

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Efficient human–object-interaction (EHOI) detection via interaction label coding and Conditional Decision

Tsung-Shan Yang^{a,*}, Yun-Cheng Wang^a, Chengwei Wei^a, Suya You^b, C.-C. Jay Kuo^a

^a University of Southern California, Los Angeles, 90089, CA, USA ^b DEVCOM Army Research Laboratory, Adelphi, 20783, MD, USA

ARTICLE INFO

Communicated by Mohamed Daoudi

Dataset link: https://github.com/keevin60907/ EHOI---Efficient-Human-Object-Interaction-Det ector

Keywords: Human–object interaction (HOI) detection Error correction code Green learning Image understanding

ABSTRACT

Human–Object Interaction (HOI) detection is a fundamental task in image understanding. While deeplearning-based HOI methods provide high performance in terms of mean Average Precision (mAP), they are computationally expensive and opaque in training and inference processes. An Efficient HOI (EHOI) detector is proposed in this work to strike a good balance between detection performance, inference complexity, and mathematical transparency. EHOI is a two-stage method. In the first stage, it leverages a frozen object detector to localize the objects and extract various features as intermediate outputs. In the second stage, the first-stage outputs predict the interaction type using the XGBoost classifier. Our contributions include the application of error correction codes (ECCs) to encode rare interaction cases, which reduces the model size and the complexity of the XGBoost classifier in the second stage. Additionally, we provide a mathematical formulation of the relabeling and decision-making process. Apart from the architecture, we present qualitative results to explain the functionalities of the feedforward modules. Experimental results demonstrate the advantages of ECC-coded interaction labels and the excellent balance of detection performance and complexity of the proposed EHOI method. The codes are available: https://github.com/keevin60907/EHOI—Efficient-Human-Object-Interaction-Detector.

1. Introduction

Human–Object Interaction (HOI) detection is essential for image understanding (Chao et al., 2018; Gupta et al., 2019). The labels in HOI datasets are triplets in the form of <Human-Interaction-Object>. The detector needs to know not only the bounding boxes of the human and objects but also the class labels of the objects and the types of interaction. HOI detection is a human-centric application and can be further applied to the image understanding task, such as visual question answering and image captioning. HOI detection can be challenging, as illustrated in Figs. 1 and 2. The examples are taken from a popular HOI detection dataset, HICO-DET (Chao et al., 2018). Images may contain the label '*no_interaction*', but not every '*no_interaction*' human–object pair is labeled. Furthermore, some images may share the same verb in different scenarios, known as verb polysemy (Zhong et al., 2021).

Another challenge in HOI detection is the imbalanced distribution of interaction pairs in the training samples. As shown in Fig. 3, the HICO-DET dataset exhibits a skewed distribution of annotations, leading to highly biased predictions. To prevent overfitting on the relationship annotations, we propose a hybrid coding scheme to address this problem. That is, we partition interaction pairs into rare and non-rare cases.

For non-rare cases, we adopt the traditional one-hot coding. For rare cases, we group them into one super-class and then adopt binary codes and error correction codes (ECCs) to encode these rare cases. The repartitioning and data relabeling mitigate the imbalanced distribution in the original dataset. This is one of the significant contributions of this work.

Prevailing HOI detection solutions can be categorized into onestage and two-stage methods. One-stage methods (Lim et al., 2023; Chen and Yanai, 2023; Ma et al., 2023; Kim et al., 2021) are trained via end-to-end optimization of specific neural network architectures, where the objective function relates to the <human-interaction-object> labels. The models fit the training dataset. Yet, labels in the dataset are not aligned perfectly with human understanding. An imperfect label is shown in Fig. 1, which indicates that the algorithms may be biased due to the imbalanced labels, making them difficult to interpret. Two-stage methods (Zhang et al., 2022, 2021; Hou et al., 2021) include object detection and relationship prediction steps, namely, (1) identifying where humans and objects are located and (2) determining the type of interaction between them. In the first stage, it conducts object detection using a pre-trained object detector and extracts various features from

* Corresponding author. E-mail address: tsungsha@usc.edu (T.-S. Yang).

https://doi.org/10.1016/j.cviu.2025.104390

Received 6 November 2024; Received in revised form 18 January 2025; Accepted 13 May 2025 Available online 22 May 2025 1077-3142/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



Fig. 1. Illustration of challenges in the HOI problem with images from the HICO-DET dataset: images labeled as 'no_interaction. The missing labels are for the background people and the detected objects.

input images. In the second stage, it leverages the first-stage outputs, such as object classes, bounding boxes, spatial relationships, etc., to predict the interaction type. Deep Learning (DL) models are widely applied in both stages. However, when deployed on edge applications, the cascaded model structures are computationally expensive. Typically, one-stage models outperform two-stage models in detection accuracy at the expense of larger model sizes and higher training/inference complexities. Fig. 4 visualizes the model complexity and inference cost. On the other hand, the decoupled modular design makes two-stage models more straightforward to understand.

Fig. 4 compares the model performance, model size, and computational complexity of several state-of-the-art (SOTA) HOI detectors. The model performance is represented by mAP (%) in the *y*-axis, the model size is represented by the number of model parameters in the *x*-axis, and the computational complexity is represented by inference floating-point operation (FLOP) numbers, respectively. This figure shows one-stage transformer-based models (QAHOI Chen and Yanai, 2023 and ERNet Lim et al., 2023) and one-stage models based on interaction point prediction (IP-Net Wang et al., 2020 and PPDM Liao et al., 2020). We also visualize two SOTA two-stage methods — UPT (Zhang et al., 2022) and SCG (Zhang et al., 2021).

One-stage models achieve astonishing detection performance. However, the end-to-end optimization algorithm conceals the reasoning processing in the high-dimensional numerical latent spaces. Though the attention heatmaps provide visual clues, the numerical vectors do not correspond to interpretable semantic representations. Besides, Zhu et al. (2023) pointed out that one-stage models could be biased in detection results under a skewed data distribution. On the other hand, two-stage methods have concrete intermediate results. Starting from the given conditions, the second-stage interaction prediction is an extension of the semantic understanding.

