QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information

Masato Tamura[1], Hiroki Ohashi[2], and Tomoaki Yoshinaga[1]

[1]Lumada Data Science Lab., Hitachi, Ltd.
[2]Center for Technology Innovation - Artificial Intelligence, Hitachi, Ltd.
{*masato.tamura.sf, hiroki.ohashi.uo, tomoaki.yoshinaga.xc*}*@hitachi.com*

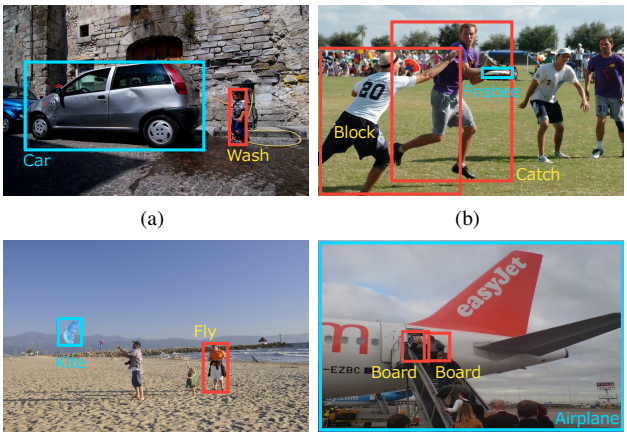
Abstract

We propose a simple, intuitive yet powerful method for human-object interaction (HOI) detection. HOIs are so diverse in spatial distribution in an image that existing CNN-based methods face the following three major drawbacks; they cannot leverage image-wide features due to CNN's locality, they rely on a manually defined location-of-interest for the feature aggregation, which sometimes does not cover contextually important regions, and they cannot help but mix up the features for multiple HOI instances if they are located closely. To overcome these drawbacks, we propose a transformer-based feature extractor, in which an attention mechanism and query-based detection play key roles. The attention mechanism is effective in aggregating contextually important information image-wide, while the queries, which we design in such a way that each query captures at most one human-object pair, can avoid mixing up the features from multiple instances. This transformer-based feature extractor produces so effective embeddings that the subsequent detection heads may be fairly simple and intuitive. The extensive analysis reveals that the proposed method successfully extracts contextually important features, and thus outperforms existing methods by large margins (5.37 mAP on HICO-DET, and 5.6 mAP on V-COCO). The source codes are available at https://github.com/hitachi-rd-cv/qpic.

1. Introduction

Human-object interaction (HOI) detection has attracted enormous interest in recent years for its potential in deeper scene understanding [3–6, 8, 11–13, 15–17, 19–21, 24, 27, 29–37]. Given an image, the task of HOI detection is to localize a human and object, and identify the interactions between them, typically represented as ⟨*human bounding box, object bounding box, object class, action class*⟩.

Conventional HOI detection methods can be roughly divided into two types: two-stage methods [3–6, 8, 11, 13, 15, 16, 19–21, 24, 27, 29–31, 33–37] and single-stage methods [12, 17, 32]. In the two-stage methods, humans and objects are first individually localized by off-the-shelf object detectors, and then the region features from the localized area are used to predict action classes. To incorporate contextual information, auxiliary features such as the features from the union region of a human and object bounding box, and locations of the bounding boxes in an image are often utilized. The single-stage methods predict interactions using the features of a heuristically-defined position such as a midpoint between a human and object center [17].



(a) (b) (c) (d)

Figure 1. Observed failure cases of conventional methods. The ground-truth human bounding boxes, object bounding boxes, object classes, and action classes are drawn with red boxes, blue boxes, blue characters, and yellow characters, respectively.

While both two- and single-stage methods have shown significant improvement, they often suffer from errors attributed to the nature of convolutional neural networks (CNNs) and the heuristic way of using CNN features. Figure 1 shows observed failure cases of conventional methods. In Fig. 1a, we can easily recognize from an entire image

that a boy is washing a car. It is difficult, however, for two-stage methods to predict the action class "wash" since they typically use only the cropped bounding-box regions. The regions sometimes miss contextually important cues located outside the human and object bounding box such as the hose in Fig. 1a. Even though the features of union regions may contain such cues, these regions are frequently dominated by disturbing contents such as background and irrelevant humans and objects. Figure 1b shows an example where multiple HOI instances are overlapped. In this example, the region features of the catching person inevitably contain the features of the blocking person because the hand of the latter, which is important to predict "block", is in the bounding box of the former, ending up in contaminated features. The detection based on the contaminated features easily results in failures. The single-stage methods attempt to capture the contextual information by pairing a target human and object from an early stage in feature extraction and extracting integrated features rather than individually treating the targets. To determine the regions from which integrated features are extracted, they rely on heuristically-designed location-of-interest such as a midpoint between a human and object center [17]. However, such reliance sometimes causes a problem. Fig. 1c shows an example where a target human and object are located distantly. In this example, the midpoint is located close to the man in the middle, who is not relevant to the target HOI instance. Therefore, it is difficult to detect the target on the basis of the features around the midpoint. Fig. 1d is an example where the midpoints of multiple HOI instances are close to each other. In this case, CNN-based methods tend to make mis-detection due to the contaminated features as they do in the case of Fig. 1b.

