spatial relationships. However, some complicated interactions are very fine-grained, which make it hard to distinguish only by appearance and relative locations. On the other hand, the proposed iHOI jointly takes advantages of the gazed scene context and subtle differences of the body movements. A discriminative pattern between the positive and hard negative samples is also learnt. Thus it achieves better overall results, albeit the less effective performance on the rare split due to the limited training samples.

To study the effectiveness on various interaction classes, we analyse the mAP for each action-target type defined in V-COCO. Figure 3 shows the result of InteractNet, the base framework VTransE [30], and iHOI. We observe consistent actions with leading mAP, such as *surf, skateboard*. Our proposed framework improves most action-target categories compared to VTransE. The actions with the largest improvement are those closely related with human intention such as *look, work-on-computer*, and those likely to be mis-grouped such as *skateboard, ski*. The three actions (i.e. *catch, throw, lay*) where our iHOI shows no improvement over VTransE are the confusing ones, requiring more discriminative patterns.

Comparing the proposed iHOI with InteractNet shows that

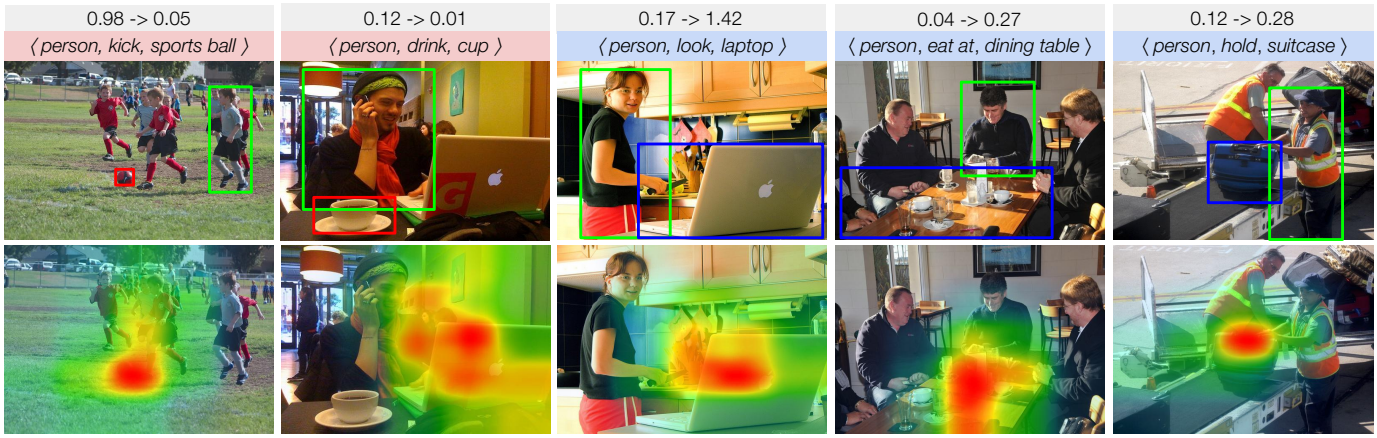

Fig. 4: The effect of human intention in V-COCO and HICO-DET. HOI predictions together with the triplet scores (with gray headings) are shown. After leveraging intention (*VTransE w/ P+G* vs. *VTransE*), we show the change in triplet scores for the detections. Using intention suppresses the false prediction scores (i.e. column 1 – 2 with red headings), whereas improves the correct ones (i.e. column 3 – 5 with blue headings). The corresponding gaze density heatmaps intuitively demonstrate that fixated regions are informative of HOIs. The pose information is not plotted. Triplet scores are obtained as in Section III-C.

iHOI can achieve overall better performance on most action-target categories, whereas showing notably worse performance on a small proportion of the categories such as *hit-obj, cut-instr, drink-instr*. We observe that iHOI performs worse mostly on actions with small objects, mainly due to inaccurate object detection compared to InteractNet with target re-localization.

D. Ablation Studies

In this section, we examine the impact of each proposed component with the following iHOI variants upon the base framework VTransE, shown in Table I and Table II:

(a) **w/ pose locations**: The relative locations of body joints *w.r.t* the image size are used to compute an additional set of action scores.
(b) **w/ P (pose distances)**: The relative distances from body joints to the instance are concatenated into the respective human and object feature spaces, as in iHOI.
(c) **w/ sorted** r: An additional context-aware branch is implemented, and the top-5 scene regions are selected by detection scores (w/o pose).
(d) **w/ G (gazed** r**)**: An additional gaze-driven context-aware branch is implemented (w/o pose).
(e) **w/ P+G**: Body joints distances and gaze information are incorporated into the two-branch model, equivalent to the proposed iHOI without hard negative triplet mining.
(f) **w/ P+G+an alternative mining [17]**: A general mining method [17] is used in addition to (f), to compare with our proposed mining strategy.