Aiming at interpretability and lower carbon footprints, we propose a new two-stage method called Efficient HOI (EHOI) in this work. As shown in Fig. 4, its mAP performance is worse than other two-stage models, UPT and SCG, yet its FLOP number is significantly lower. As reported in Section 4, the FLOP number of EHOI is 4500 times smaller than SCG and 15,800 times smaller than UPT per query. EHOI outperforms some one-stage models (IP-Net and PPDM) and underperforms others (QAHOI and ERNet) in mAP. However, it has tremendous advantages in the model size and FLOP numbers. Regarding carbon footprint (FLOP number) and memory (model size), EHOI offers an attractive AI/ML solution for mobile and edge devices. EHOI extends green learning (Kuo and Madni, 2023) to the HOI detection problem. The detector aims to provide a transparent decision-making process and maintain low power consumption regarding the number of floating point operations.

The contributions of this work are summarized below.

- EHOI offers comparable mAP performance with significantly lower (1/15,800x) inference complexity for the HICO-DET dataset.
- EHOI is explainable. All of its modules can be interpreted as probability estimators. We formulate the learning process as the aggregation of conditional probabilities.
- The data distribution in the HOI dataset is analyzed. We shed light on individual modules with visual examples and explain modules' decoupling statistically.

2. Related work

2.1. One-stage HOI detection

One-stage HOI detectors are inspired by one-stage object detectors (Ren et al., 2015; Carion et al., 2020). Apart from the human and object bounding boxes, researchers define the interaction vectors from the center of the human bounding box to the corresponding object bounding box. Starting from the detector backbones, deep learning models can perform HOI detection with auxiliary interaction points or vectors. Wang et al. (2020) cropped the intersection of the proposed human and object bounding boxes as the interaction region. The model can predict the internship label by exploiting the overlapped features. Liao et al. (2020) proposed a model optimized by the auxiliary output of predicting the interaction points. Leveraging the spatial information, the supervised loss from interaction points helps the model obtain better features and remove unlikely interactions.

With the thriving visual transformers (ViTs), transformer-based detectors (Zhu et al., 2020; Jia et al., 2023) have been developed. The encoder-decoder structure conducts the detection by trainable query tensors. To merge the information from human and object bounding boxes, Kim et al. (2021) developed a DETR (Carion et al., 2020) backbone and combined it with pairwise human/object queries for the interaction decoder. The merge human and object queries are interaction queries for further decoding. Chen et al. (2021) introduced auxiliary interaction vector prediction to optimize the interaction decoder. Tamura et al. (2021) used the Hungarian algorithm (Kuhn, 1955) to calculate the loss of matched human/object pairs in the loss function. By exploiting overlapped patches between humans and objects, the auxiliary output can utilize human priors' supervision by minimizing the loss function.

Besides the decoding process, utilizing the detectors in conjunction with pre-trained models significantly enhances performance. Liao et al. (2022) combined the SOTA language and visual transformer, CLIP (Radford et al., 2021). Matching the textual embeddings and the interaction representations improves the relation decision significantly. Inspired by the deformable operations (Dai et al., 2017; Xia et al., 2022), Chen and Yanai (2023) replaced the ResNet backbone with deformable transformers. The backbone can focus on the objects not limited to the square grids. To enrich the presentations, Lim et al. (2023) used the EfficientNet as the backbone to extract multi-scale features. Yuan et al. (2023) consider the interaction as semantic labels by leveraging Large Language Models (LLMs). Sun et al. (2024) introduce the human guidance of the language concept perception to further improve the detector performance.

Although one-stage models offer SOTA HOI detection performance, they have shortcomings. First, the training of one-stage models is highly



Fig. 2. Illustration of challenges in the HOI problem with images from the HICO-DET dataset: three images with the same verb, 'wash,' but humans behave differently. From left to right: 'wash airplane', 'wash apple', and 'wash carrot'.



Fig. 3. The occurrence of <human-interaction-object> labels, where if the number of annotations exceeds 1000, it is clipped to 1000 for simplicity. Half of the relationship labels in the HICO-DET dataset have fewer than 200 annotations.



Fig. 4. Complexity comparison between the proposed EHOI and several other stateof-the-art (SOTA) detectors for the HICO-DET dataset, where the *x*-axis is the model size in the log scale, the *y*-axis is mAP (%), and the bubble size is proportional to the inference FLOP numbers.

dependent on the dataset. Zhu et al. (2023) pointed out that onestage models could be biased in detection results under a skewed data distribution. Second, it is challenging to interpret one-stage methods since the semantic information in images is hidden in numerous cascading latent spaces. Researchers indirectly analyzed the models using convolutional filter responses and attention matrices in Visual Transformers (ViTs). Third, they suffer from a large model size and high computational complexity.

2.2. Two-stage HOI detection

HOI is a human-centric classification task. The linkage between

human and object representations is crucial. With the stagewise algorithm, humans can understand the decision-making process clearly. In two-stage models, the first stage extracts various human and object representations. In contrast, the second stage is a multi-pair (i.e., humanobject pairs) and multi-label (i.e., interaction labels) classification problem.

To understand the human representations, Gupta et al. (2019) used pose estimation models to obtain semantic information. The addition pose-estimation model provides a three-dimensional spatial understanding. By exploiting the human–object correlation, Hou et al. (2020) constructed a model to yield human and object streams and handled the relation between the two streams based on the co-occurrence of <human-relation-object> triplets. The triplets formation can be optimized by the composition learning (Kato et al., 2018). The merging process removes the impossible triplets and enhances the probabilities of possible triplets.

The human-object relationship can also be formulated as a graph, where human and object features can be viewed as the vertices in the graph. Hence, the interaction detection can be formulated as an edge production problem. Gao et al. (2020) proposed a dual structure to model the relation. It combined the human-centric and object-centric graphs to predict the relation. The relationship is the weighted sum of the two sub-graphs. Zhang et al. (2021) incorporated the coordinate information between objects in the graph convolution structure to capture spatial information within images. The spatial information can be defined as the vector between the human and object bounding boxes and their intersection. With the development of transformers, Zhang et al. (2022) adopted the FFN decoder structure for the pairwise relation classification. The decoder takes the features of the region of interest from the first stage as the inputs and performs the classification.

