QAHOI: Query-Based Anchors for Human-Object Interaction Detection

Junwen Chen Keiji Yanai
Department of Informatics, The University of Electro-Communications, Tokyo, Japan
chen-j@mm.inf.uec.ac.jp, yanai@cs.uec.ac.jp

Abstract

Human-object interaction (HOI) detection as a downstream of object detection task requires localizing pairs of humans and objects and recognizing the interaction between them. Recent one-stage approaches focus on detecting possible interaction points or filtering human-object pairs, ignoring the variability in the location and size of different objects at spatial scales. In this paper, we propose a transformer-based method, QAHOI (Query-Based Anchors for Human-Object Interaction detection), which leverages a multi-scale architecture to extract features from different spatial scales and uses query-based anchors to predict all the elements of an HOI instance. We further investigate that a powerful backbone significantly increases accuracy for QAHOI, and QAHOI with a transformer-based backbone outperforms recent state-of-the-art methods by large margins on the HICO-DET benchmark.

1 Introduction

Human-object interaction (HOI) detection has recently received increasing attention as a field with great potential applications. HOI detection approaches need to extract the semantic relationships between humans and objects and predict a set of ⟨human, object, action⟩ triplets within an image. Specifically, an HOI instance is a pair of human and object bounding boxes, and a corresponding action class represents the relationship between them. HOI detection can be divided into two parts: human-object pair detection and interaction recognition.

To achieve high efficiency, one-stage approaches [5, 9, 7, 19, 3, 14, 17, 6, 15] detect human-object pairs and recognize the corresponding action class in parallel. A commonly adopted way is to make use of the interaction point, which is between the human and object boxes [9, 17, 15], and a matching process is required to match the interaction point with a pair of human and object boxes. Although interaction points combine the HOI instance detection and recognition together, there are mainly two drawbacks such as the semantic features are ambiguous when the interaction point is far apart from the human and object, and the lack of a multi-scale architecture which is commonly used in object detection.

To extract the semantic features between the human-object pairs with more contextual information and less irrelevant local information, Transformer [16] is introduced into HOI detection [7, 19, 3, 14]. As query embeddings in the transformer decoder represent HOI instances and incorporate object detection and interaction recognition together, the transformer-based HOI detection methods also can be seen as query-based methods which belong to the one-stage approach. However, the transformer-based methods [7, 19, 3, 14] are built upon the CNN backbone, and the multi-head attention used in transformer suffers from a quadratic complexity with the growth of the feature map size. Besides, the training of the high complexity transformer suffers from slow convergence, and pre-training the model in object detection task and fine-tuning in HOI detection task are always used to obtain a fine result.

In this paper, we proposed a transformer-based method, which leverages a hierarchical backbone to extract multi-scale context features, and a deformable transformer [18] to encode the multi-scale semantic features and decode the HOI instances. The reference points in the deformable transformer decoder act as the anchors for aggregating the HOI embeddings from the multi-scale context features. With the base loacation of anchors and corresponding HOI embeddings, an interaction detection head can predict the HOI instances directly. As the anchors are used throughout the HOI embeddings' decoding and the final prediction process, we call our method **Q**uery-Based **A**nchors for **HOI** detection, QAHOI. Furthermore, with the efficient attention mechanism of the deformable transformer, QAHOI with a large transformer-based backbone can be trained from scratch and outperform recent state-of-the-art methods by large margins.

2 Method

Our purpose is to address the drawbacks in the recent one-stage approaches that lack a multi-scale architecture and suffers from a poor CNN backbone for the HOI detection task. The deformable DETR [18] develops the deformable multi-scale attention module to reduce the complexity of attention in DETR to the linear complexity with the spatial size, which achieves a multi-scale transformer-based object detector. Our proposed method, QAHOI, further improves this idea

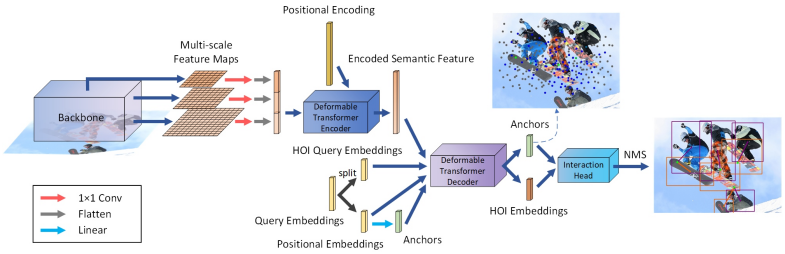

Figure 1. The overall architecture of QAHOI.