To overcome these drawbacks, we propose QPIC, a query-based HOI detector that detects a human and object in a pairwise manner with image-wide contextual information. QPIC is extended from a recently proposed object detector, DETR [2], and has the attention mechanism and query-based detection in a transformer [28] as key components. The attention mechanism scans through the entire area of an image and is expected to selectively aggregate contextually important features according to the contents of an image. Moreover, we design QPIC's queries in such a way that they can separately extract features of multiple HOI instances without contaminating them even when the instances are located closely. We realize this by making each query capture at most one human-object pair. This enables to calculate attentions query-wise as opposed to location-wise, and to clarify each query's target human-object-pair through the self-attention mechanism. These key designs of the attention mechanism and query-based pairwise detection make QPIC robust even under the difficult conditions such as the case where contextually important information appears outside the human and object bounding box (Fig. 1a), the target human and object are located distantly (Fig. 1c), and multiple instances are close to each other (Fig. 1b and 1d). The key designs produce so effective embeddings that the subsequent detection heads may be fairly simple and intuitive.

To summarize, our contributions are three-fold: (1) We propose a simple yet effective query-based HOI detector, QPIC. To the best of our knowledge, this is the first work to introduce an attention- and query-based method to HOI detection. (2) We achieve significantly better performance than state-of-the-art methods on two challenging HOI detection benchmarks. (3) We conduct detailed analysis on the behavior of QPIC, and reveal some of the important characteristics of HOI detection tasks that conventional methods could not capture, but QPIC does relatively well.

2. Related Work

Two-stage HOI detection methods [3–6, 8, 11, 13, 15, 16, 19–21, 24, 27, 29–31, 33–37] utilize Faster R-CNN [25] or Mask R-CNN [9] to localize targets. Then, they crop features of backbone networks inside the localized regions. The cropped features are typically processed with multi-stream networks. Each stream processes features of target humans, those of objects, and some auxiliary features such as spatial configurations of the targets, and human poses either alone or in combination. Some of the two-stage methods [21, 24, 27, 30, 34, 36] utilize graph neural networks to refine the features. These methods mainly focus on the second stage architecture, which uses cropped features to predict action classes. However, the cropped features sometimes lack contextual information outside the cropped regions or are contaminated by features of irrelevant targets, which results in the degradation of the performance.

Recently, single-stage methods [12, 17, 32] that utilize integrated features from a pair of a human and object have been proposed to solve the problem in the individually cropped features. Liao *et al*. [17] and Wang *et al*. [32] proposed a point-based interaction detection method that utilizes CenterNet [38] as a base detector. This method predicts action classes using integrated features collected at a midpoint between a human and object center. In particular, Liao *et al*.'s PPDM [17] achieves simultaneous object and interaction detection training, which is the most similar to our training approach. Kim *et al*. [12] proposed UnionDet, which predicts the union bounding box of a human-object pair to extract integrated features. Although these methods attempt to capture contextual information by integrated features, they are still insufficient and sometimes contaminated due to the CNN's locality and heuristically-designed location-of-interests.

Our method differs from conventional methods in that we leverage a transformer in DETR [2] to aggregate image-wide contextual features in a pairwise manner.

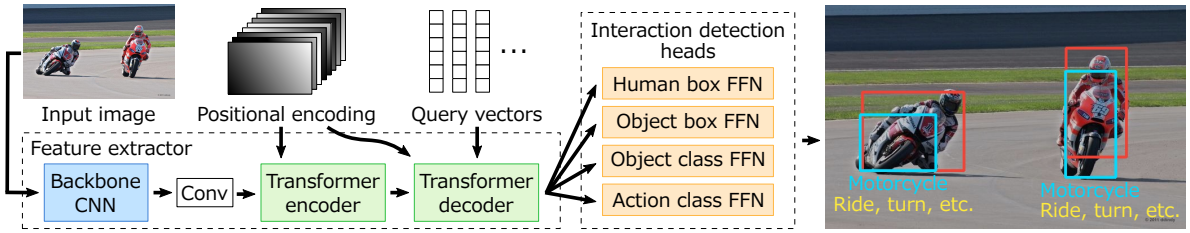


Figure 2. Overall architecture of the proposed QPIC.

3. Proposed Method

To effectively extract important features for each HOI instance taking image-wide contexts into account, we propose to leverage transformer-based architecture as a base feature extractor. Because of the limited space, we explain our method with a transformer as a module. For the details of the calculation in a transformer, we encourage readers to refer to the paper of DETR [2].

3.1. Overall Architecture

Figure 2 illustrates the overall architecture of QPIC. Given an input image $\boldsymbol{x} \in \mathbb{R}^{3\times H\times W}$, a feature map $\boldsymbol{z_b} \in \mathbb{R}^{D_b\times H'\times W'}$ is calculated by an arbitrary off-the-shelf backbone network, where H and W are the height and width of the input image, H' and W' are those of the output feature map, and D_b is the number of channels. Typically $H' < H$ and $W' < W$. $\boldsymbol{z_b}$ is then input to a projection convolution layer with a kernel size of 1×1 to reduce the dimension from D_b to D_c.

The transformer encoder takes this feature map with the reduced dimension $\boldsymbol{z_c} \in \mathbb{R}^{D_c\times H'\times W'}$ to produce another feature map with richer contextual information on the basis of the self-attention mechanism. A fixed positional encoding $\boldsymbol{p} \in \mathbb{R}^{D_c\times H'\times W'}$ [1, 2, 23] is additionally input to the encoder to supplement the positional information, which the self-attention mechanism alone cannot inherently incorporate. The encoded feature map $\boldsymbol{z_e} \in \mathbb{R}^{D_c\times H'\times W'}$ is then obtained as $\boldsymbol{z_e} = f_{enc}\left(\boldsymbol{z_c}, \boldsymbol{p}\right)$, where $f_{enc}\left(\cdot,\cdot\right)$ is a set of stacked transformer encoder layers.

