Interactiveness Field in Human-Object Interactions*

Xinpeng Liu[1†] Yong-Lu Li[2†] Xiaoqian Wu[1] Yu-Wing Tai[3] Cewu Lu[1‡] Chi-Keung Tang[2]
[1]Shanghai Jiao Tong University [2]HKUST [3]Kuaishou Technology
{xinpengliu0907, yuwing}@gmail.com, {yonglu_li, enlighten, lucewu}@sjtu.edu.cn, cktang@cs.ust.hk

Abstract

Human-Object Interaction (HOI) detection plays a core role in activity understanding. Though recent two/one-stage methods have achieved impressive results, as an essential step, discovering interactive human-object pairs remains challenging. Both one/two-stage methods fail to effectively extract interactive pairs instead of generating redundant negative pairs. In this work, we introduce a previously overlooked ***interactiveness bimodal prior****: given an object in an image, after pairing it with the humans, the generated pairs are either mostly non-interactive, or mostly interactive, with the former more frequent than the latter. Based on this interactiveness bimodal prior we propose the* ***"interactiveness field"****. To make the learned field compatible with real HOI image considerations, we propose new energy constraints based on the cardinality and difference in the inherent "interactiveness field" underlying interactive versus non-interactive pairs. Consequently, our method can detect more precise pairs and thus significantly boost HOI detection performance, which is validated on widely-used benchmarks where we achieve decent improvements over state-of-the-arts. Our code is available at https://github.com/Foruck/Interactiveness-Field.*

1. Introduction

Human-Object Interaction (HOI) detection consists of distinguishing human-object (H-O) pairs that have interactions from still images and classifying the interactions into various verbs. In practice, an HOI instance is represented as a triplet: $\langle human, verb, object \rangle$. Considering its important role in recent advances in robot manipulation [14], surveillance event detection [1, 30], and so on, HOI detection has been attracting continuous attention in computer vision.

Overall, HOI detection can be divided into *H/O localization*, *interactive H-O pairing*, i.e., localizing the interactive humans and objects and pairing them correctly, and

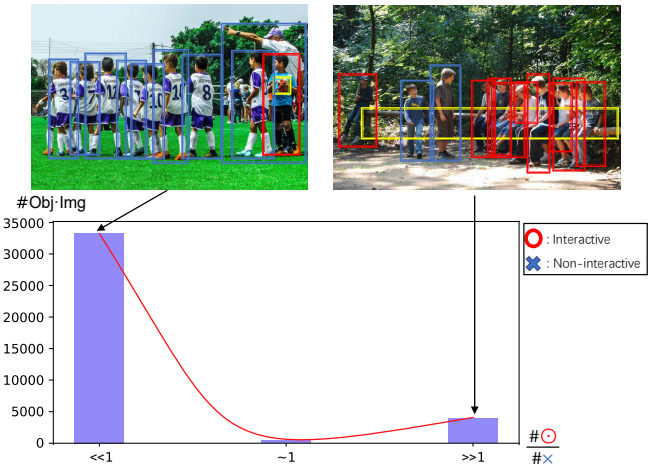

Figure 1. Distribution of interactive ratio between interactive and non-interactive H-O pairs in HICO-DET [4], where two representative samples are shown. For the pairs containing a given object (yellow), either non-interactive pairs or interactive pairs dominates, with the former much more frequent.

verb classification. The most conventional approach is the two-stage paradigm [10, 13, 22, 26, 37] proposed in HO-RCNN [5], where an object detector is first adopted to detect all the human/object instances in a given image, followed by *exhaustive pairing* and verb classification. The major issue of this straightforward approach is that, in practice, only a small portion of human/object instances are involved in HOI relationships, making the exhaustive object detection and pairing excessive and seemingly unnecessary.

The other approach consists of one-stage methods [18, 35] represented by PPDM [27]. One-stage approach adopts an *end-to-end* manner following the one-stage object detection [21, 44], where the object boxes are replaced by H-O pair boxes and the object category by HOI category. This circumvents the exhaustive instance detection and explicit pairing while achieving the same goal. However, given that a typical image, e.g., HICO-DET [4] contains 2.47 HOIs on average, it is still unsatisfactory that a recent state-of-the-art one-stage method QPIC [35] still needs **100** output pairs per image to achieve a recall of 70%.

Though significant progress has been made, the two

*The research is supported in part by the Hong Kong Research Grant Council under grant number 16201420.
†The first two authors contribute equally.
‡Corresponding author.

paradigms are still bottle-necked by **H-O pairing**: they fail to effectively extract interactive pairs but generate excessively redundant and negative pairs. One of the early studies to address this problem is TIN [23, 26], where the pairing problem is addressed by interactiveness learning. A *pairwise* interactiveness binary classifier is inserted to discriminate whether a human and an object should be paired (i.e., interactive or otherwise). Despite its simple design, the improvement is rather decent, indicating the great potential of such proper pairing strategies.

Given this early promise, here, we aim at improving HOI detection by studying the interactiveness problem from a **global** and **distribution** point of view. Specifically, we propose a previously overlooked but powerful prior: the **bimodal** property of interactiveness. In Figure 1, the dominating proportion of H-O pairs given the same object in an image are either interactive or non-interactive, while most of the time they are non-interactive. This phenomenon of interactiveness distribution is closely related to Zipf's Law [2]: *informative events are rarer than non-informative events*. To exploit this prior, we pursue a verb-agnostic measurement of interactiveness. In line with the notion of field and its global measurement [8] as such, we introduce the **"interactiveness field"** to model the global interactiveness distribution of HOI images. Specifically, we encode the H-O pairs in a complex scene as a field. Each pair is encoded as a point with an "energy" value, indicating its difference from other pairs. The field is expected to obey the bimodal prior, *i.e*., the high-energy pairs should be rare. Based on this, we analyze the change of the field with the modification on a *single* pair and impose energy constraints on the field modeling: modification on high-energy pairs should bring more salient influence. Then, the interactiveness labels are bounded with the modeled field following the prior.