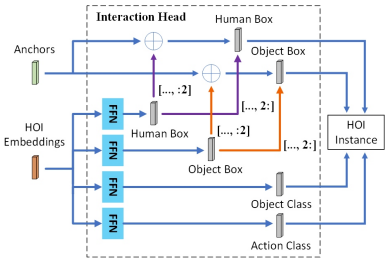

Figure 3. The interaction head predicts the HOI instances based on the anchors.

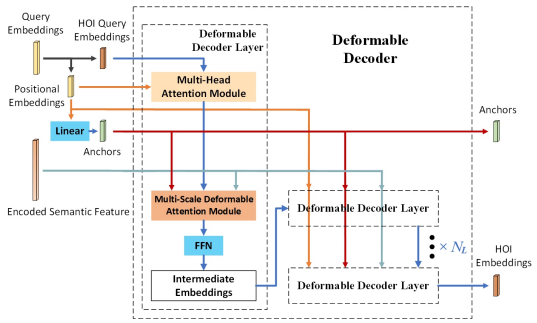

Figure 2. The decoding process of the deformable transformer decoder.

to solve HOI detection as a dense prediction problem. QAHOI adapts the deformable transformer decoder to an HOI instance detector by using the query embeddings to generate anchors and decode the HOI information. The overall architecture of QAHOI is shown in Figure 1.

2.1 Multi-Scale Feature Extractor

To improve the model's expression ability, QAHOI constructs a multi-scale feature extractor by combining a hierarchical backbone and a deformable transformer encoder [18] as shown in Figure 1. The hierarchical backbone extracts four stages' feature maps for the deformable transformer encoder, which is well designed for processing multi-scale feature maps. Specifically, given an image of size $3 \times H \times W$, QAHOI uses the last three stages' feature maps $x_1 \in \mathbb{R}^{2C_s \times \frac{H}{8} \times \frac{W}{8}}$, $x_2 \in \mathbb{R}^{4C_s \times \frac{H}{16} \times \frac{W}{16}}$ and $x_3 \in \mathbb{R}^{8C_s \times \frac{H}{32} \times \frac{W}{32}}$ of the backbone. The 1×1 convolution is used to project the feature map x_1, x_2 and x_3 from dimension C_s to dimension C_d. Then, the multi-scale feature maps x_1, x_2 and x_3 are flattened and concatenated to N_S vectors with C_d dimensions as the input of the deformable transformer encoder, where N_S is the sum of pixel numbers of the three feature maps from the backbone. A fixed positional encoding is used to indicate the scale level of the input. The deformable transformer encoder extracts the semantic feature $S \in \mathbb{R}^{N_S \times C_d}$ in a multi-scale manner and provides it for the deformable transform decoder to decode the HOI instances.

2.2 Predicting HOI with Query-Based Anchors

According to the deformable DETR, the query embeddings of the deformable transformer decoder in QAHOI are split equally into two parts, one as the HOI query embeddings $Q_{HOI} \in \mathbb{R}^{N_q \times C_d}$ and the other as the positional embeddings $Q_{Pos} \in \mathbb{R}^{N_q \times C_d}$, and the anchors $P \in \mathbb{R}^{N_q \times 2}$ are generated from positional embeddings Q_{Pos} via a linear layer. With the HOI query embeddings and the anchors, the HOI embeddings $E \in \mathbb{R}^{N_q \times C_d}$ are decoded by the deformable transformer decoder's attention mechanism with the source of the encoded semantic feature from the deformable transformer encoder. The decoding process of the deformable transformer decoder is shown in Figure 2. The self-attention of the HOI query embeddings are calculated by the multi-head attention module [16] with the positional embeddings, and the anchors aggregate the semantic feature from the output of the deformable transform encoder to calculate the multi-scale deformable attention [18] with the HOI query embeddings. Besides, after the calculation of the multi-scale deformable attention, a feed-forward network (FFN) composed of linear layers is used to process the output embeddings. The self-attention and the multi-scale attention are calculated in the stacked decode layer for N_L times, and the last layer outputs the HOI embeddings for the interaction detection head to predict the HOI instances.