The transformer decoder transforms a set of learnable query vectors $\boldsymbol{Q} = \{\boldsymbol{q}_i | \boldsymbol{q}_i \in \mathbb{R}^{D_c}\}_{i=1}^{N_q}$ into a set of embeddings $\boldsymbol{D} = \{\boldsymbol{d}_i | \boldsymbol{d}_i \in \mathbb{R}^{D_c}\}_{i=1}^{N_q}$ that contain image-wide contextual information for HOI detection, referring to the encoded feature map $\boldsymbol{z_e}$ using the attention mechanism. N_q is the number of query vectors. The queries are designed in such a way that one query captures at most one human-object pair and an interaction(s) between them. N_q is therefore set to be large enough so that it is always larger than the number of actual human-object pairs in an image. The decoded embeddings are then obtained as $\boldsymbol{D} = f_{dec}\left(\boldsymbol{z_e}, \boldsymbol{p}, \boldsymbol{Q}\right)$, where $f_{dec}\left(\cdot,\cdot,\cdot\right)$ is a set of stacked transformer decoder layers. We use a positional encoding $\boldsymbol{p}$ again to incorporate the spatial information.

The subsequent interaction detection heads further processes the decoded embeddings to produce N_q prediction results. Here, we note that one or more HOIs corresponding to a human-object pair are mathematically defined by the following four vectors: a human-bounding-box vector normalized by the corresponding image size $\boldsymbol{b}^{(h)} \in [0,1]^4$, a normalized object-bounding-box vector $\boldsymbol{b}^{(o)} \in [0,1]^4$, an object-class one-hot vector $\boldsymbol{c} \in \{0,1\}^{N_{obj}}$, where N_{obj} is the number of object classes, and an action-class vector $\boldsymbol{a} \in \{0,1\}^{N_{act}}$, where N_{act} is the number of action classes. Note that $\boldsymbol{a}$ is not necessarily a one-hot vector because there may be multiple actions that correspond to a human-object pair. Our interaction detection heads are composed of four small feed-forward networks (FFNs): human-bounding-box FFN f_h, object-bounding-box FFN f_o, object-class FFN f_c, and action-class FFN f_a, each of which is dedicated to predict one of the aforementioned 4 vectors, respectively. This design of the interaction detection heads is fairly intuitive and simple compared with a number of state-of-the-art methods such as the point-detection and point-matching branch in PPDM [17] and the human, object, and spatial-semantic stream in DRG [4]. Thanks to the powerful embeddings that contain image-wide contextual information, QPIC does not have to rely on a rather complicated and heuristic design to produce the prediction. One thing to note is that unlike many existing methods [3–6, 8, 11, 13, 15, 16, 19, 20, 24, 27, 29–31, 33–37], which first attempt to detect humans and objects individually and later pair them to find interactions, it is crucial to design queries in such a way that one query directly captures a human and object as a pair to more effectively extract features for interactions. We will experimentally verify this claim in Sec. 4.4.1.

The prediction of normalized human bounding boxes $\{\hat{\boldsymbol{b}}_i^{(h)} | \hat{\boldsymbol{b}}_i^{(h)} \in [0,1]^4\}_{i=1}^{N_q}$, that of object bounding boxes $\{\hat{\boldsymbol{b}}_i^{(o)} | \hat{\boldsymbol{b}}_i^{(o)} \in [0,1]^4\}_{i=1}^{N_q}$, the probability of object classes $\{\hat{\boldsymbol{c}}_i | \hat{\boldsymbol{c}}_i \in [0,1]^{N_{obj}+1}, \sum_{j=1}^{N_{obj}+1} \hat{\boldsymbol{c}}_i(j) = 1\}_{i=1}^{N_q}$, where $\boldsymbol{v}(j)$ denotes the j-th element of $\boldsymbol{v}$, and the probability of action classes $\{\hat{\boldsymbol{a}}_i | \hat{\boldsymbol{a}}_i \in [0,1]^{N_{act}}\}_{i=1}^{N_q}$, are calculated as $\hat{\boldsymbol{b}}_i^{(h)} = \sigma\left(f_h\left(\boldsymbol{d}_i\right)\right), \hat{\boldsymbol{b}}_i^{(o)} = \sigma\left(f_o\left(\boldsymbol{d}_i\right)\right), \hat{\boldsymbol{c}}_i = \varsigma\left(f_c\left(\boldsymbol{d}_i\right)\right), \hat{\boldsymbol{a}}_i = \sigma\left(f_a\left(\boldsymbol{d}_i\right)\right)$, respectively. σ, ς are the sigmoid and softmax functions, respectively. Note that $\hat{\boldsymbol{c}}_i$ has the $(N_{obj}+1)$-th element to indicate that the i-th query has no correspond-

ing human-object pair, while an additional element of $\hat{\boldsymbol{a}}_i$ to indicate "no action" is not necessary because we use the sigmoid function rather than the softmax function to calculate the action-class probabilities for co-occuring actions.

3.2. Loss Calculation

The loss calculation is composed of two stages: the bipartite matching stage between predictions and ground truths, and the loss calculation stage for the matched pairs.