To use the interactiveness field, we propose a novel paradigm. First, instead of exhaustive human/object detection, a DETR [3] structure detector is adopted to directly detect initial H-O pairs organized in an object-centric manner. Subsequently, based on the interactiveness field subjecting to the bimodal prior, we design an interactiveness field module to further filter out non-interactive pairs. Finally, the filtered pairs are fed into a verb classifier for HOI classification. On HICO-DET [4] and V-COCO [12], we achieve state-of-the-art and significant improvements.

Our contribution includes: **1)** the interactiveness bimodal prior of HOI is identified as a key to improve the H-O pair filtering and boost the HOI detection, based on which an interactiveness field model is introduced; **2)** we achieve state-of-the-art performance on widely-used HOI benchmarks.

2. Related Works

Rapid progress has recently been made in HOI learning. Many large datasets [4, 12, 20, 25] and deep learning based methods [9–11, 13, 15, 16, 19, 22, 24–26, 32, 33, 35, 37] have been proposed. For example, Chao *et al*. [4] proposed the widely-used multi-stream framework, while GPNN [33] and Wang *et al*. [38] adopted graphs to model the HOI relationship. iCAN [10] and PMFNet [37] adopted the self-attention mechanism to correlate the human, object, and context from different levels. TIN [26] introduced interactiveness to filter out non-interactive pairs. Besides, some works [19, 32, 42] focused on the relationship between HOIs. In terms of information utilization, DJ-RN [22] introduced 3D information for better inference. PaStaNet [25] introduced part states as an intermediate semantic hierarchy for further HOI reasoning. DRG [9] considered HOI from both human-centric and object-centric point of view, while VCL [15] exploited the compositional characteristic of HOI. IDN [24] analyzed how HOI is integrated and composed from a transformation-based perspective.

Recently, several one-stage methods have been proposed [7, 27, 35, 39], where parallel HOI detectors directly detect HOIs triplets, in contrast to the conventional two-stage method [10, 26] for interaction prediction. PPDM [27], UnionDet [7], and IP-Net [39] adopted a variant of one-stage object detector [21, 44] for HOI detection.

While based on the recently proposed transformer detector DETR [3], QPIC [35] managed to achieve impressive performance. By capitalizing on the powerful transformer, DETR [3] achieved impressive performance without many hand-designed components. A fixed-size set of predictions is produced in a single pass through the decoder. The main loss is calculated by matching the predicted and ground-truth predictions via an optimal bipartite matching, followed by imposing the specific losses. QPIC [35] adapted the paradigm by regressing both the human and object box with the addition of a verb classifier to detect HOI triplets.

3. Methods

Our goal is to address the pairing problem in HOI detection, by exploiting the underlying distributional information of H-O pairs subject to the interactiveness bimodal prior. Section 3.1 first presents the preliminaries of our method and a formal definition of interactiveness field. Then, in Section 3.2, we introduce how interactiveness field is modeled with the pair distributional characteristics. In Section 3.3, we demonstrate how to design the practical system.

3.1. Preliminaries

Given an image $\mathcal{I}$, we define *interactiveness field* $\mathcal{F}$ as

$$\mathcal{F} = (\mathcal{A} \times \mathcal{A}, E(\cdot) : \mathcal{A} \times \mathcal{A} \rightarrow [0, 1]), \tag{1}$$

where $\mathcal{A}$ denotes arbitrary areas in $\mathcal{I}$, $E(\cdot)$ is the *energy* function for each area pair, indicating the *relative difference* of each pair against other pairs. Given the interactiveness

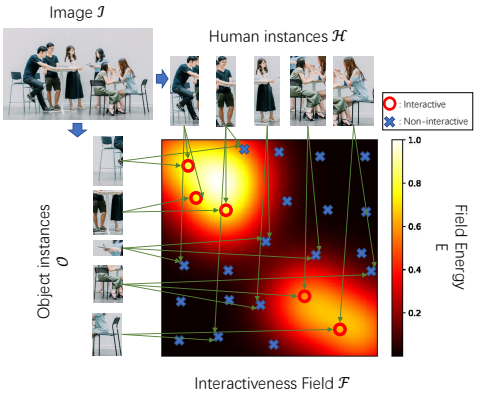

Figure 2. Interactiveness field illustration.

bimodal prior, the energy function is closely related to the interactiveness: *when the pairs are mostly non-interactive, interactive pairs would possess high energy and vice versa.*

Since we focus on HOI detection, where only human/object instances are considered to be potentially interactive, the definition in Eq. 1 is simplified as

$$\mathcal{F} = (\mathcal{P} = \mathcal{H} \times \mathcal{O}, E(\cdot) : \mathcal{P} \rightarrow [0, 1]), \tag{2}$$

where $\mathcal{H}, \mathcal{O}$ are the human and object instance proposals in $\mathcal{I}$ respectively, as illustrated in Figure 2.

Here, we focus on the pairs concerning the same given object o_i. Each pair $\langle h_i, o_i \rangle$ is represented by the extracted feature $f^i_{\mathcal{P}} \in f_{\mathcal{P}}$, and $E(\cdot)$ is implemented by specially designed neural networks. Thus, the interactiveness field $\mathcal{F}$ could be generally formulated as

$$\mathcal{F} = (f_{\mathcal{P}}, E(\cdot)),\ f_s = g(f_{\mathcal{P}}), \tag{3}$$

where f_s denotes the **summary** of the field extracted from the pairs with summary function $g(\cdot)$, the energy function $E(\cdot)$ takes the sample feature $f^i_{\mathcal{P}}$, producing the energy of the input sample. Intuitively, the binary pair-wise classifier introduced in TIN [26] could be a simple implementation of $E(\cdot)$, lacking the consideration of global interactiveness distribution and pair difference. However, in Section 4, we show that without the interactiveness bimodal prior, the simple TIN-style classifier outputs a biased interactiveness score thus performs unsatisfactorily on interactiveness discrimination. That is, for almost all the pairs in an image involving the same object, near-zero interactiveness score is produced due to the extreme imbalance in data distribution. Rather than resorting to simple modeling using a pair-wise classifier, we propose to model the interactiveness field regulated by the interactiveness bimodal prior, considering the underlying global-distribution properties.