QAHOI implements a simple interaction head which is similar to the QPIC [14], and the difference is that QAHOI combines each HOI embedding with a certain anchor. Hence, QAHOI feeds the decoded HOI embeddings into the interaction head to predict the HOI instances based on the anchors. Figure 3 shows the predicting process of the interaction head in QAHOI. Following the deformable DETR, each anchor (p_x, p_y) of the anchor set $P \in \mathbb{R}^{N_q \times 2}$ acts as the base point for

the bounding boxes of a pair of a human and an object. Thus, the human and object boxes B^h, $B^o \in \mathbb{R}^{N_q \times 4}$ predicted by the FFN in the interaction head are composed of $\{d_x, d_y, w, h\}$, where d_x and d_y denote the offsets between the anchor and the box's center, and w and h denote the box's width and height. Then, the final bounding boxes $\hat{B}^h$, $\hat{B}^o$ are composed of $\{d_x + p_x, d_y + p_y, w, h\}$. Finally, the object class of the object boxes $O \in \mathbb{R}^{N_q \times K_o}$ and the action class of the HOI instances $A \in \mathbb{R}^{N_q \times K_a}$ are combined with the human and object bounding boxes $\hat{B}^h$, $\hat{B}^o$ to construct the output HOI instances.

2.3 Training and Inference

Following the training procedure of the QPIC [14], the ground-truth set is padded with ϕ (no pairs) to the size of N_q, and the Hungarian algorithm [8] is used to match all of the N_q predictions with the ground-truth set. For the loss calculated on the matched pairs, the QPIC's loss function is based on the DETR [1], and because QAHOI implements the deformable DETR [18], we follow the Deformable DETR to calculate the Focal Loss [10] of the object class, which is different from the QPIC. For the anchors derived from the query embeddings, because the query embeddings are learnable parameters, the positions of the anchors are learned during training and fixed during inference.

2.4 Top K Scores and HOI NMS

QAHOI requires sufficient anchors to extract multi-scale features. In general, the number of anchors far exceeds the number of HOI instances in an image. For the HICO-DET dataset, 96% of the images contains less than 10 HOI instances. QAHOI filters the results in two steps. Firstly, the HOI instances with the top N_t object class scores are selected. Then, an HOI Non-Maximal Suppression (NMS) is used to filter out the final results. The HOI NMS is calculated based on the IoU of humans and objects between HOI instances and the HOI score. The HOI score is obtained by multiplying the object score and the action score, $c_{\text{HOI}} = c_o \cdot c_a$. And a combined IoU of human and object between an HOI instance i and j is calculated as:

$$\text{IoU}(i, j) = \text{IoU}(B_i^{(h)}, B_j^{(h)}) \cdot \text{IoU}(B_j^{(o)}, B_j^{(o)}) \qquad (1)$$

The same as the object detection task, a threshold δ is used to remove HOI instances with low scores for each action category based on the IoU.

3 Experiments

3.1 Experimental Setting

Dataset. We conduct the experiments on the HICO-DET [2] dataset, which contains 47,776 images (38,118 in the training set and 9,658 in the test set). HICO-DET has 117 action classes and 80 object classes (the object classes same as the MS-COCO [11] dataset), and the action classes and the object classes constitute 600 HOI classes. Based on the number of instances of the 600 HOI classes in the dataset, these HOI classes are divided into three categories: *Full* (all of the HOI classes), *Rare* (138 classes with less than 10 instances), and *Non-Rare* (462 classes with 10 or more than 10 instances). We report the results (in Table 1) on the Default setting (with unknown objects) and the Known Object setting (without unknown objects) of the HICO-DET.

Metric. The mean average precision (mAP) is used to evaluate the predicted HOI instances. For a true positive HOI instance, the intersection over union (IoU) between the predicted human bounding box and the ground-truth human bounding box is higher than 0.5, and the IoU between the predicted object and the ground-truth object bounding box is also higher than 0.5. As usual, we report the mAP on the *Full*, *Rare*, and *Non-Rare* categories of the HICO-DET.