For the bipartite matching, we follow the training procedure of DETR [2] and use the Hungarian algorithm [14]. Note that this design obviates the process of suppressing over-detection as described in [2]. We first pad the ground-truth set of human-object pairs with ϕ (no pairs) so that the ground-truth set size becomes N_q. We then leverage the Hungarian algorithm to determine the optimal assignment $\hat{\omega}$ among the set of all possible permutations of N_q elements $\boldsymbol{\Omega}_{N_q}$, i.e. $\hat{\omega} = \arg\min_{\omega \in \boldsymbol{\Omega}_{N_q}} \sum_{i=1}^{N_q} \mathcal{H}_{i,\omega(i)}$, where $\mathcal{H}_{i,j}$ is the matching cost for the pair of i-th ground truth and j-th prediction. The matching cost $\mathcal{H}_{i,j}$ consists of four types of costs: the box-regression cost $\mathcal{H}^{(b)}_{i,j}$, intersection-over-union (IoU) cost $\mathcal{H}^{(u)}_{i,j}$, object-class cost $\mathcal{H}^{(c)}_{i,j}$, and action-class cost $\mathcal{H}^{(a)}_{i,j}$. Denoting i-th ground truth for the normalized human bounding box by $\boldsymbol{b}^{(h)}_i \in [0,1]^4$, normalized object bounding box by $\boldsymbol{b}^{(o)}_i \in [0,1]^4$, object-class one-hot vector by $\boldsymbol{c}_i \in \{0,1\}^{N_{obj}}$, and action class by $\boldsymbol{a}_i \in \{0,1\}^{N_{act}}$, the aforementioned costs are formulated as follows.

$$\mathcal{H}_{i,j} = \mathbb{1}_{\{i \notin \boldsymbol{\Phi}\}} \left[\eta_b \mathcal{H}^{(b)}_{i,j} + \eta_u \mathcal{H}^{(u)}_{i,j} + \eta_c \mathcal{H}^{(c)}_{i,j} + \eta_a \mathcal{H}^{(a)}_{i,j} \right], \tag{1}$$

$$\mathcal{H}^{(b)}_{i,j} = \max \left\{ \left\| \boldsymbol{b}^{(h)}_i - \hat{\boldsymbol{b}}^{(h)}_j \right\|_1, \left\| \boldsymbol{b}^{(o)}_i - \hat{\boldsymbol{b}}^{(o)}_j \right\|_1 \right\}, \tag{2}$$

$$\mathcal{H}^{(u)}_{i,j} = \max \left\{ -GIoU\left(\boldsymbol{b}^{(h)}_i, \hat{\boldsymbol{b}}^{(h)}_j\right), -GIoU\left(\boldsymbol{b}^{(o)}_i, \hat{\boldsymbol{b}}^{(o)}_j\right) \right\}, \tag{3}$$

$$\mathcal{H}^{(c)}_{i,j} = -\hat{\boldsymbol{c}}_j(k) \quad s.t. \quad \boldsymbol{c}_i(k) = 1, \tag{4}$$

$$\mathcal{H}^{(a)}_{i,j} = -\frac{1}{2}\left(\frac{\boldsymbol{a}_i^{\intercal} \hat{\boldsymbol{a}}_j}{\|\boldsymbol{a}_i\|_1 + \epsilon} + \frac{(\boldsymbol{1} - \boldsymbol{a}_i)^{\intercal}(\boldsymbol{1} - \hat{\boldsymbol{a}}_j)}{\|\boldsymbol{1} - \boldsymbol{a}_i\|_1 + \epsilon} \right), \tag{5}$$

where $\boldsymbol{\Phi}$ is a set of ground-truth indices that correspond to ϕ, $GIoU(\cdot,\cdot)$ is the generalized IoU [26], ϵ is a small positive value introduced to avoid zero divide, and η_b, η_u, η_c, and η_a are the hyper-parameters. We use two types of bounding-box cost $\mathcal{H}^{(b)}_{i,j}$ and $\mathcal{H}^{(u)}_{i,j}$ following [2]. In calculating $\mathcal{H}^{(b)}_{i,j}$ and $\mathcal{H}^{(u)}_{i,j}$, instead of minimizing the average of a human and object-bounding-box cost, we minimize the larger of the two to prevent the matching from being undesirably biased to either if one cost is significantly lower than the other. We design $\mathcal{H}^{(a)}_{i,j}$ so that the costs of both positive and negative action classes are taken into account. In addition, we formulate it using the weighted average of the two with the inverse number of nonzero elements as the weights rather than using the vanilla average. This is necessary to balance the effect from the two costs because the number of positive action classes is typically much smaller than that of negative action classes.

The loss to be minimized in the training phase is calculated on the basis of the matched pairs as follows.

$$\mathcal{L} = \lambda_b \mathcal{L}_b + \lambda_u \mathcal{L}_u + \lambda_c \mathcal{L}_c + \lambda_a \mathcal{L}_a, \tag{6}$$

$$\mathcal{L}_b = \frac{1}{|\bar{\boldsymbol{\Phi}}|} \sum_{i=1}^{N_q} \mathbb{1}_{\{i \notin \boldsymbol{\Phi}\}} \left[\left\| \boldsymbol{b}^{(h)}_i - \hat{\boldsymbol{b}}^{(h)}_{\hat{\omega}(i)} \right\|_1 + \left\| \boldsymbol{b}^{(o)}_i - \hat{\boldsymbol{b}}^{(o)}_{\hat{\omega}(i)} \right\|_1 \right], \tag{7}$$

$$\mathcal{L}_u = \frac{1}{|\bar{\boldsymbol{\Phi}}|} \sum_{i=1}^{N_q} \mathbb{1}_{\{i \notin \boldsymbol{\Phi}\}} \left[2 - GIoU\left(\boldsymbol{b}^{(h)}_i, \hat{\boldsymbol{b}}^{(h)}_{\hat{\omega}(i)}\right) - GIoU\left(\boldsymbol{b}^{(o)}_i, \hat{\boldsymbol{b}}^{(o)}_{\hat{\omega}(i)}\right) \right], \tag{8}$$