3.2. Interactiveness Field Modeling

In the following, we first delve into how interactiveness field is modeled in Section 3.2.1 subject to the interactiveness bimodal prior. Notably, two main constraints are derived in Section 3.2.2 to regulate the field, where the global

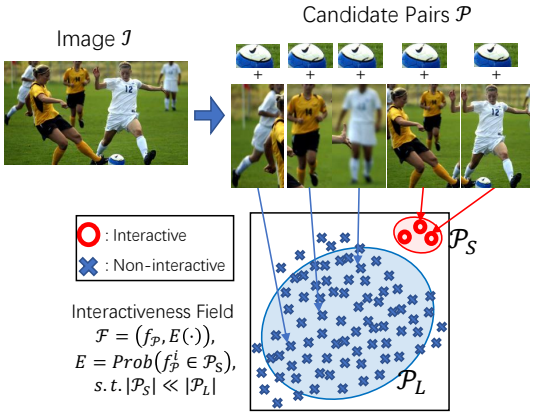

Figure 3. Interactiveness field modeling subject to the interactiveness bimodal prior.

change in $\mathcal{F}$ upon removing or modifying a single local pair will be analyzed. The modeling formulation detailed in Sections 3.2.1–3.2.2 only requires the interactiveness bimodal prior. In Section 3.2.3, we describe how the interactiveness labels can then be incorporated into the formulation to enhance the proposed field modeling.

3.2.1 Cardinality Constraint

As illustrated in Figure 1, candidate pairs involving the same object can be divided into two clusters: the *rare, high-energy* cluster and the *frequent, low-energy* cluster. Correspondingly, we argue that the interactiveness field should possess the following property: candidate pairs set $\mathcal{P}$ should consist of two diverse sets $\mathcal{P}_S$ and $\mathcal{P}_L$ with salient differences in cardinality. This property is formulated as

$$\begin{aligned} &\mathcal{P} = \mathcal{P}_L \cup \mathcal{P}_S, \\ &\text{s.t. } \mathcal{P}_L \cap \mathcal{P}_S = \emptyset,\ |\mathcal{P}_S| \ll |\mathcal{P}_L|, \end{aligned} \tag{4}$$

where $|\cdot|$ denotes cardinality. The interactiveness field is

$$\begin{aligned} &\mathcal{F} = (f_{\mathcal{P}}, E(\cdot)),\ E(f^i_{\mathcal{P}}) = Prob(\mathcal{P}^i \in \mathcal{P}_S), \\ &\text{s.t. } |\mathcal{P}_S| \ll |\mathcal{P}_L|. \end{aligned} \tag{5}$$

Thus, given the extracted pair feature $f_{\mathcal{P}} \in \mathcal{R}^{N \times C}$, the summary function $g(\cdot)$ first extracts the two clusters $\mathcal{P}_S$ and $\mathcal{P}_L$, denoted by centroids $c_s, c_l \in \mathcal{R}^C$ and assignment vectors $A_s, A_l \in \mathcal{R}^N$, where A^i_s, A^i_l respectively mean the probability that pair i belongs to cluster $\mathcal{P}_S, \mathcal{P}_L$, subjecting to $\sum_i A^i_s \ll \sum_i A^i_l$. $f_s = (c_s, c_l)$ is then adopted as the summary representation of the interactiveness field $\mathcal{F}$. The energy function $E(\mathcal{P}^i) = A^i_s$ for each pair $\mathcal{P}^i$ is given by the probability that the pair belongs to $\mathcal{P}_S$. Figure 3 illustrates the formulation:

$$c_s, c_l, A_s, A_l = g(f_{\mathcal{P}}), \text{s.t. } \sum_i A^i_s \ll \sum_i A^i_l. \tag{6}$$

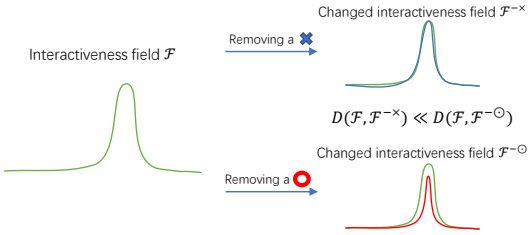

Figure 4. Field change against pair removal. Removal rare pairs (usually also interactive) brings more salient change.

To regulate the field to satisfy the interactiveness bimodal prior, a cardinality loss L_{card} is formulated as

$$L_{\text{card}} = \sum_i A_s^i - \sum_i A_l^i. \tag{7}$$

The loss corresponds to the constraint $\sum_i A_s^i \ll \sum_i A_l^i$, which encourages more pronounced cardinality difference. Noticeably, here we do not need the binary interactiveness labels [26] in modeling. Thus, the above modeling can be regarded as an **unsupervised** process using our bimodal prior. In Section 3.2.3, we introduce how to further enhance the interactiveness discrimination with the binary labels.

3.2.2 Field Change Constraints

The cardinality constraint introduced above focuses on the *static* status of the interactiveness field. We now investigate how to model the field by observing how $\mathcal{F}$ should change upon modifying local pairs with different energy level.

Field Change against Pair Removal. We first explore how the global field representation changes when a certain sample is removed. Starting from the interactiveness field $\mathcal{F}$ in Section 3.2.1, we can tell the removal of a high-energy point would affect the overall representation of $\mathcal{F}$ more than the removal of a low-energy point (Figure 4). So we adopt a difference indicator D_r to encode the global field change when a certain sample is removed, which is formulated as

$$\begin{aligned} D_r^i &= D(\mathcal{F}, \mathcal{F}^{-i}), \\ \mathcal{F} &= (f_\mathcal{P}, E(\cdot)),\ \mathcal{F}^{-i} = (f_\mathcal{P}^{-i}, E(\cdot)), \end{aligned} \tag{8}$$

where $D(\cdot,\cdot)$ denotes the difference between the two fields, and $f_\mathcal{P}^{-i} = f_\mathcal{P} / f_\mathcal{P}^i$ denotes the pair features minus $f_\mathcal{P}^i$.