Implementation Details. For the backbone, we train QAHOI with Swin-Transformer [12] pre-trained on ImageNet [4] as our best model. Specifically, we use Swin-Tiny and Swin-Base pre-trained on ImageNet-1K, and Swin-Base and Swin-Large pre-trained on ImageNet-22K. Following the setting of the Deformable DETR, the deformable transformer encoder and decoder both have 6 layers ($N_L = 6$), the number of the query embeddings is $N_q = 300$, and top $N_t = 100$ HOI instances are selected by object scores. In the NMS process, $\delta = 0.5$ is used to filter the HOI instances by the combined IoU. We use the AdamW [13] optimizer with the backbone's learning rate of 10^{-5} and other's 10^{-4}, and the weight decay of 10^{-4}. We train the model for 150 epochs with a batch size of 16 (two images per GPU, 8 GPUs), and the learning rates of the backbone and others are decayed at 120 epochs.

3.2 Comparison with State-of-the-Arts

The results compared with the state-of-the-art methods on the HICO-DET are shown in Table 1. We use QAHOI with the Swin Transformer as our best model to compare with other state-of-the-art methods. Compared with recent one-stage approaches, with the multi-scale feature maps and multi-scale deformable attention, even we do not train a detector on the MS-COCO dataset, which is beneficial for the object detection part of the model, QAHOI with Swin-Large backbone still outperforms the state-of-the-art one-stage method, QPIC with 5.88 mAP (relatively 19.7%). We found that the better the performance of the pre-trained backbone in the classification task became, the further improvement in accuracy we achieved in the HOI detection. The mAP of QAHOI with Swin-Base backbone pre-trained on ImageNet-20K is 4.1 (rel-

Method	Backbone	Fine-tuned Detection	Default			Known Object		
			Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
IP-Net [15]	ResNet-50-FPN	✗	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [9]	Hourglass-104	✓	21.73	13.78	24.10	24.58	16.65	26.84
GGNet [17]	Hourglass-104	✓	23.47	16.48	25.60	27.36	20.23	29.48
HOITrans [19]	ResNet-101	✓	26.61	19.15	28.84	29.13	20.98	31.57
HOTR [7]	ResNet-50	✗	23.46	16.21	25.65	-	-	-
HOTR [7]	ResNet-50	✓	25.10	17.34	27.42	-	-	-
AS-Net [3]	ResNet-50	✗	24.40	22.39	25.01	27.41	25.44	28.00
AS-Net [3]	ResNet-50	✓	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [14]	ResNet-101	✓	29.90	23.92	31.69	32.38	26.06	34.27
QAHOI	**Swin-Tiny**	✗	28.47	22.44	30.27	30.99	24.83	32.84
QAHOI	**Swin-Base**	✗	29.47	22.24	31.63	31.45	24.00	33.68
QAHOI	**Swin-Base***	✗	31.83	26.27	33.49	33.53	27.72	35.26
QAHOI	**Swin-Base**$^{*+}$	✗	33.58	25.86	35.88	35.34	27.24	37.76
QAHOI	**Swin-Large**$^{*+}$	✗	**35.78**	**29.80**	**37.56**	**37.59**	**31.66**	**39.36**

Table 1. Comparison with state-of-the-art on HICO-DET. Using fine-tuned detection means initializing the weights of the detection part from a model pre-trained on the MS-COCO dataset and fine-tuning the whole model on the HICO-DET dataset. The Swin-Base and Swin-Large backbone with the * and $^{+}$ are pre-trained on ImageNet-22K with 384×384 input resolution.

Arch.	#	Backbone	Fine-tuned Detection	Multi-scale	Default		
					Full	*Rare*	*Non-Rare*
QPIC	(1)	ResNet-50	✗	x_3	24.21	17.51	26.21
	(2)	ResNet-50	✓	x_3	29.07	21.85	31.23
	(3)	Swin-Tiny	✗	x_3	27.19	21.32	28.95
QAHOI	(4)	ResNet-50	✗	x_1, x_2, x_3, x_4	24.35	16.18	26.80
	(5)	ResNet-50	✓	x_1, x_2, x_3, x_4	26.18	18.06	28.61
	(6)	Swin-Tiny	✗	x_1, x_2, x_3, x_4	28.09	21.65	30.01
	(7)	Swin-Tiny	✗	x_1, x_2, x_3	28.47	22.44	30.27
	(8)	Swin-Tiny	✗	x_2, x_3	28.12	20.43	30.41
	(9)	Swin-Tiny	✗	x_3	26.65	19.13	28.89