$$\mathcal{L}_c = \frac{1}{N_q} \sum_{i=1}^{N_q} \left\{ \mathbb{1}_{\{i \notin \boldsymbol{\Phi}\}} \left[-\log \hat{\boldsymbol{c}}_{\hat{\boldsymbol{\omega}}(\boldsymbol{i})}(k) \right] + \mathbb{1}_{\{i \in \boldsymbol{\Phi}\}} \left[-\log \hat{\boldsymbol{c}}_{\hat{\boldsymbol{\omega}}(\boldsymbol{i})}(N_{obj}+1) \right] \right\} \quad s.t. \quad \boldsymbol{c}_i(k) = 1, \tag{9}$$

$$\mathcal{L}_a = \frac{1}{\sum_{i=1}^{N_q} \mathbb{1}_{\{i \notin \boldsymbol{\Phi}\}} \|\boldsymbol{a}_i\|_1} \sum_{i=1}^{N_q} \left\{ \mathbb{1}_{\{i \notin \boldsymbol{\Phi}\}} \left[l_f\left(\boldsymbol{a}_i, \hat{\boldsymbol{a}}_{\hat{\boldsymbol{\omega}}(\boldsymbol{i})}\right) \right] + \mathbb{1}_{\{i \in \boldsymbol{\Phi}\}} \left[l_f\left(\boldsymbol{0}, \hat{\boldsymbol{a}}_{\hat{\boldsymbol{\omega}}(\boldsymbol{i})}\right) \right] \right\}, \tag{10}$$

where λ_b, λ_u, λ_c and λ_a are the hyper-parameters for adjusting the weights of each loss, and $l_f(\cdot,\cdot)$ is the element-wise focal loss function [18]. For the hyper-parameters of the focal loss, we use the default settings described in [38].

3.3. Inference for Interaction Detection

As previously mentioned, the detection result of an HOI is represented by the following four components, $\langle$*human bounding box, object bounding box, object class, action class*$\rangle$. Our interaction detection heads are designed so intuitively that all we need to do is to pick up the corresponding information from each head. Formally, we set the prediction results corresponding to the i-th query and j-th action as $\langle \hat{\boldsymbol{b}}^{(\boldsymbol{h})}_{\boldsymbol{i}}, \hat{\boldsymbol{b}}^{(\boldsymbol{o})}_{\boldsymbol{i}}, \arg\max_k \hat{\boldsymbol{c}}_{\boldsymbol{i}}(k), j \rangle$. We define a score of the HOI instance as $\{\max_k \hat{\boldsymbol{c}}_{\boldsymbol{i}}(k)\}\, \hat{\boldsymbol{a}}_{\boldsymbol{i}}(j)$, and regard this instance to be present if the score is higher than a threshold.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conducted extensive experiments on two HOI detection datasets: HICO-DET [3] and V-COCO [7]. We followed the standard evaluation scheme. HICO-DET contains

38,118 and 9,658 images for training and testing, respectively. The images are annotated with 80 object and 117 action classes. V-COCO, which originates from the COCO dataset, contains 2,533, 2,867, and 4,946 images for training, validation, and testing, respectively. The images are annotated with 80 object and 29 action classes.

For the evaluation metrics, we use the mean average precision (mAP). A detection result is judged as a true positive if the predicted human and object bounding box have IoUs larger than 0.5 with the corresponding ground-truth bounding boxes, and the predicted action class is correct. In the HICO-DET evaluation, the object class is also taken into account for the judgment. The AP is calculated per object and action class pair in the HICO-DET evaluation, while that is calculated per action class in the V-COCO evaluation.

For HICO-DET, we evaluate the performance in two different settings following [3]: *default setting* and *known-object setting*. In the former setting, APs are calculated on the basis of all the test images, while in the latter setting, each AP is calculated only on the basis of images that contain the object class corresponding to each AP. In each setting, we report the mAP over three set types: a set of 600 HOI classes (*full*), a set of 138 HOI classes that have less than 10 training instances (*rare*), and a set of 462 HOI classes that have 10 or more training instances (*non-rare*). Unless otherwise stated, we use the default full setting in the analysis. In V-COCO, a number of HOIs are defined with no object labels. To deal with this situation, we evaluate the performance in two different scenarios following V-COCO's official evaluation scheme. In scenario 1, detectors are required to report cases in which there is no object, while in scenario 2, we just ignore the prediction of an object bounding box in these cases.

4.2. Implementation Details

We use ResNet-50 and ResNet-101 [10] as a backbone feature extractor. Both transformer encoder and decoder consist of 6 transformer layers with a multi-head attention of 8 heads. The reduced dimension size D_c is set to 256, and the number of query vectors N_q is set to 100. The human- and object-bounding-box FFNs have 3 linear layers with ReLU activations, while the object- and action-class FFNs have 1 linear layer.

For training QPIC, we initialize the network with the parameters of DETR [2] trained with the COCO dataset. Note that for the V-COCO training, we exclude the COCO's training images that are contained in the V-COCO test set when pre-training DETR[1]. QPIC is trained for 150 epochs using the AdamW [22] optimizer with the batch size 16, initial learning rate of the backbone network 10^{-5}, that of the others 10^{-4}, and the weight decay 10^{-4}. Both learning rates are decayed after 100 epochs. The hyper-parameters for the Hungarian costs η_b, η_u, η_c, and η_a, and those for the loss weights $\lambda_b, \lambda_u, \lambda_c$, and λ_a are set to 2.5, 1, 1, 1, 2.5, 1, 1, and 1, respectively. Following [17], we select 100 high scored detection results from all the predictions for fair comparison. Please see the supplementary material for more details.

[1]A few previous works inappropriately use COCO train2017 set for pre-training, whose images are contained in the V-COCO test set.