Based on this, given the pair feature $f_\mathcal{P}$, $g(\cdot)$ (defined in Section 3.2.1) first extracts the field summary representation $f_s = (c_s, c_l)$ for $\mathcal{F}$. Then, each pair i is removed, and the rest pair features $f_\mathcal{P}^{-i}$ are fed to $g(\cdot)$, which produces the modified field representation f_s^{-i}. The L2 distance between f_s and f_s^{-i} is then defined as the difference indicator D_r^i. Larger D_r^i indicates that the pertinent pair is more likely to have higher energy level (or more different from the other pairs). The above process can be summarized as

$$\begin{aligned} f_s &= (c_s, c_l) = g(f_\mathcal{P}), \\ f_s^{-i} &= (c_s^{-i}, c_l^{-i}) = g(f_\mathcal{P}^{-i}), \\ D_r^i &= \|f_s, f_s^{-i}\|_2. \end{aligned} \tag{9}$$

Since the removal of a pair will definitely change the field, instead of enforcing D_r to be zero for frequent low-energy pairs, a rank loss L_{rank}^r is imposed as

$$\begin{aligned} L_{\text{rank}}^r &= \sum_{i\in\mathcal{P}_S} \sum_{j\in\mathcal{P}_L} D_r^j - D_r^i, \\ \mathcal{P}_S &= \{i : A_s^i > A_l^i\},\ \mathcal{P}_L = \{i : A_l^i > A_s^i\}, \end{aligned} \tag{10}$$

where A_l, A_s are the assignment vectors produced by $g(f_\mathcal{P})$. L_{rank}^r only encourages the assumed high-energy pairs to cause more field change with their removal than the low-energy pairs.

Field Change against Pair Modification. Another worthwhile constraint to explore is how the field changes when a pair is modified, in our case, replaced by the **mean** pair representation. Still referring to the interactiveness field $\mathcal{F}$ in Section 3.2.1, given a field with most areas possessing low energy, we could tell that the mean representation of this field should also carry low energy. Thus, if we replace a high-energy pair with the mean, the overall field representation should change significantly. On the other hand, the overall field representation should not change much when a low-energy pair is replaced by the mean. Thus, we can obtain another difference indicator D_m as

$$\begin{aligned} D_m^i &= D(\mathcal{F}, \hat{\mathcal{F}}^i), \\ \mathcal{F} &= (f_\mathcal{P}, E(\cdot)),\ \hat{\mathcal{F}}^i = (\hat{f}_\mathcal{P}^i, E(\cdot)). \end{aligned} \tag{11}$$

$\hat{f}_\mathcal{P}^i$ denotes $f_\mathcal{P}$ replacing $f_\mathcal{P}^i$ with mean representation $\bar{f}_\mathcal{P}$.

To implement the above, the field representation f_s is first extracted by $g(\cdot)$ in Section 3.2.1. Then we obtain the modified field $\hat{f}_s^i$ by feeding $\hat{f}_\mathcal{P}^i$ to $g(\cdot)$. The difference between f_s and $\hat{f}_s^i$ is defined as the difference indicator $D_m^i = \|f_s - \hat{f}_s^i\|$. Again, larger difference indicates

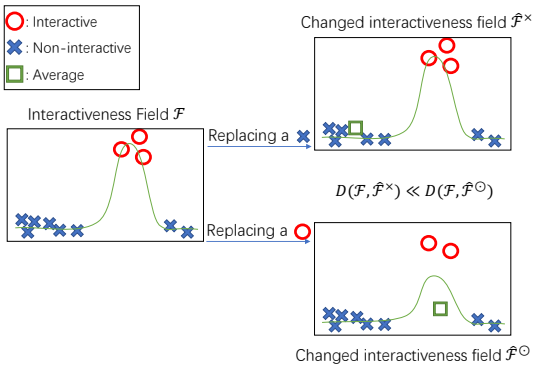

Figure 5. Field change against pair modification. Modification on rare pairs (usually also interactive) brings more change.

the sample is more likely to be a high-energy pair. The rank loss L^m_{rank} with the same formulation as Eq. 10 is computed.

3.2.3 Binding with Interactiveness Labels

The previous modeling formulation only adopts the interactiveness bimodal prior, functioning in an unsupervised manner. For further enhancement, we can bind the field with the interactive semantics via specially designed losses to connect interactiveness labels transferred from HOI labels, following TIN [26]. This encourages the modeled field to simultaneously approach the ground truth distribution while following the prior when applicable.

Following the set-based training procedure in QPIC [35], the interactiveness labels are assigned to the candidate pairs. Given the assigned labels, we obtain the correspondence between $\{\mathcal{P}_S, \mathcal{P}_L\}$ and {interactive pairs, non-interactive pairs}. In the following, we assume $\mathcal{P}_S$ is interactive for ease of description, which is most of the cases. Analogous descriptions apply when $\mathcal{P}_L$ is interactive. A simple cross entropy loss L_{ce} is imposed on A_s, A_l. Then, the cardinality loss in Section 3.2.1 is enriched with an additional term:

$$L_{\text{card}} = \sum_i A^i_s - \sum_i A^i_l + \|n_T - \sum_i A^i_s\|, \tag{12}$$

where n_T is the number of interactive pairs for this object in this image. This added term regulates the cardinality of $\mathcal{P}_S$ to be the same as the number of interactive pairs. Moreover, a clustering loss L_{clus} inspired by [34] is formulated as

$$\begin{aligned} p_{ij} &= A^i_s A^j_s + A^i_l A^j_l, \\ L_{\text{clus}} &= \sum_{i,j} \left((\alpha_{ij} - 1)\log(1 - p_{ij}) - \alpha_{ij} \log p_{ij}\right), \end{aligned} \tag{13}$$

where $\alpha_{ij} = 1$ if pair i, j are both interactive or non-interactive, otherwise $\alpha_{ij} = 0$. This loss encourages pairs with the same interactiveness label to be clustered together.