The reimplemented base framework VTransE [30] achieves solid performance on both datasets, due to the effective pairwise embedding. Our iHOI achieves gains in mAP of +4.78 on V-COCO and +2.1 on HICO-DET, which are relative improvements of 13.42% and 26.68% over VTransE.

We analyze the effect of each component as follows.

1) Gazed Context: Human gaze explicitly conveys his/her intention, which drives the attended scene regions in the intention-driven branch, forming a key component of our method. The ablation results are colored with blue. (c) leverages the scene regions sorted by detection scores without gaze guidance, which improves upon the base framework VTransE with +1.15 and +0.52 on V-COCO and HICO-DET, respectively. The improvements indicate that scene regions are informative for HOI detection. By utilizing the actor's fixated regions guided by gaze, (d) further achieves improvements of +0.71 and +0.26 compared to (c). This demonstrates that the actor's fixated regions can reasonably provide information in detecting HOIs even with some ambiguous gaze predictions. The effectiveness of using the gazed context is also demonstrated by (e) vs. (b).

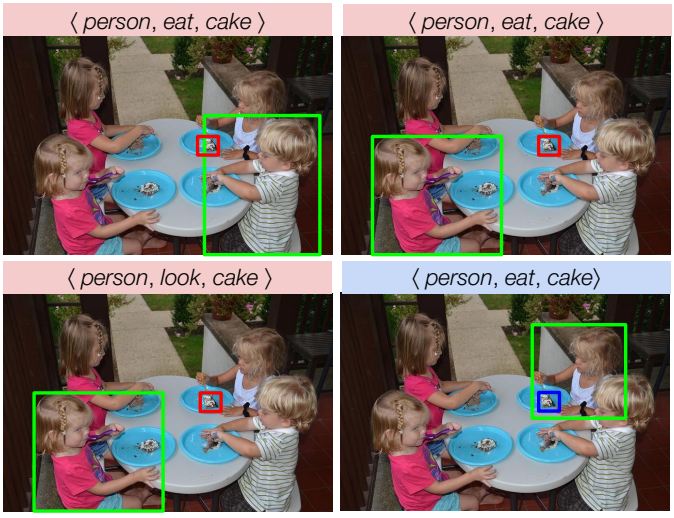

Fig. 5: Detections with triplet score larger than 0.7 are displayed. Without the proposed negative triplet mining, the model gives all four predictions, in which three of them (i.e. with red headings) are mis-grouped. Model with negative triplet mining reduces the prediction to only the correct triplet (i.e. the bottom right with blue heading).