2.3. Green learning

Kuo and Madni (2023) proposed a statistical-based learning framework called Green Learning (GL) to address the increasing computational burdens of DL networks. The GL paradigm does not have neurons, neural networks, and end-to-end optimization via backpropagation. Instead, it adopts a feedforward and modular design in both training and inference based on data statistics. The whole processing pipeline is purely data-driven and transparent.

The green learning paradigm is composed of representation learning and decision learning. Representation learning is conducted by the transformations of the image and spectrum domains. Kuo and Chen (2018) proposed the Subspace approximation with augmented kernels (Saak) Transform to construct the features from sampled patches. The transformation is a variant of the Principal Component Analysis (PCA). Saak transformation compacts the information from the sample patches and keeps the signs of responses without nonlinear functions. Kuo et al. (2019) further proposed Subspace approximation with adjusted bias (Saab) Transform extending from the Saak transform. Saab transform divides the information into DC and AC terms and recenters the responses using the sample means. The Saab and Saak transformations are calculated from the eigenvalue decomposition, which purely depends on the sample statistics.

Decision learning is finding the correlations between representations and labels. Yang et al. (2022) proposed a Discriminant Feature Test (DFT) to remove the redundant features from the raw features set. DFT computes the loss between a single dimension feature and the given labels. The more discriminant the feature, the more the loss function can be minimized. Aiming for explainability, GL methods adopt XGBoost classifiers (Chen and Guestrin, 2016) in the decision-making stage. The XGBoost classifier can lower the cross-entropy loss iteratively through a sequence of boosting trees to achieve loss optimization. In addition, the tree structure can demonstrate the importance of the input features and provide the statistical reasoning process from the features to the output labels.

The GL solution addresses environmental concerns by reducing FLOPs to relieve power consumption and carbon footprint. Our work follows this principle, as detailed in the next section.

3. EHOI method

3.1. System overview

The system diagram of the proposed EHOI method is shown in Fig. 5, which is a two-stage method. The first stage is a pre-trained object detector, where we select DETR (Carion et al., 2020) as the object detector. It uses ResNet50 as the backbone and achieves good object detection performance trained by suitable object detection datasets. Since our main contributions lie in the second stage, we will emphasize the data processing pipeline of the second stage in this section. It consists of the following four tasks in cascade.

Module A: Visual Features Construction

We utilize the Region of Interest (RoI) alignment and pooling (Girshick, 2015) to generate human and object representations. It yields an input query pair that contains human and object features and their associated spatial information.

Module B: Hybrid Interaction Coding

We propose a hybrid coding scheme to address the training samples' imbalanced distribution of interaction pairs. That is, we partition interaction pairs into rare and non-rare cases. For nonrare cases, we adopt the traditional one-hot coding. For rare cases, we group them into one super-class and then adopt binary codes with error correction codes (ECCs) to encode rare cases within the super-class.

- Module C: Discriminant Features Selection
- The discriminant feature selection process is conducted based on the interaction codes. We identify a subset of discriminant features against every bit assignment of the interaction type.
- Module D: Conditional Decision on the Interaction Type The final prediction is the aggregation of the probabilities outputs from each interaction bit.

The model in the second stage is efficient regarding the number of model parameters and Floating-point Operations (FLOP) numbers. Modules B–C are statistics-based, allowing interpretability. Furthermore, applying error correction codes (ECCs) to encode interaction labels is a novel contribution to the AI/ML literature. Its advantages are demonstrated in the experiments section.

3.2. Processing modules in the second stage

3.2.1. Module A: Visual features construction

Constructing a rich feature set for human and object representations is critical. Intuitively, the relative distances and other scenarios between the human and object locations in images are helpful for HOI detection. Utilizing the first-stage model, we can capture the features of corresponding regions by RoI pooling and alignment (Girshick, 2015). The human and object features can be obtained from aggregating different layer feature maps in the detector. The relative spatial information includes the interaction vector and the relative sizes of human and object bounding boxes. The interaction vector is the difference between human- and object-bounding box centers. It can be represented in Euclidean or polar coordinate systems.

Furthermore, the background information can be extracted from the whole image features obtained from the backbone network. To summarize, the features under consideration comprise human RoI features, object RoI features, relative spatial features, and whole image features. We also split human and object representations as individual queries and determine their spatial features accordingly. Some detailed descriptions are illustrated in Fig. 5. Yet, we should point out that these features are far from perfect since they lack precise semantic information, and discriminant features may be concealed in noisy training samples.

To deal with imbalanced labels, our model fits subsets of the interaction samples instead of the whole dataset. The classifiers are trained by subsets containing a common object for human queries. That is, the desired outputs of a classifier can be denoted as

$$P(relation \mid human) = \sum_{c \in \{Object\}} P(relation, object = c \mid human),$$
(1)

where *c* denotes an object type and {*Object*} denotes the whole object set. The clustering process repartition the skewed data distribution into subset distribution. Under the constraint, *object* = *c*, we can reduce the overall long-tail distribution to a few short-tail conditional distributions. Similarly, the classifier's performance for object query can also be improved by conditional probabilities. We employ clustering algorithms such as KMeans to create subsets and use them to train classifiers in each subset. We can assign a pseudo-label for each subset, and then the classifier can be formulated similarly. The ultimate classifier for object queries can be obtained by combining multiple subset classifiers in a weighted manner. We can aggregate the results of both human and object queries to make relation decisions in the inference stage.