With these losses, we force the field $\mathcal{F}$ to simultaneously follow the interactiveness bimodal prior while approaching the ground-truth interactiveness distribution. More discussions on the generalization of our interactiveness bi-modal prior would be included in the supplementary material.

3.3. Practical System Design

Next, we introduce how the interactive field is incorporated into a practical HOI detection system. Such system contains four components: visual feature extractor, pair decoder, interactiveness field module defined in Section 3.2, and verb classifier. Figure 6 shows the overall pipeline.

3.3.1 Visual Feature Extractor

Our feature extractor is a combination of a CNN and a transformer encoder. In detail, given an image $\mathcal{I} \in \mathcal{R}^{H\times W\times 3}$, the CNN encodes it into feature map $f_C \in \mathcal{R}^{H'\times W'\times C_C}$, which is linearly projected to a lower dimension of C_T, flattened into $\mathcal{R}^{(H'W')\times C_T}$, which is then fed into the transformer encoder with sinusoidal positional embedding $E \in \mathcal{R}^{(H'W')\times C_T}$ to output the final visual feature $f \in \mathcal{R}^{(H'W')\times C_T}$. The CNN encoder aggregates the local information into patch tokens, while the transformer encoder, leveraging the power of multi-head self-attention, generates a feature map with rich global contextual information.

3.3.2 Pair Decoder

A transformer decoder is adopted as the pair decoder. With visual feature f as K, V, a learned query embedding $Q \in \mathcal{R}^{M\times C_T}$ is utilized to decode the candidate pairs $\mathcal{P}$ along with feature $f_{\mathcal{P}}$. A fully-connected layer is imposed on $f_{\mathcal{P}}$ to classify the corresponding object class o, and two two-layer MLPs regress the human and object box coordinates b^h, b^o. Following previous set-based training process [3, 35], with the Hungarian bipartite matching algorithm, ground truth labels are assigned to the pair predictions. Multiple loss items are computed, including generalized IoU (Intersection over Union) loss $L^h_{\text{giou}}, L^o_{\text{giou}}$, box regression L1 loss $L^h_{\text{reg}}, L^o_{\text{reg}}$, and object class cross-energy loss L_o. The pair decoder is first trained along with the visual feature extractor with target loss

$$L_{\text{pair}} = \lambda_1(L^h_{\text{giou}} + L^o_{\text{giou}}) + \lambda_2(L^h_{\text{reg}} + L^o_{\text{reg}}) + \lambda_3 L_o, \tag{14}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weighting coefficients.

3.3.3 Implementation of Interactiveness Field Module

To implement the interactiveness field module, multiple choices for $E(\cdot)$ and $g(\cdot)$ are proposed. A toy design is first used, where $E(\cdot)$ and $g(\cdot)$ are implemented as a hierarchical cluster followed by a *soft* two-means cluster with the hierarchical centroids as the initial centroids. For both clustering procedures, Euclidean distance is adopted. By "soft", we mean the distance vectors $D_s, D_l \in \mathcal{R}^N$ are respectively processed by a softmax function along each column to obtain the assignment vectors A_s, A_l.

For a more advanced version, the two-means clustering is replaced by a modified multi-head attention layer. In detail, it takes $f_{\mathcal{P}}$ as K, V, and the two hierarchical cluster centroids as Q to extract C. To obtain the assignment matrix, the original softmax function used to generate attention from logits is replaced by sigmoid function following averaging, where the attention value before averaging is adopted as assignment matrix A. In this way, the multi-head attention module is adapted for clustering by regarding the attention mechanism as a soft assignment procedure, thus acquiring a more powerful mean field representation. The

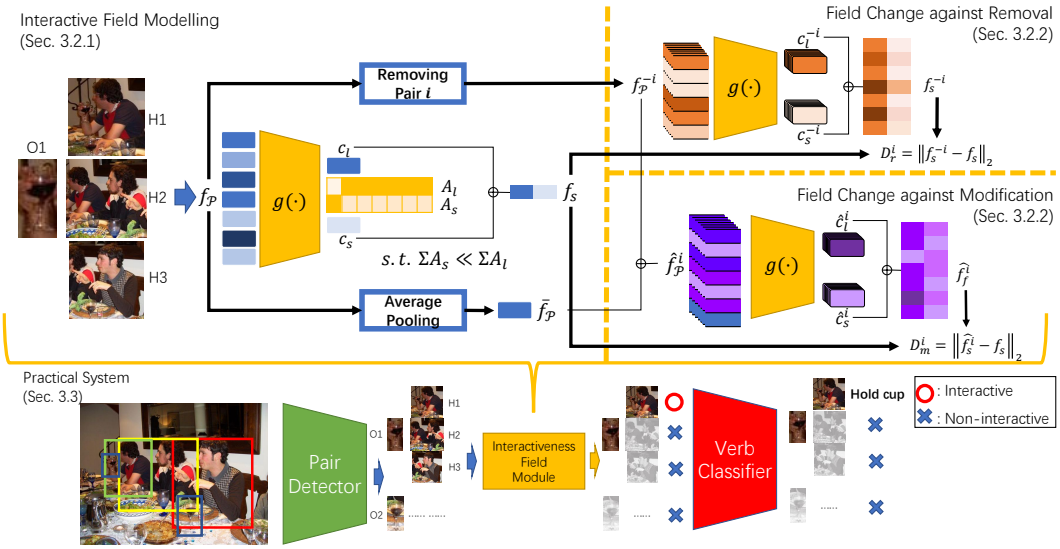

Figure 6. Our pipeline for HOI detection with interactiveness field modeling which is composed of four components. Visual feature extractor generates visual feature map f, based on which pair decoder decodes the candidate pairs $\mathcal{P}$ along with feature $f_{\mathcal{P}}$. Our proposed interactiveness field module models the interactiveness field and assigns interactiveness score S_b for each pair. The verb decoder infers the verb score S_v for generating the final score as $S = S_v \cdot S_b$.

target loss is formulated as

$$L_{\text{field}} = \lambda_4 L_{\text{card}} + \lambda_5 L_{\text{ce}} + \lambda_6 L_{\text{clus}} + \lambda_r (L^r_{\text{rank}} + L^m_{\text{rank}}), \quad (15)$$

where $\lambda_4, \lambda_5, \lambda_6, \lambda_r$ are weighting coefficients, and the different loss terms have already been defined in Section 3.2.