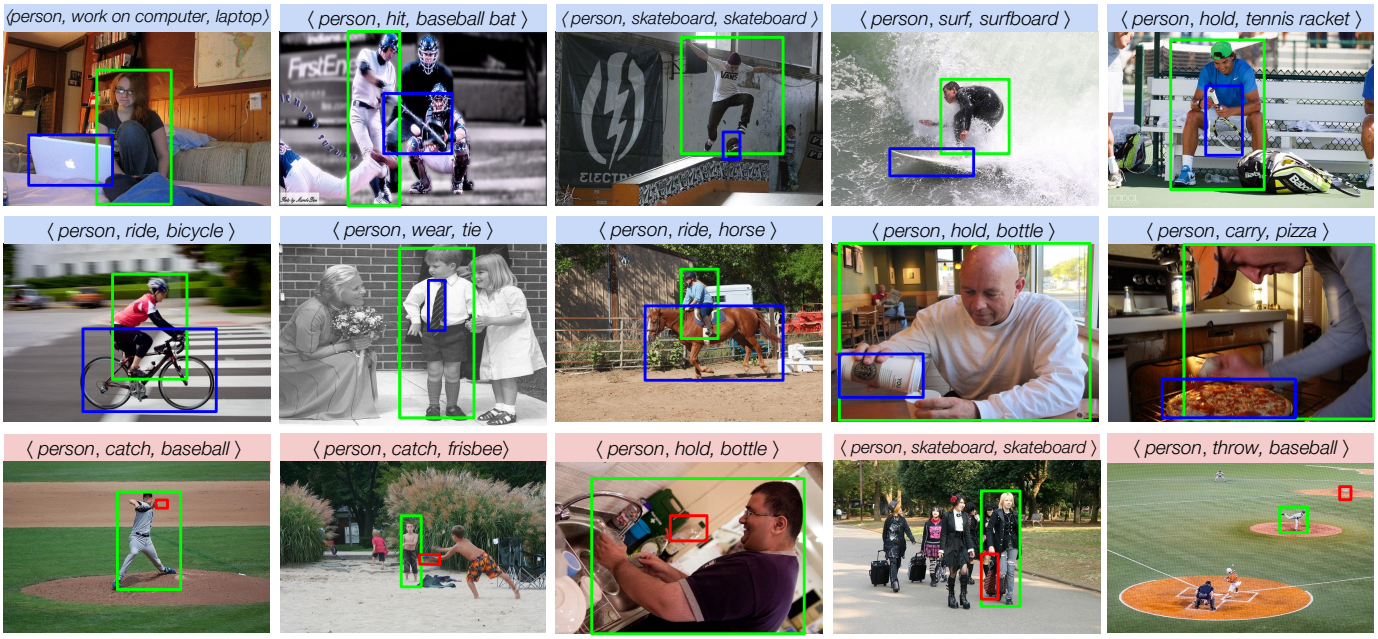

Fig. 6: Samples of human-object interactions detected by our proposed framework. Each image displays one $\langle human, action, object \rangle$ triplet. The first two rows present correct detections, and the last row presents false positives.

2) Human Pose: Human pose implicitly conveys his/her intention, which bridges the action with the interacting object. Comparing (b) to VTransE, incorporating joints distance information achieves +0.20 and +0.08 in mAP on the two datasets. Comparing method (e) to (d), the improvements are +0.16 and +0.07, respectively. It shows that HOI recognition is likely to be benefited from capturing the subtle differences of body movements. Yet the performance improvement is slight, possibly because the pose prediction could be inaccurate due to scale variation, crowding, occlusion.

An alternative implementation of human pose information has also been conducted, which directly incorporating the body joints coordinates, shown in (a) with yellow color. The advantage of (b) over (a) demonstrates the efficacy of the proposed implementation of pose information. In contrast to directly utiliforzing the locations of body joints, our proposed iHOI can capture the spatial differences of movements relative to the human and object.

3) Modeling Human Intention: Human intention can be jointly modeled using both gaze (G) and pose (P). Comparing (e) to VTransE, considering both gaze and pose achieves +2.02 and +0.85 in mAP for the two datasets.

Qualitatively, Figure 4 shows five HOI predictions with notable changes in the triplet scores after joint modeling intention using gaze and pose. The false triplet predictions (i.e. with red headings) are suppressed by incorporating human intention. For example, in the first image, it is unlikely that the detected boy is *kicking the sports ball* due to the large distances between his body joints to the target ball, as well as there is another boy nearer to the ball with a kicking pose. In the second image, the score of *drinking with cup* is significantly decreased when the model learns that the person is looking at a laptop.

Meanwhile, leveraging human intention increases the confidence of correct HOI predictions, shown by examples with blue headings. It indicates that human intention can reasonably help by leveraging the gazed context and the spatial differences of pose.

4) Hard Negative Triplet Mining: The ablation results for an alternative negative mining and the proposed one are colored with green. The proposed iHOI framework utilizes a hard negative triplet mining during training. Without the proposed mining strategy, shown in (e), mAP is decreased by -2.76 and -1.25 on the two datasets. This demonstrates that the examined hard negative samples are essential for the model to learn a more discriminative pattern.

Our method specifically targets the hard triplets that are likely to be mis-grouped, therefore outperforms the general negative mining of inaccurate localization and false classification [17], shown in (f). Our proposed mining method can be applied to other tasks that require pairing of proposals.

Figure 5 shows the effectiveness of the proposed negative triplet mining strategy for HOI detection. If no negative sampling is used, there exists interaction hallucination (i.e. *eat the other person's cake*). Model trained with the proposed strategy manages to reject the mis-grouped pairs and only predicts the correct triplet (i.e. the bottom right prediction).

E. Qualitative Examples

Figure 6 shows the examples of HOI detections generated with the proposed iHOI method, including correct predictions and false positives. The incorrect detections can be caused by confusing actions (e.g. *catch* and *throw* sports ball), inaccurate object detections (e.g. object detected on the background, false object classification or localization), and incomplete HOI annotations.

V. Conclusion

In this work, we introduce a human intention-driven framework, namely iHOI, to detect human-object interactions in social scene images. We provide an unique computational perspective to explore the role of human intention, i.e. iHOI jointly models the actor's attended contextual regions, and the differences of body movements. In addition, we propose an effective hard negative triplet mining strategy to address the mis-grouping problem. We perform extensive experiments on two benchmark datasets, and validates the efficacy of the proposed components of iHOI. Specifically, iHOI can take advantages of human gaze and pose information. Human gaze is more effective to convey intention by guiding the attended regions, whereas human pose shows less advantage.

For future work, gaze prediction on small objects could be explored, which the current model is weak at. Another direction could be studying human intention for HOI detection in videos, where intention is a more dynamic signal conveyed through multi-modality data.