Table 1. Comparison against state-of-the-art methods on HICO-DET. The top, middle, and bottom blocks show the mAPs of the two-stage, single-stage, and our methods, respectively.

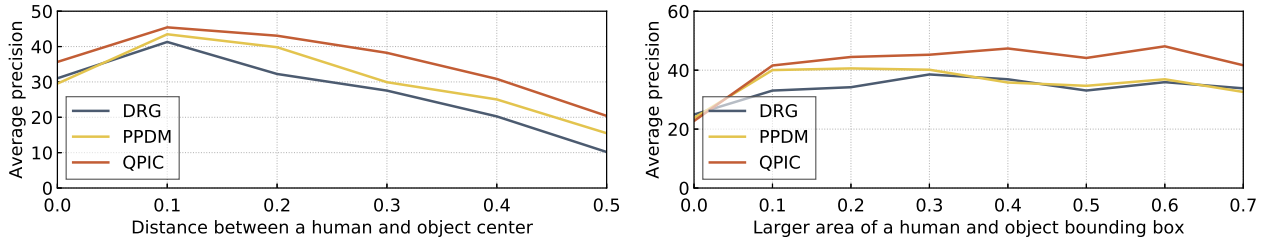
	Default			Known object		
Method	full	rare	non-rare	full	rare	non-rare
VSGNet [27]	19.80	16.05	20.91	–	–	–
FCMNet [20]	20.41	17.34	21.56	22.04	18.97	23.13
VCL [11]	23.63	17.21	25.55	25.98	19.12	28.03
ConsNet [21]	24.39	17.10	26.56	–	–	–
DRG [4]	24.53	19.47	26.04	27.98	23.11	29.43
UnionDet [12]	17.58	11.72	19.33	19.76	14.68	21.27
Wang *et al*. [32]	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [17]	21.73	13.78	24.10	24.58	16.65	26.84
Ours (ResNet-50)	29.07	21.85	31.23	31.68	24.14	33.93
Ours (ResNet-101)	**29.90**	**23.92**	**31.69**	**32.38**	**26.06**	**34.27**

Table 2. Comparison against state-of-the-art methods on V-COCO. The split of the blocks are the same as Table 1.

Method	Scenario 1	Scenario 2
VCL [11]	48.3	–
DRG [4]	51.0	–
VSGNet [27]	51.8	57.0
FCMNet [20]	53.1	–
ConsNet [21]	53.2	–
UnionDet [12]	47.5	56.2
Wang *et al*. [32]	51.0	–
Ours (ResNet-50)	**58.8**	**61.0**
Ours (ResNet-101)	58.3	60.7

4.3. Comparison to State-of-the-Art

We first show the comparison of our QPIC with the latest HOI detection methods including both two- and single-stage methods in Table 1. As seen from the table, QPIC outperforms both state-of-the-art two- and single-stage methods in all the settings. QPIC with the ResNet-101 backbone yields an especially significant gain of 5.37 mAP (relatively 21.9%) compared with DRG [4] and 8.17 mAP (37.6%) compared with PPDM [17] in the default full setting. Table 2 shows the comparison results on V-COCO. QPIC achieves state-of-the-art performance among all the baseline methods. QPIC with the ResNet-50 backbone achieves a 5.6 mAP (10.5%) gain over ConsNet [21], which is the strongest baseline. Unlike in the HICO-DET result, the ResNet-50 backbone shows better performance than



(a) AP depending on the distance between a human and object center. (b) AP depending on the larger area of a human and object bounding box.

Figure 3. Performance analysis on different spatial distribution of HOIs evaluated on HICO-DET.

the ResNet-101 backbone probably because the number of training samples in V-COCO is insufficient to train the large network. Overall, these comparison results demonstrate the dataset-invariant effectiveness of QPIC.

We then investigate in which cases QPIC especially achieves superior performance compared with the strong baselines. To do so, we compare QPIC in detail with DRG [4] and PPDM [17], which are the strongest baselines of the two- and single-stage methods, respectively. We use the ResNet-50 backbone for QPIC in this comparison. Note that hereinafter the distance and area are calculated in normalized image coordinates. Figure 3a shows how the performances change as the distance between the center points of a paired human and object bounding box grows. We split HOI instances into bins of size 0.1 according to the distances, and calculate the APs of each bin that has at least 1,000 HOI instances. As shown in Fig. 3a, the relative gaps of the performance between QPIC and the other two methods become more evident as the distance grows. The graph suggests three things; HOI detection tends to become more difficult as the distance grows, the distant case is especially difficult for CNN-based methods, and QPIC relatively better deals with this difficulty. The possible explanation for these results is that the features of the CNN-based methods, which rely on limited receptive fields for the feature aggregation, cannot include contextually important information or are dominated by irrelevant information in the distant cases, while the features of QPIC are more effective thanks to the ability of selectively extracting image-wide contextual information. Figure 3b presents how the performances change as the areas of target human and object bounding boxes grow. We pick up the larger area of a target human and object bounding box involved in each HOI instance. We then split HOI instances into bins of size 0.1 according to the area, and calculate the APs of each bin that has at least 1,000 HOI instances. As illustrated in Fig. 3b, the gaps of the APs between the conventional methods and QPIC tend to grow as the area increases. This is probably because of the combination of the following two reasons; if the area becomes bigger, the area tends to more often include harmful regions such as another HOI instance, and the conventional methods mix up the irrelevant features in such situation, whereas the attention mechanism and the query-based framework enable to selectively aggregate effective features in a separated manner for each HOI instance. These results reveal that the QPIC's significant improvement shown in Table 1 and Table 2 is likely to be brought by its nature of robustness to diverse spatial distribution of HOIs, probably originating from its capability of aggregating image-wide contextual features for each HOI instance. This observation is further confirmed qualitatively in Sec. 4.5.