3.3.4 Verb Decoder

Another transformer decoder takes f (whole image feature) as K, V, $f_{\mathcal{P}}$ as Q, followed by a fully-connected verb classifier, which is used to produce the verb score S_v. The verb classifier is attached with verb label cross-energy loss L_{verb}.

3.3.5 Training and Inference on HOI Datasets

The training is divided into three stages. First, we train the pair decoder along with the visual feature extractor using L_{pair}. Then, the interactiveness field module is introduced and the three components are fine-tuned together with loss $L = L_{\text{pair}} + L_{\text{field}}$. Finally, the verb classifier is included, and the whole system is trained with $L = L_{\text{pair}} + L_{\text{field}} + L_{\text{verb}}$.

In some cases interactive pairs dominate, e.g., in a restaurant, several humans are *sitting beside* the dinner table except for the waiter. We consider this special situation in training. Since these cases only account for less than 10% in HICO-DET [4], we assume that the interactive pairs are always minorities in inference. Thus, the energy and difference indicators can be directly adopted to compute interactiveness binary score S_b. The difference indicators are aggregated and normalized to $[0, 1]$, and then combined with A_s, producing $S_b = (A_s + (\sigma(D_r) + \sigma(D_m) - 1))/2 \in [0, 1]$, where $\sigma(\cdot)$ is sigmoid function. The final prediction is constructed as $(b^h, b^o, o, S) \in \mathcal{P}_r$, where $S = S_v \cdot S_b$. Our experimental results show even with this compromised strategy, the improvement is still substantial.

Notwithstanding, a possible problem is that though the interactiveness bimodal prior is statistically reasonable, there still exist exceptions, e.g., an image contains only one person. For the practical system here, we cover the situation with sparse scene in two ways. First, the human proposals $\mathcal{H}$ generated by the model are abundant most of the time, making the prior still applicable. Second, pairs with the same object category are aggregated and modeled by the same field, as they share similar interactiveness patterns.

4. Experiments

4.1. Dataset and Metric

We adopt two large-scale HOI detection benchmarks: HICO-DET [4] and V-COCO [12] for evaluation. **HICO-DET [4]** consists of 38,118 training images, 9,658 testing images, 600 HOI categories (comprising of 80 COCO [28] objects and 117 verbs), and more than 150 K annotated HOI pairs. We use mAP for evaluation: true positive is required to contain accurate human and object locations (box IoU with reference to GT box is larger than 0.5) and accurate interaction classification. Following [4], mAP for three sets: Full (600 HOIs), Rare (138 HOIs), Non-Rare (462 HOIs) under both Default and Known Object modes are reported. **V-COCO [12]** contains 10,346 images (2,533 in train set, 2,867 in validation set, and 4,946 in test set), and covers 29 verb categories (25 HOIs and 4 body motions) and 80 objects from COCO [28]. We use role mean average precision under both scenario 1 and scenario 2 as evaluation metrics, where only the 25 HOIs are taken into consideration.

Method	mAP Default ↑			mAP Known Object ↑		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
iCAN [10]	14.84	10.45	16.15	16.26	11.33	17.73
TIN [26]	17.03	13.42	18.11	19.17	15.51	20.26
PMFNet [37]	17.46	15.65	18.00	20.34	17.47	21.20
DJ-RN [22]	21.34	18.53	22.18	23.69	20.64	24.60
PPDM [27]	21.73	13.78	24.10	24.58	16.65	26.84
VCL [15]	23.63	17.21	25.55	25.98	19.12	28.03
DRG [9]	24.53	19.47	26.04	27.98	23.11	29.43
IDN [24]	26.29	22.61	27.39	28.24	24.47	29.37
Zou *et al*. [45]	26.61	19.15	28.84	29.13	20.98	31.57
ATL [16]	28.53	21.64	30.59	31.18	24.15	33.29
AS-Net [6]	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [35]	29.07	21.85	31.23	31.68	24.14	33.93
FCL [17]	29.12	23.67	30.75	31.31	25.62	33.02
GGNet [43]	29.17	22.13	30.84	33.50	26.67	34.89
SCG [41]	31.33	24.72	33.31	34.37	27.18	36.52
CDN [40]	31.78	27.55	33.05	34.53	29.73	35.96
Ours	**33.51**	**30.30**	**34.46**	**36.28**	**33.16**	**37.21**

Table 1. Results on HICO-DET [4]. The first part adopted COCO pre-trained detector. HICO-DET fine-tuned or one-stage detector is used in the second part. All the results are with **ResNet-50**.

Method	AP_{role}(Scenario 1)	AP_{role}(Scenario 2)
iCAN [10]	45.3	52.4
TIN [26]	47.8	54.2
VSGNet [36]	51.8	57.0
IDN [24]	53.3	60.3
HOTR [18]	55.2	64.4
QPIC [35]	58.8	61.0
CDN [40]	62.3	64.4
Ours	**63.0**	**65.2**

Table 2. Results with **ResNet-50** on V-COCO [12].