3.2.2. Module B: Hybrid interaction coding

The foundation of modern machine learning models is to capture the distribution in the training dataset and generalize it to unseen samples. Finding a representation space that generalizes well between training and testing samples is essential. Representations could be highly diversified, and the labeled data may possess a long-tail distribution in



Fig. 5. The overall system diagram of the proposed EHOI. Its first stage uses a pre-trained object detector circled by blue dashed lines to obtain the object classes and locations. The object detector is kept frozen when training modules of the second stage. The main contributions of EHOI lie in the data processing pipeline in the second stage, which is circled by red dashed lines. It consists of four modules: (A) visual features construction, (B) interaction label coding, (C) discriminant features selection, and (D) conditional decision on the interaction type. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

real-world applications. Adopting the one-hot vector to represent the classes of interest in the context of AI/ML is typical. There are two problems with the one-hot representation. First, if the class number is large, the dimension of these one-hot vectors can be high. Second, the labeled data possess a long-tail distribution, as mentioned above. Take the HOI benchmark, HICO-DET, as an example. It has 600 interaction triplets. However, 138 have less than ten samples and are called rare cases. If we adopt the one-hot encoding scheme for all, we need 600-dimensional vectors to represent them and have to train 600 one-versue-the-rest binary classifiers. The classification performance of each rare case is expected to be poor due to high data imbalance since it is challenging for a classifier to learn from less than ten samples among more than 10^6 queries.

To handle this challenge, we merge several rare cases into a superclass and adopt the traditional one-hot coding to encode non-rare cases plus this super-class. Then, we adopt the binary coding scheme to differentiate rare cases inside the super-class. Instead of training from less than ten positive samples, the super-class gathers several categorical labels, which prevents training a complicated classifier to fit on imperfect data distribution. To compare the difference between the one-hot and binary coding, we take a 4-class classification problem as an example. The four classes are {1000,0100,0010,0001} in the one-hot encoding and as {00,01,10,11} in the binary coding. Each bit represents a binary split. We can train four binary classifiers for the one-hot coding that handles the one-versus-the-rest classification problem. For the binary coding, we only train two binary classifiers. The first one separates {00,01} from {10,11} based on the first bit while the second one splits {00, 10} from {01, 11} based on the second bit. Each binary classifier can be viewed as a decoder. Nevertheless, each classifier may have mistakes, leading to wrong aggregated results. We use error correction codes (ECC) to enhance the robustness of a remedy. To follow the above example, we can assign three-bit codewords to them, i.e., {000,011,101,110}. After adding the error correction bit, every codeword pair has a Hamming distance of 2 (i.e., have two different bits). Here, we use Hamming codes (Hamming, 1950) to improve the performance of straightforward binary codes and ensure that each representation differs from others with a Hamming distance of no less than 3 in EHOI.

The performance of four coding schemes is compared in Table 1 for the HICO-DET dataset. They are one-hot codes, binary codes, Hamming codes, and a hybrid coding scheme. The hybrid coding scheme adopts the one-hot coding for non-rare cases, the super-class of all rare cases, and the Hamming codes for rare cases. The table shows that the hybrid coding scheme achieves the best results, with a higher mAP value than one-hot codes. It is also worthwhile to point out that Hamming codes

Table 1

Performance comparison between four coding schemes for interaction labels in mAP (%). Under the same architecture, the Hamming codes perform best for rare cases, while the one-hot codes offer the best performance for non-rare cases. The hybrid coding schemes yield the best overall performance.

Methods	Default		Model size	
	Full	Rare	Non-Rare	
EHOI (one-hot codes)	20.55	13.47	22.66	56.4M
EHOI (binary codes)	16.35	7.97	18.86	15.4M
EHOI (Hamming codes)	19.19	14.50	20.59	27.7M
EHOI (hybrid)	24.53	19.26	26.09	45.9M

have a smaller model size, which helps reduce the model size of the hybrid scheme.

3.2.3. Module C: Discriminant features selection

Each bit representation in Hamming codes corresponds to a partition of labeled interactions into two sets. In other words, we relabel interactions of rare cases into two types denoted by 0 and 1, respectively. If we do not change the features, this relabeling will become a subset of the original coding scheme. Hence, each bit assignment prompts the utilization of a feature selection module to identify the pivotal feature and exclude others. Such dimension reduction can prevent overfitting the estimator with the low-variance features.

We must select discriminant features for a given binary label to facilitate the classifier in the next module. This can be achieved by applying the Discriminant Feature Test (DFT) (Yang et al., 2022) to all input features individually. For a given 1D input feature, we place the feature value of each labeled training sample in a line segment bounded by the range of the maximum and minimum values, as shown in Fig. 6. Then, we search for the optimal partition point on this line segment to minimize the loss function, which is defined as the weighted sum of binary cross-entropies of the left and right partitions.

$$H = -\sum_{i \in \{\text{partition}\}} y_i \log(y_i), \tag{2}$$

where H denotes the binary entropy of the partitioned samples and y represents the samples' labels.

$$L = \frac{N_{left}H_{left} + N_{right}H_{right}}{N_{left} + N_{right}},$$
(3)

where *L* denotes the loss of a partition point, and N_{left} , N_{right} , H_{left} , H_{right} denote the number of the samples and the binary cross-entropy in the left and right partition, respectively. A feature is more discriminant



Fig. 6. Visualization of DFT, where pink and orange dots represent the "0" and "1" binary labels, and the loss function is the weighted cross-entropy sum of samples in the left and right parts of the partition line.

if it has a lower loss value, reducing the number of miss-classified cases. Then, we can plot the loss value curve from the lowest to the highest and use the elbow point to select discriminant features from the whole feature set.