Table 3. Evaluation results of the various detection heads.

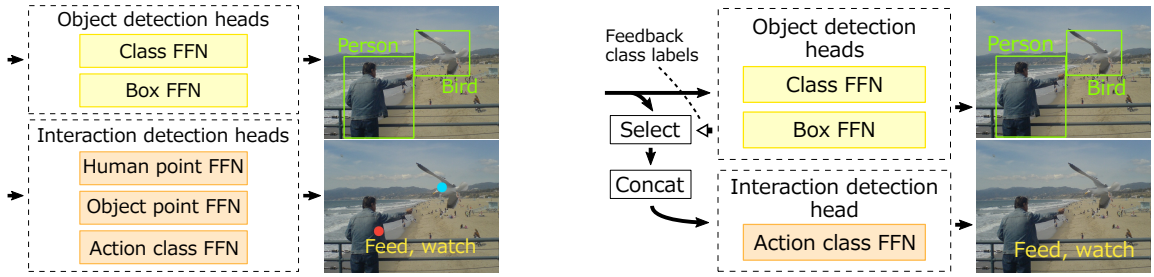
Base method	Detection heads	HICO-DET (mAP)
Ours (ResNet-50)	Simple (original)	29.07
	Point matching (Fig. 4a)	29.04
	Two-stage like (Fig. 4b)	26.18
PPDM [17] (Hourglass-104)	Simple	17.45
	Point matching (original)	21.73

4.4. Ablation Study

To understand the key ingredients of QPIC's superiority shown in Sec. 4.3, we analyze the key building blocks one by one in detail. We first analyze the interaction detection heads in Sec. 4.4.1 and subsequently analyze the transformer-based feature extractor in Sec. 4.4.2.

4.4.1 Analysis on Detection Heads

Feasibility of simple heads. As previously mentioned, the inference process of QPIC is simplified thanks to the enriched features from the transformer-based feature extractor. To confirm that this simple prediction is sufficient for QPIC, we investigated if the detection accuracy increases by leveraging a typical point-matching-based detection heads presented in [17], which is one of the best performing heuristically-designed heads. Figure 4a represents the implemented heads. A notable difference from the original simple heads lies in that the interaction detection heads output center points of target humans and objects instead of bounding boxes. Consequently, the outputs from the interaction detection heads need to be fused with the outputs from the object detection heads with point matching. Note that in this implementation, duplicate detection



(a) Interaction detection with point matching.

(b) Interaction detection with two-stage like approach.

Figure 4. Implemented variants for analyzing detection heads. These heads are on top of our transformer-based feature extractor.

results that share an identical human-object pair needs to be suppressed by some means such as non-maximum suppression. Table 3 shows the evaluation results. As seen from Table 3, the point-matching-based heads exhibit no performance improvement over the simple heads, which indicates that the simple detection heads are enough and we do not have to manually design complicated detection heads.

Importance of pairwise detection. Although the detection heads can be as simple as we present, we claim that there is a crucial aspect that must be covered in the design of the heads. It is to treat a target human and object as a pair from early stages rather than to first detect them individually and later integrate the features from the cropped regions corresponding to the detection, as typically done in two-stage approaches. We verify this claim by looking into the performance of the two-stage like detection-heads on top of our transformer-based feature extractor, which is exactly the same as original QPIC. Figure 4b illustrates the implemented detection heads. This model first derives object detection results from the object detection heads. Then, the results are used to create all the possible human-object pairs. The features of each pair is constructed by concatenating the features from the human and object bounding boxes. The interaction detection head predicts action classes of all the pairs on the basis of the concatenated features. As seen from Table 3, two-stage like method yields worse performance than the original. This observation indicates that the two-stage methods, which rely on individual feature extraction, do not perform well even with our strong feature extractor, and suggests the importance of the pairwise feature extraction in heads for HOI detection.

4.4.2 Analysis on Feature Extractor

Importance of a transformer. To confirm that a transformer-based feature extractor is key to make the simple heads sufficiently work for HOI detection as discussed in Sec. 4.4.1, we replace QPIC's transformer-based feature extractor by a CNN-based counterpart and examine how the performance changes. We utilize the Hourglass-104 backbone used in PPDM [17] in this experiment. Table 3 shows the performance of the original point-matching-based PPDM as well as its simple-heads variant. The simple-heads variant directly predicts all the information corresponding to a human-object pair on the basis of the features extracted in the feature-extraction stage, just as QPIC's simple heads do. More concretely, not only a human point, an object point, and action classes, but also a human-bounding-box size, an object-bounding-box size, and an object class are directly predicted on the basis of the features at the midpoint between the human and object centers. As Table 3 shows, the simple-heads variant exhibits far worse performance than QPIC. This implicates that the CNN-based feature extractor is not as powerful as our transformer-based feature extractor, so the simple heads cannot be leveraged with it. In addition, we find that the point-matching-based heads, which is the original version of PPDM, achieve higher performance than the simple ones, implying that there is a room for increasing accuracy by heuristically designing the heads if the feature extractor is not so powerful, which is not the case with our powerful transformer-based feature extractor.

Table 4. Effect of the transformer encoder and decoder.