4.2. Implementation Details

We adopt ResNet-50 followed by a six-layer transformer encoder as our visual feature extractor. The pair decoder and the verb decoder are both implemented as a six-layer transformer decoder. During training, AdamW [29] with the weight decay of 1e-4 is used. The visual feature extractor and pair decoder are initialized from COCO [3] pre-trained DETR [3]. The query size is set as 64 for HICO-DET [4] and 100 for V-COCO [12] following CDN [40]. The loss weight coefficients $\lambda_1, \lambda_2, \lambda_3$ are respectively set as 1, 2.5, 1, exactly the same as QPIC [35]. The visual feature extractor and pair decoder are fine-tuned for 90 epochs with a learning rate of 1e-4 which is decreased by 10 times at the 60th epoch. Then, the interactiveness field module is introduced and fine-tuned for another 9 epochs with learning rate of 1e-4. Finally, the verb decoder is added and the whole model is trained for 30 epochs. All experiments are conducted on four NVIDIA GeForce RTX 3090 GPUs with batch size of 16. In inference, a pair-wise NMS with threshold of 0.6 is conducted. That is, low-score predictions with both human and object IoU > 0.6 compared to the same category high-score pair is suppressed.

4.3. Results

Results on HOI Detection Benchmarks We first report the results on HICO-DET [4]. Table 1 compares our methods with previous state-of-the-art methods. We outperform all of them with Default Full mAP of **33.51**. Even compared with methods like ATL [16] which adopted additional object attribute information, we achieve an impressive advantage of **4.98** mAP. When comparing to other transformer-based methods such as HOTR [18], [45], AS-Net [6], QPIC [35], and CDN [40] our method manages to attain relative improvements of **30.2%**, **16.1%**, **15.3%**, and **5.4%**, respectively. To fully verify the effectiveness of our method, we also adopt the very recent CDN [40] and outperform it significantly. Note that even compared with CDN-L [40] (Default Full mAP 32.07) with more parameters, our model still maintains a significant advantage.

Table 2 compares our result on V-COCO [12] with those of previous state-of-the-arts, which indicates that our method achieves impressive advantage over previous methods with **63.0** and **65.2** mAP under Scenario 1 and 2.

Results on Interactiveness Detection To better demonstrate our contribution to the H-O pair filtering, we evaluate our interactiveness detection [26] on HICO-DET [4].

First, following the interactiveness AP proposed in [26], we evaluate our interactiveness detection, comparing with open-source state-of-the-arts [26, 27, 35, 40]. In detail, we adopt S_b as the interactiveness score for our model. For TIN [26], the inherent interactiveness score is adopted. For PPDM [27], QPIC [35], and CDN [40], the mean of 520 HOI scores is used as an approximation.Table 3 tabulates the results, which shows the interactiveness AP of TIN is significantly lower, echoing our analysis that it suffers from the mass of exhaustively generated negative H-O pairs even with the non-interaction suppression [26]. In terms of the one-stage PPDM [27] directly detecting H-O pairs, the performances are better since the avoid of exhaustive pairing. Surprisingly, the interactiveness performance gap between QPIC [35] and CDN [40] is negligible, while our method demonstrates to be considerably better than previous methods with interactiveness AP of **37.39**.

To verify that our method is superior on pair filtering, we select previous open-source state-of-the-arts and compare the Default Full mAP in a Top-k manner [27] in Table 4. That is, we only select the predictions with top-k confidence for each image. Even with only 5 predictions per image, the advantage is still impressive over other methods.

Furthermore, we explore how our pair filtering can boost the performance of two-stage methods. Following CDN [40], we feed the representative two-stage method iCAN [10] (using exhaustive pairing without pair filtering) with our detected pairs, and compare the result produced by feeding exhaustive pairs as input. In addition, the results using CDN [40] and QPIC [35] pairs as input are also compared. Here, mAP under Default mode for the three sets (Full, Rare, Non-Rare) are reported. Table 5 shows that the performance of iCAN is significantly boosted with the pairs of high-quality, especially of the ones from our method.

	TIN [26]	PPDM [27]	QPIC [35]	CDN [40]	ours
AP	14.35	27.34	32.96	33.55	**37.39**

Table 3. Interactiveness detection on HICO-DET [4].

Methods	Top-5	Top-10	All
PPDM [27]	18.92	20.35	21.10
QPIC [35]	29.07	29.29	29.07
CDN [40]	30.19	30.40	31.78
Ours	**32.65**	**33.07**	**33.51**

Table 4. Top-K result on HICO-DET [4]. "All" indicates Top-100 for PPDM [27] and QPIC [35], and Top-64 for CDN [40] and ours.

Methods	Full	Rare	Non-Rare
iCAN [10]	14.16	12.26	14.73
iCAN [10]$^{\text{QPIC}}$	21.78	13.18	24.35
iCAN [10]$^{\text{CDN}}$	24.05	18.32	25.76
iCAN [10]$^{\text{Ours}}$	**26.07**	**21.03**	**27.58**

Table 5. Performance of iCAN [10] on HICO-DET [4] with different pair detection. Superscripts indicate the source of pair detection, where no superscript indicates the exhaustive pairing [10].

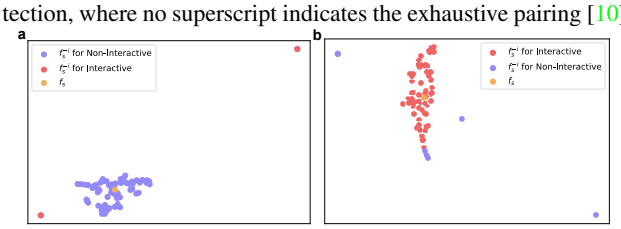

Figure 7. Field change visualization. f_s (orange) is the field summary feature, while f_s^{-i} of non-interactive pairs (purple) are in majority in the left; f_s^{-i} of interactive pairs (red) are in majority in the right. As shown, f_s^{-i} of minority pairs locates far from f_s.

	Full	Rare	Non-Rare
Ours	**33.51**	**30.30**	**34.46**
w/o IFM	30.54	26.04	31.88
w/o S_b	33.30	29.76	34.35
$g(\cdot)$ via FC	30.70	25.68	32.20
$g(\cdot)$ via clustering	30.97	26.86	32.20
cardinality only	32.38	27.99	33.69
field change only	32.76	28.82	33.94
Unsup-IFM	31.62	27.38	32.88

Table 6. Ablation studies on HICO-DET [4].