However, the detection dataset accompanies plenty of false detections. The imbalance distribution also occurs in the HOI detection. To enhance the correlation between features and positive labels, we change the binary cross-entropy to focal loss (Lin et al., 2020).

$$H^{f} = -\alpha(1-p)^{\gamma}\log(p), \tag{4}$$

where H^f denotes the focal loss in the partition, and $\alpha, \gamma \in \mathbf{R}$ are hyperparameters. The exponential parameter can emphasize the hard positive cases. Accordingly, the loss of the partition points can be denoted as the weighted sums of the two focal losses in the left and right partition.

$$L^{f} = \frac{N_{left}H_{left}^{J} + N_{right}H_{right}^{J}}{N_{left} + N_{right}}.$$
(5)

3.2.4. Module D: Conditional decision on the interaction type

We divide the desired decisions into sequential subproblems instead of training a complex classifier for the skewed data distribution. Each subproblem can be expressed clearly, step by step. First, we attempt to maximize the conditional probability of an interaction (or relation) conditioned on the human and object representations, which can be written as

$$\begin{aligned} P(relation \mid human, object) \\ &= P(relation \mid human) * \frac{P(object \mid relation, human)}{P(object \mid human)} \\ &= P(relation \mid object) * \frac{P(human \mid relation, object)}{P(human \mid object)} \\ &\sim \alpha P(relation \mid human) + \beta P(relation \mid object), \end{aligned}$$

where α , β are learnable parameters. The conditional probability estimator can be decoupled into the linear combination of two simpler conditional estimators. As shown in Section 4.4.3, the human and object queries provide different types of interaction predictions. Then, the two conditional probabilities in the last equation can be further expressed as

$$P(relation|human) = \sum_{c \in \{Object\}} P(relation, object = c|human)$$

$$P(relation|object) = \sum_{d \in \{Human\}} P(relation, human = d|object),$$
(7)

where *c* and *d* are class labels and {*Object*}, {*Human*} are the object sets and clustered human representations. Furthermore, the probability estimator can be decomposed into the piecewise conditional probability by using the clustering results in the previous stage. Suppose we use a bit stream, $B = (b_0, b_1, \dots, b_{n-1})$, to represent a relation. Then, we would like to maximize P(relation, object = c|human) and P(relation, human = d|object), which are called the human query and the object query,

respectively. The conditional probability of human queries can be written as

P(relation, object = c | human)

$$= P(B, object = c|human)$$

$$= \bigcap_{0 \le i < n} P(b_i, object = c|human).$$
(8)

The sequential decoding process is equivalent to aggregation of the label prediction, which is described in module D in Section 3.2.4. The conditional probability of object queries can be found in the same manner.

All the probability estimators in EHOI are XGBoost (Chen and Guestrin, 2016). XGBoost is a state-of-the-art classification algorithm and can demonstrate the feature importance by the tree node construction, making the overall pipeline more transparent. For each bit classifier, we set the number of estimators and the depth of the tree to 300 and 3, respectively. The aggregation of the bit stream prediction is conducted by Linear Discriminant Analysis (LDA).

4. Experiments

4.1. Datasets

(6)

V-COCO (Gupta and Malik, 2015) and HICO-DET (Chao et al., 2018) are two commonly used HOI detection datasets. V-COCO is a subset of the MS-COCO dataset. It contains 2533 training images, 2867 validation images, 4946 test images, and 24 actions. HICO-DET is larger than V-COCO. It comprises 37,633 training images, 9546 test images, 117 actions, and 600 interactions for various action–object pairs. Its training set has 117,871 human–object pairs with annotated bounding boxes, while its testing set contains 33,405 such pairs. HICO-DET is a challenging dataset. The 600 labeled interactions can be divided into 138 rare cases and 462 non-rare cases. Rare cases have less than ten samples in the training set. We use the mean Average Precision (mAP) as the evaluation metric, which is the mean of the average precision of all classes.

The HICO-DET dataset has two settings: Default and Known Object. In the Default setting, the algorithms must simultaneously label human and object bounding boxes, object type, and relationship. In the Known Object setting, the algorithms already have object information and only need to label human and object bounding boxes and their relationship.

In the V-COCO setting, there are two scenarios: S1 and S2. Scenario 1 (S1) applies to test cases with missing role annotations. An agent role prediction is correct if the action is accurate, the overlap between the person boxes is >0.5, and the corresponding role is empty. Scenario 2 (S2) also applies to test cases with missing role annotations. An agent role prediction is correct if the action is correct and the overlap between the person boxes is >0.5 (ignoring the corresponding role). This scenario is suitable for cases with roles outside the COCO categories.

Table 2

Detection performance comparison of SOTA two-stage models in mAP (%) for the HICO-DET dataset under the default and known object settings and for the V-COCO dataset. The model sizes in the parameter number (M) are also compared, where the numbers are taken from Lim et al. (2023).

Architecture	Method	$Param(M)(\downarrow)$	Backbone	Default(†)		Known Object(†)		V-COCO(↑)			
				Full	Rare	Non-Rare	Full	Rare	Non-Rare	AP^{S1}_{role}	AP^{S2}_{role}
			Two-Sta	ge Metho	İs						
	No-Frill (Gupta et al., 2019)	72.3	ResNet152	17.18	12.17	18.08	-	-	-	-	-
Multi-Stream PMFNet (Wan et al., 2019) ACP (Bansal et al., 2020) PD-Net (Zhong et al., 2021) VCL (Hou et al., 2020)	49.3	ResNet50	17.46	15.65	18.00	20.34	17.47	21.20	-	-	
	-	ResNet101	21.96	16.43	23.62	-	-	-	53.2	-	
	PD-Net (Zhong et al., 2021)	-	ResNet152	22.37	17.61	23.79	26.86	21.70	28.44	52.0	-
	-	ResNet50	23.63	17.21	25.55	25.98	19.12	28.03	48.3	-	
	RPNN (Zhou and Chi, 2019)	-	ResNet-50	17.35	12.78	18.71	-	-	-		
Graph-Based VSGNet (Ulutan et al., 202 DRG (Gao et al., 2020) SCG (Zhang et al., 2021)	VSGNet (Ulutan et al., 2020)	84.9	ResNet-152	19.80	16.05	20.91	-	-	-	51.8	57.0
	DRG (Gao et al., 2020)	46.1	ResNet50-FPN	21.66	19.66	22.25	-	-	-	51.0	-
	SCG (Zhang et al., 2021)	53.9	ResNet50-FPN	29.26	24.61	30.65	32.87	27.89	34.35	54.2	60.9
Green Learning	EHOI (Ours)	45.9	ResNet50-FPN	24.53	19.26	26.09	27.64	22.70	29.12	50.8	56.3

Table 3

Detection performance comparison in mAP (%) between EHOI and three end-to-end trained one-stage methods (without finetuning object and action detectors individually) for the HICO-Det dataset, where the mAP results of HOTR, AS-Net, and QPIC are taken from Ma et al. (2023).