Transformer encoder	Transformer decoder	HICO-DET (mAP)	COCO (mAP)
		18.89	34.6
✓		20.07	35.1
	✓	26.75	38.7
✓	✓	29.27	43.5

Importance of a decoder. To further dig into the transformer to find out the essential component for HOI detection, we compare four variants listed in Table 4. The model without the decoder leverages the point-matching-based method like PPDM [17] on top of the encoder's output (with encoder) or on top of the base features (without encoder). The model with the decoder utilizes the point-matching-based heads (Fig. 4a) for fair comparison. We use the ResNet-101 backbone for all the variants. As seen from Table 4, the transformer encoder yields merely slight improvement on HICO-DET (2.52 and 1.18 mAP with and without the decoder, respectively), while the decoder remarkably boosts the performance (9.20 and 7.86 mAP with

(a) (b) (c) (d)

Figure 5. Failure cases of conventional detectors (top row, same as Fig. 1) and attentions of QPIC (bottom row). In (b) and (d), the attentions corresponding to different HOI instances are drawn with blue and orange, and the areas where two attentions overlap are drawn with white.

and without the encoder, respectively). These results indicate that the decoder plays a vital role in HOI detection. Additionally, we evaluate the performance on COCO to compare the degrees of improvement in object detection and HOI detection. As seen in the table, the relative performance improvement brought by the decoder for object detection (on COCO) is 23.9% and 11.8% with and without the encoder, respectively, while that for HOI detection (on HICO-DET) is 45.8% and 41.6% with and without the encoder, respectively. This means that the decoder is more effective in an HOI detection task than in an object detection task. This is probably because the regions of interest (ROI) are mostly consolidated in a single area in object detection tasks, while in HOI detection tasks, the ROI can be diversely distributed image-wide. CNNs, which rely on localized receptive fields, can deal with the former case relatively easily, whereas the image-wide feature aggregation of the decoder is crutial for the latter case.

4.5. Qualitative Analysis

To qualitatively reveal the characteristics of QPIC and the main reasons behind its superior performance over existing methods, we analyze the failure cases of existing methods and QPIC's behavior in the cases. The top row in Fig. 5 shows the failure cases shown in Fig. 1, and the bottom row illustrates the attentions of QPIC on the images.

Figure 5a and 5b show the cases where DRG fails to detect the action classes, but QPIC does not. As previously discussed, the regions in an image other than a human and object bounding box sometimes contain useful information. Figure 5a is a typical example, where the hose held by the boy is likely to be the important contextual information. Two-stage methods that utilize only the region features, namely the human and object bounding box (and sometimes the union region of the two), cannot fully leverage the contextual information, whereas QPIC successfully places the distinguishing focus on such information and leverages it as shown in the attention map. Furthermore, the region features are sometimes contaminated by other region features when target bounding boxes are overlapped. Figure 5b shows such an example, where the hand of the blocking man is contained in the bounding box of the catching man. The typical two-stage methods, which rely on region features, cannot exclude this disturbing information, resulting in incorrect detection. QPIC, however, can selectively aggregate only the helpful information for each HOI as shown in the attention map, resulting in the correct detection.

Figure 5c and 5d illustrate the failure cases of PPDM, whose detection points are drawn in yellow circles. As discussed in Sec. 4.3, features of heuristic detection points are sometimes dominated by irrelevant information such as the non-target human in Fig. 5c and another HOI features in Fig. 5d. Consequently, the detection based on those confusing features tends to result in failures. QPIC alleviates this problem by incorporating the attention mechanism that selectively captures image-wide features as shown in the attention maps, and thus correctly detects these HOIs.

Overall, these qualitative analysis demonstrates the QPIC's capability of acquiring image-wide contextual features, which lead to its superior performance over the existing methods.

5. Conclusion

We have proposed QPIC, a novel detector that can selectively aggregate image-wide contextual information for HOI detection. QPIC leverages an attention mechanism to effectively aggregate features for detecting a wide variety of HOIs. This aggregation enriches HOI features, and as a result, simple and intuitive detection heads are realized. The evaluation on two benchmark datasets showed QPIC's significant superiority over existing methods. The extensive analysis showed that the attention mechanism and query-based detection play a crucial role for HOI detection.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks. In *ICCV*, October 2019.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, September 2020.

[3] Yu-Wei Chao, Yunfan Liu, Michael Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, March 2018.

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

[5] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, September 2018.

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

[7] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. May 2015. arXiv:1505.04474.

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

[9] Kaiming He, Georgia Gkioxari, P. Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, October 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016.

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

[12] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J. Kim. UnionDet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, August 2020.

[13] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, August 2020.

[14] H. W. Kuhn and Bryn Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

[15] 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*, June 2020.

[16] 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*, June 2019.

[17] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. PPDM: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, June 2020.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, October 2017.

[19] Xue Lin, Qi Zou, and Xixia Xu. Action-guided attention mining and relation reasoning network for human-object interaction detection. In *IJCAI*, July 2020.

[20] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *ECCV*, August 2020.

[21] Ye Liu, Junsong Yuan, and Chang Wen Chen. ConsNet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM Multimedia*, October 2020.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, May 2019.

[23] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, July 2018.

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

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

[26] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, June 2019.

[27] Oytun Ulutan, A S M Iftekhar, and B. S. Manjunath. VSGNet: Spatial attention network for detecting human object interactions using graph convolutions. In *CVPR*, June 2020.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, December 2017.

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

[30] Hai Wang, Wei shi Zheng, and Ling Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *ECCV*, August 2020.

[31] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, October 2019.

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

[33] Bingjie Xu, Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Interact as you intend: Intention-driven human-object interaction detection. *TMM*, 22(6):1423–1432, June 2020.

[34] Dongming Yang and Yuexian Zou. A graph-based interactive reasoning for human-object interaction detection. In *IJCAI*, July 2020.

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

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

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

[38] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points, April 2019. arXiv:1904.07850.