Acknowledgment

This research was carried out at the NUS-ZJU SeSaMe Centre. It is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centre in Singapore Funding Initiative.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[2] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015, pp. 91–99.
[3] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 936–944.
[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[5] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Transactions on Multimedia*, 2018.
[6] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugeson, D. P. Kennedy, R. Adolphs, and Q. Zhao, "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604–616, 2015.
[7] B. Hayes and J. A. Shah, "Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks," in *ICRA*, 2017, pp. 6586–6593.
[8] C. Lu, R. Krishna, M. S. Bernstein, and F. Li, "Visual relationship detection with language priors," in *ECCV*, 2016, pp. 852–869.
[9] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Dual-glance model for deciphering social relationships," in *ICCV*, 2017, pp. 2669–2678.
[10] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *ICCV*, 2017, pp. 1270–1279.
[11] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, F.-F. Li, and C. L. Z. R. Girshick, "Inferring and executing programs for visual reasoning," in *ICCV*, 2017, pp. 2989–2998.
[12] B. Wu, J. Jia, Y. Yang, P. Zhao, J. Tang, and Q. Tian, "Inferring emotional tags from social images with user demographics," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1670–1684, 2017.
[13] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.
[14] L. Wang, X. Zhao, Y. Si, L. Cao, and Y. Liu, "Context-associative hierarchical memory model for human activity recognition and prediction," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 646–659, 2017.
[15] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *CVPR*, 2018.
[16] S. Gupta and J. Malik, "Visual semantic role labeling," *arXiv preprint arXiv:1505.04474*, 2015.
[17] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *WACV*, 2017.
[18] X. Yang, T. Zhang, and C. Xu, "Deep-structured event modeling for user-generated photos," *IEEE Transactions on Multimedia*, vol. 20, no. 8, 2018.
[19] T. Dong, S. Nishimura, and J. Liu, "Diversified and summarized video search system," in *ACM Multimedia*, 2017, pp. 1263–1264.
[20] B. Yao and F. Li, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, 2012.
[21] J. Hu, W. Zheng, J. Lai, S. Gong, and T. Xiang, "Recognising human-object interaction via exemplar based modelling," in *ICCV*, 2013, pp. 3144–3151.
[22] Y. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *ICCV*, 2015, pp. 1017–1025.
[23] B. Zhuang, Q. Wu, C. Shen, I. Reid, and A. v. d. Hengel, "HCVRD: A benchmark for large-scale human-centered visual relationship detection," in *AAAI*, 2018.
[24] B. F. Malle and J. Knobe, "The folk concept of intentionality," *Journal of Experimental Social Psychology*, vol. 33, no. 2, pp. 101–121, 1997.
[25] J. Pelz, M. Hayhoe, and R. Loeber, "The coordination of eye, head, and hand movements in a natural task," *Experimental Brain Research*, vol. 139, pp. 266–277, 2001.
[26] P. Wei, Y. Liu, T. Shu, N. Zheng, and S.-C. Zhu, "Where and why are they looking? Jointly inferring human attention and intentions in complex tasks," in *CVPR*, 2018.
[27] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, 2017.
[28] G. Gkioxari, R. B. Girshick, and J. Malik, "Contextual action recognition with R*CNN," in *ICCV*, 2015, pp. 1080–1088.
[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.
[30] H. Zhang, Z. Kyaw, S. Chang, and T. Chua, "Visual translation embedding network for visual relation detection," in *CVPR*, 2017, pp. 3107–3115.
[31] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and F. Li, "Human action recognition by learning bases of action attributes and parts," in *ICCV*, 2011, pp. 1331–1338.
[32] L. Shen, S. Yeung, J. Hoffman, G. Mori, and F. Li, "Scaling human-object interaction recognition through zero-shot learning," in *WACV*, 2018.
[33] J. J. Van Boxtel, N. Tsuchiya, and C. Koch, "Consciousness and attention: on sufficiency and necessity," *Frontiers in Psychology*, vol. 1, p. 217, 2010.
[34] M. F. Land and M. Hayhoe, "In what ways do eye movements contribute to everyday activities?" *Vision Research*, vol. 41, no. 25-26, pp. 3559–3565, 2001.
[35] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 7572, 2012, pp. 314–327.
[36] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba, "Where are they looking?" in *NIPS*, 2015, pp. 199–207.
[37] S. Gorji and J. J. Clark, "Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes," in *CVPR*, 2017, pp. 3472–3481.
[38] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015.
[39] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 1302–1310.
[40] R. B. Girshick, "Fast R-CNN," in *ICCV*, 2015, pp. 1440–1448.
[41] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, ser. Lecture Notes in Computer Science, vol. 8693, 2014, pp. 740–755.
[42] X. Chen and A. Gupta, "An implementation of faster rcnn with study for region sampling," *arXiv preprint arXiv:1702.02138*, 2017.