Methods	Default					
	Full	Rare	Non-Rare			
HOTR (Kim et al., 2021)	23.46	16.21	25.65			
AS-Net (Chen et al., 2021)	24.40	22.39	25.01			
QPIC (Tamura et al., 2021)	24.21	17.51	26.21			
EHOI (ours)	24.53	19.26	26.09			

Table 4

The comparison of FLOP numbers per query between EHOI and two SOTA two-stage models.

Two-stage models	Default Parameters		FLOPs	
	Full		(per query)	
SCG (Graph)	29.26	53.9M (1.2x)	54M (4,500x)	
UPT (Transformer)	32.62	54.7M (1.2x)	190M (15,800x)	
EHOI (Ours)	24.53	45.9M (1x)	12K (1x)	

4.2. Parameter settings

We used the pre-trained DETR (Carion et al., 2020) model as the first-stage object detector provided by SCG (Zhang et al., 2021) and UPT (Zhang et al., 2022). The human and object representations are the ROI pooling features from the CNN backbones. The feature selection processes are set to choose the top 1000 discriminant features from the human and object queries. The focal loss in the feature selection module is set to $\alpha = 1.0$ and $\gamma = 2.0$.

The conditional classifiers are based on the HICO-DET dataset in the second stage. The 600 interaction triples can be divided into 80 objects and 117 relations. The possible relation classes range from 2 to 16 by giving the object class. Hence, the Hamming codes can be represented by four binary bits with three error correlation bits. The bit classifiers are the XGBoost classifiers (Chen and Guestrin, 2016) with 700 estimators and a depth of 5. The aggregation processes are conducted by Linear Discriminant Analysis (LDA). The one-hot coding classifiers follow the same setting as the bit classifier.

4.3. Experimental results

4.3.1. Performance benchmarking against other two-stage models

We compare the performance of EHOI against other SOTA two-stage models, including multi-stream and graph-based models, in Table 2, where the top and the second performers are in bold and underlined, respectively. EHOI achieves the second-best mAP values in most categories for HICO-DET. Our EHOI model also has the smallest number of learnable parameters. It is important to note that EHOI relies solely on visual features and does not incorporate external word embeddings or pose estimation information during training.

DRG (Gao et al., 2020) uses extra language models in the training and prediction, which increases the computational burden in the training stage. Besides, graph-based HOI models need iterations of operations in graph convolutional networks, leading to a higher computation complexity, which will be discussed at the end of this subsection. SCG (Zhang et al., 2021) provides a similar mAP performance, but the classifier is a black box. The hidden layers are trained through end-to-end optimizations. However, our EHOI can give a statistical reasoning process for every intermediate classifier. The decision process is feedforward and can provide a sensible decision-making process for human understanding.

4.3.2. Performance benchmarking against one-stage models

Transformer models achieve impressive performance in various computer vision tasks at the expense of high computational complexities in training and inference. In the encoder–decoder-based detector, the model requires auxiliary queries for detection. There is no rule of thumb to determine the hyperparameters of the queries and save on computation requirements. Besides model efficiency, the training process is nontrivial for transformer-based one-stage HOI models.

Especially, HICO-DET is a dataset with a long-tailed distribution. The model will become biased if we enforce the model to fit the skewed label distribution. The number of rare cases with less than ten samples has more than 10^6 labeled pairs. The performance of one-stage models would drop dramatically if no finetuning were conducted on object detection and relation detection individually. Compared to the object labels, the interaction labels are incomplete 1. Hence, the imbalance labels will make the interaction detector the biased objected detector.

To verify the biased one-stage detectors, we compare the performance of our EHOI method and those without delicate finetuning in Table 3. It was observed in Zhu et al. (2023) that imbalanced triples tend to decrease the object detection performance. In contrast, our twostage EHOI method does not need iterative finetuning on object and interaction detection. Thus, our architecture is more robust in terms of imbalanced triples. It can retain the performance of object detectors and plug-on relation detection features.

4.3.3. Comparison of computational complexity and carbon footprint

One of the critical aspects of our approach is the computational cost, with Green Learning explicitly focusing on energy consumption in the algorithms. Our claim of being "Green" is verified by comparing the computational cost of EHOI, UPT (Zhang et al., 2022), and SCG (Zhang et al., 2021) in Table 4. UPT is a transformer-based method, while SCG is a graph-based method. They are SOTA two-stage HOI methods (see Fig. 4 and Table 2).

Both of them demand a large number of iterated computations. The table shows that in the inference stage, the FLOP number per query of

Table 5

The ablation studies of the single modules. Feature Selection refers to Module C in Fig. 5. Two-query is the feature construction process, which divides the features into two predictions. The hybrid coding scheme involves coding with binary codes and hamming error correction codes.

Experiments	Feature selection	Two-query	Coding	Default		
				Full	Rare	Non-Rare
Exp #1				11.83	9.78	12.44
Exp #2	1			16.01	10.97	17.52
Exp #3	1	1		20.55	13.47	22.66
EHOI (ours)	1	1	1	24.53	19.26	26.09

EHOI is 1/4500 and 1/15,800 of that of SCG and UPT, respectively. Despite the possibility of accelerating tensor operations through parallel computing, the two deep learning models require much more electricity during the inference stage. EHOI is the most eco-friendly option in terms of carbon footprint.

4.4. Ablation studies

We conduct an ablation study to evaluate the individual module's functionality in Section 4.4.1. Table 5 compares the performance of our model under different settings. In Section 4.4.2, we provide the detailed difference in detection results between one-hot and Hamming coding with error correction codes. Fig. 8 shows the advantage of the hybrid coding in the proposed HOI. The qualitative visual examples are illustrated in Section 4.4.3. Figs. 9 and 10 visualize the effect of the two-query structure in the proposed EHOI detector.

In this section, all the detection experiments and the visualizations are conducted in the HICO-DET dataset under the default setting. We follow the result format in Section 4 and use the mean Average Precision evaluation metric in the following subsections.