4.4. Visualization

Figure 7 visualizes the field change under the constraints (Section 3.2.2). The field summary feature f_s and the changed summary feature f_s^{-i} of different pairs are visualized with t-SNE [31], where f_s^{-i} corresponding to minority pairs follow the constraints well, validating our design.

4.5. Ablation Studies

We conduct ablation studies on HICO-DET [4] under the Default mode, with the results in Table 6.

First, we show how the model is influenced if the interactiveness field module (IFM) is removed. The considerable mAP drop of 2.97 validates the key role of IFM. We then reveal the influence of interactive score S_b on performance. We find that removing S_b only results in a minor drop. This demonstrates that the IFM functions more than merely in results fusion: it also contributes to feature learning.

Dataset	$\frac{\#inter}{\#no-inter} \ll 1$	$\frac{\#inter}{\#no-inter} \approx 1$	$\frac{\#inter}{\#no-inter} \gg 1$
w/o IFM	0.38	0.55	2.34
$g(\cdot)$ via FC	0.32	0.57	2.12
$g(\cdot)$ via clustering	0.28	0.51	2.09
Ours	**0.19**	**0.42**	**1.88**

Table 7. Error of #interactive pairs between prediction and GT.

Second, different implementations of IFM are compared. Replacing IFM with a fully-connected layer as done in TIN [26], we obtain 30.70 mAP ($g(\cdot)$ via fully-connected in Table 6), which is slightly better than removing IFM while still insignificant. By implementing $g(\cdot)$ via clustering as proposed in Section 3.3, we achieve a marginal improvement compared to a model w/o IFM, far below the advanced version of $g(\cdot)$, showing the efficacy of our design. This experiment on the other hand shows the importance of the bimodal prior even with a straightforward $g(\cdot)$ implementation. Moreover, we evaluate the influence of different constraints. With only cardinality constraint (Section 3.2.1), we suffer 1.13 mAP drop (cardinality only in Table 6). While the mAP drop is 0.75 if only field change constraints (Section 3.2.2) are preserved (field change only in Table 6).

Third, we demonstrate the performance of IFM operating in the unsupervised mode, referred to as Unsup-IFM. That is, we zero out the loss items proposed in Section 3.2.3. Then, IFM is only restrained by the bimodal prior. Even without supervision using interactiveness labels, we can achieve good improvement with only the bimodal prior.

Moreover, we validate IFM by the error between the number of predicted and GT interactive pairs per image of different implementation of $g(\cdot)$. The predicted interactive pair number is calculated by summing the predicted interactive probability of each pair. The results in Table 7 show the advanced implementation does exploit the prior. The impressive gap with and w/o IFM proves that the raw *data-driven* methods fail to model the bimodal distribution well.

Finally, we demonstrate the performance under different interactive ratios. The IFM brings relative improvement as 9.23% (30.68 to 33.52), 0.11% (52.98 to 53.04), 3.04% (51.42 to 52.98), respectively with interactive ratio $\ll 1, \approx 1, \gg 1$. These show our impressive improvement upon valid cases and ignorable harm on invalid cases. For more limitation and social impact discussion, please refer to the supplementary.

5. Conclusion

This paper focuses on previously overlooked interactiveness bimodal prior in HOI learning. To utilize this prior, the interactiveness field is proposed and modeled. Multiple properties of the proposed field are explored to match the learned field and realistic HOI scenes. Our method effectively discriminates interactive human-object pairs and achieves significant improvements, validated on widely-used benchmarks. Though interactiveness field prompts H-O pairing and boosts HOI detection, we believe the room for H-O pairing is still large and needs more explorations.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):555–560, 2008. 1

[2] Felix Auerbach. Das gesetz der bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59:74–76, 1913. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, 2020. 2, 5, 7

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 1, 2, 6, 7, 8

[5] Yu Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 1

[6] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. Reformulating hoi detection as adaptive set prediction. In *CVPR*, 2021. 7

[7] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In *AAAI*, 2018. 2

[8] Richard P Feynman, Robert B Leighton, and Matthew Sands. The feynman lectures on physics; vol. i. *American Journal of Physics*, 33(9):750–752, 1965. 2

[9] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 2, 7

[10] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 1, 2, 7, 8

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

[12] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 6, 7

[13] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, appearance and layout encodings, and training techniques. In *ICCV*, 2019. 1, 2

[14] Bradley Hayes and Julie A Shah. Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6586–6593. IEEE, 2017. 1

[15] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. *ECCV*, 2020. 2, 7

[16] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021. 2, 7

[17] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 7

[18] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, 2021. 1, 7

[19] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. *arXiv preprint arXiv:2007.08728*, 2020. 2

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 2

[21] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018. 1, 2

[22] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 1, 2, 7

[23] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Xijie Huang, Liang Xu, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. *TPAMI*, 2021. 2

[24] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, and Cewu Lu. Hoi analysis: Integrating and decomposing human-object interaction. In *NeurIPS*, 2020. 2, 7

[25] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 2

[26] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 1, 2, 3, 4, 5, 7, 8

[27] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 1, 2, 7, 8

[28] 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

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[30] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2013. 1

[31] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[32] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting rare visual relations using analogies. In *ICCV*, 2019. 2

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

[34] Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Lsd-c: Linearly sepa-

rable deep clusters. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1038–1046, 2021. 5

[35] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 1, 2, 5, 7, 8

[36] Oytun Ulutan, ASM Iftekhar, and BS Manjunath. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, 2020. 7

[37] 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. 1, 2, 7

[38] Hai Wang, Wei-shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. *arXiv preprint arXiv:2010.10001*, 2020. 2

[39] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020. 2

[40] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *arXiv preprint arXiv:2108.05077*, 2021. 7, 8

[41] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, 2021. 7

[42] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for human-object interaction detection. In *ECCV*, 2020. 2

[43] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, 2021. 7

[44] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1, 2

[45] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 7