4.4.1. Modular design

Modular design is one of the highlights of our method. The feedforward components include feature selection, representation construction, and hybrid coding.

Two experiments are compared to study the feature selection effect: one using the raw features from the first-stage detector model and the other using the 1000 most discriminative features selected through feature selection (i.e., Exp #1 and 2 in Table 5). We observe that the performance improves when more discriminative and less noisy features are used as input.

Exp #2 and 3 demonstrate the impact of decoupling in conditional probability by linear combination proposed Eq. (6). Without the decomposition, the estimator takes all the selected features from the human and object RoI representations as the input and predicts the possible relationship among the 117 classes. In contrast, the two-query structure comprises two estimators taking human and object features, respectively.

Exp #3 and the proposed EHOI compare the conventional one-hot and the proposed hybrid coding schemes. The hybrid coding scheme can benefit all categories in the HICO-DET dataset. The detailed comparison between the coding scheme and the number of parameters is revealed in Table 1. The modularized experiments are the evidence for the decomposition process in Section 3.2.4. The subproblem formulations can improve the estimators. In summary, the modules in the proposed EHOI are crucial to achieve the best results in HOI detection.

4.4.2. Effect of label coding

The relations between different coding schemes and detection performance are shown in Table 1. We can further discuss precisions over the specific relation detection. We visualize the coding effects on the detection performance. One-hot and hamming coding can compensate for the performance in rare and non-rare cases. The results of one-hot Table 6

The precision, recall, and F1 scores for the picked rare cases ('set umbrella' and 'lose umbrella').

Interaction	Coding Scheme	Precision	Recall	F1-score	AP
set umbrella	one-hot	0.03	0.20	0.04	0.077
	Hamming	0.52	0.65	0.58	0.515
lose umbrella	one-hot	0.02	0.56	0.04	0.051
	Hamming	0.39	0.83	0.53	0.400

coding, Hamming coding, and hybrid coding are illustrated in Fig. 7. As the proposed learning scheme, hybrid coding can provide the best detection performance. The two coding schemes can provide robustness and improve the detection performance by providing counterparts to each other.

In the detailed performance, one-hot coding and Hamming coding benefit different categories. Hamming coding can reach comparable performance when one-hot coding performs well. On the other hand, when the one-hot coding fails, the Hamming coding can provide the counterparts of the detection.

Zooming on the difference between two coding schemes and hybrid coding, Fig. 8 shows the performance increase along the relation categories. With the histogram, the performance does not result from using the two codes simultaneously. On the contrary, the relationship classes prefer a particular coding scheme. We can remove the extra bits using the performance boost as the criteria. Consequently, we can have the variable hybrid coding schemes for the corresponding relationship classes.

For the intuitive understanding of Hamming coding and one-hot coding impact on the detection, we picked the *'set umbrella'* and *'lose umbrella'* from the rare interactions for detailed analysis. As shown in Table 6, the Hamming coding could improve the precision of the rare case precisions. Due to the imbalanced data distribution, the recalls with one-hot coding are significantly lower than the Hamming coding. There is a great amount of missing detection of the ground truth. By relabeling the interaction with ECCs, the detection performance increases in the two rare cases.

4.4.3. Visual explaination

To verify the decoupling in probability in Eq. (6), we visualize the outputs from the human and object queries. In the example, we use the object type *umbrella* and the corresponding relations {*carry*, *hold*, *lose*, *no_interaction*, *open*, *repair*, *set*, *stand_under*}. From human understanding, {*carry*, *hold*, *open*, *stand_under*} may relate to the human representations. On the other hand, {*lose*, *repair*, *set*} may relate to the object representations.

Fig. 9 demonstrates the prediction from the human and object queries. The first row is the prediction from the human query, and the second is the prediction from the object query. The human query may confuse the predictions in the *lose* and *set* relations. In contrast, the object query can precisely capture the object status and correctly predict the *lose* and *set*.

Aiming to further understand the classifiers' behaviors, we give an example in Fig. 10. The human representations are similar in Figs. 9 and 10, which can be revealed in the ground truth labels *hold* and *stand_under*. However, human queries cannot detect the relations between the *set* and *lose*. The object queries can detect the two relations. In the object-query classifiers, the *set* and *lose* owe the highest probability scores in the results, leading to high confidence in the predictions.

The decision process can be treated as human understanding. We can take the triple <human, relation, object> as two pairs: <human, relation> and <relation, object>. The human query is based on the information that describes the <human, relation> pairs. In contrast, the object query captures the visual features of the objects and expresses the <relation, object> pairs. The final outputs of the triple can be calculated as the weighted sum of the two predictors using Eq. (6).



Fig. 7. The APs respect to different coding schemes. The one-hot coding in the interaction labels is represented by the blue histogram in the AP performance, while the green histogram depicts the Hamming coding performance. Additionally, the hybrid coding, which is proposed in our methodology, is illustrated by the gray histogram.



Fig. 8. The boost between the hybrid and on-hot coding and Hamming coding scheme. The difference between hybrid coding and one-hot coding is shown in blue. Similarly, the difference between hybrid and Hamming coding is shown in green. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Fig. 9. The classifier in the human query cannot find the critical representations in the umbrella region of interest. Fortunately, the object query can provide the counterparts of the prediction and make the outputs robust. Here, the 'set umbrella' is a rare interaction in the dataset. The proposed EHOI can successfully detect the three ground truth relations.

5. Conclusion and future work

An efficient HOI detector, called EHOI, was proposed in this work. It is both mathematically transparent and computationally efficient while offering competitive detection performance. The use of ECC for the coding of rare interaction types helps improve the robustness of EHOI. The divide-and-conquer feedforward design can reduce the complexity of the modules. Meanwhile, the intermediate results can provide interpretable insights into the decision-making processes, enhancing user comprehension and reducing the potential for misuse.

Nevertheless, the modular design can yield only partial results, limiting the ability to fix the error when the intermediate results go wrong. In the future, we can extend the green learning scheme to image understanding applications and create human-sensible learning algorithms. We can build up the controllable module to complete



Fig. 10. The human representation is similar to Fig. 9. The human query provides a probability distribution similar to the previous example. In contrast, the object query provides the *lose* label. Here, the 'lose umbrella' is a rare interaction in the dataset. The proposed EHOI can successfully detect the three ground truth relations.

the whole system and provide better performance by aligning human intelligence with machine learning algorithms.

CRediT authorship contribution statement

Tsung-Shan Yang: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Data curation, Conceptualization. Yun-Cheng Wang: Writing – review & editing. Chengwei Wei: Writing – review & editing. Suya You: Investigation. C.-C. Jay Kuo: Writing – review & editing, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Army Research Laboratory (ARL) under agreement W911NF2020157. Computation for the work was supported by the University of Southern California's Center for Advanced Research Computing (carc.usc.edu).

Data availability

Open source codes: https://github.com/keevin60907/EHOI---Effici ent-Human-Object-Interaction-Detector.

References

- Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R., 2020. Detecting humanobject interactions via functional generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10460–10469.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. Endto-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.
- Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., Deng, J., 2018. Learning to detect human-object interactions. In: 2018 IEEE Winter Conference on Applications of Computer Vision. Wacv, IEEE, pp. 381–389.
- Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794.
- Chen, M., Liao, Y., Liu, S., Chen, Z., Wang, F., Qian, C., 2021. Reformulating hoi detection as adaptive set prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9004–9013.
- Chen, J., Yanai, K., 2023. Qahoi: Query-based anchors for human-object interaction detection. In: 2023 18th International Conference on Machine Vision and Applications. MVA, IEEE, pp. 1–5.
- Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773.

- Gao, C., Xu, J., Zou, Y., Huang, J.-B., 2020. Drg: Dual relation graph for human-object interaction detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer, pp. 696–712.
 Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448.
- Gupta, S., Malik, J., 2015. Visual semantic role labeling. arXiv preprint arXiv:1505. 04474.
- Gupta, T., Schwing, A., Hoiem, D., 2019. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9677–9685.
- Hamming, R.W., 1950. Error detecting and error correcting codes. Bell Syst. Tech. J. 29 (2), 147–160.
- Hou, Z., Peng, X., Qiao, Y., Tao, D., 2020. Visual compositional learning for humanobject interaction detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer, pp. 584–600.
- Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D., 2021. Detecting human-object interaction via fabricated compositional learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14646–14655.
- Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H., 2023. Detrs with hybrid matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19702–19712.
- Kato, K., Li, Y., Gupta, A., 2018. Compositional learning for human object interaction. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 234–251.
- Kim, B., Lee, J., Kang, J., Kim, E.-S., Kim, H.J., 2021. Hotr: End-to-end humanobject interaction detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 74–83.
- Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Nav. Res. Logist. Q. 2 (1–2), 83–97.
- Kuo, C.-C.J., Chen, Y., 2018. On data-driven saak transform. J. Vis. Commun. Image Represent. 50, 237–246.
- Kuo, C.-C.J., Madni, A.M., 2023. Green learning: Introduction, examples and outlook. J. Vis. Commun. Image Represent. 90, 103685.
- Kuo, C.-C.J., Zhang, M., Li, S., Duan, J., Chen, Y., 2019. Interpretable convolutional neural networks via feedforward design. J. Vis. Commun. Image Represent. 60, 346–359.
- Liao, Y., Liu, S., Wang, F., Chen, Y., Qian, C., Feng, J., 2020. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 482–490.
- Liao, Y., Zhang, A., Lu, M., Wang, Y., Li, X., Liu, S., 2022. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20123–20132.
- Lim, J., Baskaran, V.M., Lim, J.M.-Y., Wong, K., See, J., Tistarelli, M., 2023. ERNet: An efficient and reliable human-object interaction detection network. IEEE Trans. Image Process. 32, 964–979.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Trans. Pattern Anal. Mach. Intell. 42 (2), 318–327. http: //dx.doi.org/10.1109/TPAMI.2018.2858826.
- Ma, S., Wang, Y., Wang, S., Wei, Y., 2023. Fgahoi: Fine-grained anchors for human-object interaction detection. arXiv preprint arXiv:2301.04019.
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.
- Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

- Sun, M., Suo, W., Wang, J., Wang, P., Zhang, Y., 2024. CHA: Conditional hyper-adapter method for detecting human-object interaction. Pattern Recognit. 111075.
- Tamura, M., Ohashi, H., Yoshinaga, T., 2021. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10410–10419.
- Ulutan, O., Iftekhar, A., Manjunath, B.S., 2020. Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13617–13626.
- Wan, B., Zhou, D., Liu, Y., Li, R., He, X., 2019. Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9469–9478.
- Wang, T., Yang, T., Danelljan, M., Khan, F.S., Zhang, X., Sun, J., 2020. Learning human-object interaction detection using interaction points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4116–4125.
- Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G., 2022. Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4794–4803.
- Yang, Y., Wang, W., Fu, H., Kuo, C.-C.J., et al., 2022. On supervised feature selection from high dimensional feature spaces. APSIPA Trans. Signal Inf. Process. 11 (1).

- Yuan, H., Zhang, S., Wang, X., Albanie, S., Pan, Y., Feng, T., Jiang, J., Ni, D., Zhang, Y., Zhao, D., 2023. Rlipv2: Fast scaling of relational language-image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21649–21661.
- Zhang, F.Z., Campbell, D., Gould, S., 2021. Spatially conditioned graphs for detecting human-object interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13319–13327.
- Zhang, F.Z., Campbell, D., Gould, S., 2022. Efficient two-stage detection of humanobject interactions with a novel unary-pairwise transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20104–20112.
- Zhong, X., Ding, C., Qu, X., Tao, D., 2021. Polysemy deciphering network for robust human-object interaction detection. Int. J. Comput. Vis. 129, 1910–1929.
- Zhou, P., Chi, M., 2019. Relation parsing neural network for human-object interaction detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 843–851.
- Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.
- Zhu, F., Xie, Y., Xie, W., Jiang, H., 2023. Diagnosing human-object interaction detectors. arXiv preprint arXiv:2308.08529.