Table 2. Evaluations of the training strategies and the effect of multi-scale feature maps and transformer-based backbone.

method	Default		
	Full	*Rare*	*Non-Rare*
base	26.64	20.62	28.44
+ topk scores	26.70	20.89	28.43
+ NMS	28.47	22.44	30.27

Table 3. Ablation study of the filtering steps. QAHOI with Swin-Tiny is used as the base method.

atively 13.9%) higher than the same backbone pre-trained on ImageNet-1K.

3.3 Ablation Study

We conduct ablation studies using CNN-based and Transformer-based backbones. For the CNN-based backbone, we use ResNet-50 and investigate the performance of two training strategies, starting from scratch and fine-tuning the weights of the detector.

Training strategies. The same as QPIC, we use the deformable DETR's weight which is trained on the MS-COCO dataset, to initialize QAHOI and then fine-tune QAHOI on the HICO-DET dataset. Following the deformable DETR's implementation, An additional low-resolution feature map $x_4 \in \mathbb{R}^{C_d \times \frac{H}{64} \times \frac{W}{64}}$ is generated by using a 3×3 convolution on the feature map x_3. We also train QAHOI and QPIC with ResNet-50 and Swin-Tiny from scratch, respectively. From the results in Table 2, without training a detector, (4) QAHOI with ResNet-50 or (7) Swin-Tiny achieves better results on the *Full* and *Non-Rare* categories compared with (1) QPIC with ResNet-50 or (3) Swin-Tiny.

Multi-scale feature maps. We use the Swin-Tiny backbone to investigate the effect of different combinations of feature maps on the accuracy of the proposed method. From the results in Table 2(6), the additional feature map does not improve the accuracy. For methods (7)(8)(9) of QAHOI, the accuracy decreases with the removal of multi-scale feature maps. Comparing (9) to (7), using the feature maps of three stages gives a model accuracy improvement of 1.82 mAP (relatively 6.8%) on the *Full* category.

CNN-based backbone vs Transformer-based backbone. The Swin-Tiny has the model size and the computation complexity similar to ResNet-50, but the accuracy on ImageNet is higher than ResNet-50. Without training an object detector, compared with the model trained with ResNet-50 in Table 2(1)(4), the transformer-based backbone Swin-Tiny improves the accuracy of both (3) QPIC (2.98 mAP, relatively 12.3%) and (7) QAHOI (4.12 mAP, relatively 16.9%), and (7) QAHOI with Swin-Tiny is better than (3) QPIC with Swin-Tiny both of the accuracy and improvement, which means our method has a great potential based on well-designed backbones. The results of QAHOI trained with Swin-Base and Swin-Large in Table 1 also show that using a backbone with higher accuracy on classification tasks can improve the accuracy of HOI detection significantly. The result of (5) QAHOI fine-tuned from Deformable DETR is lower than (2) QPIC fine-tuned from DETR. One of the reasons is that QPIC uses the DETR with 500 epochs of training, while we use the deformable DETR with only 50 epochs of training. QAHOI would have achieved better results if we have fine-tuned the deformable DETR with more epochs.

Top K scores and HOI NMS. The filtering process is important to QAHOI, in Table 3, the top K scores step and NMS step improve the accuracy on the *Full* category by 1.83 mAP.

4 Conclusion and Future Work

In this paper, we propose a transformer-based one-stage method for HOI detection, which leverages a hierarchical backbone and transformer encoder to extract the multi-scale semantic feature, a transformer decoder to decode the HOI embeddings and an interaction head to predict the HOI instances. The transformer decoder and the interaction head leverage the query-based anchors to decode the HOI embeddings and predict the HOI instances. Transformer-based backbones with the attention mechanism show a great advance for HOI detection, and the query-based anchors are also flexible in detecting the HOI instances. In the future, we will develop our method with better object detectors and further reduce the training cost.

References

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

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

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

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

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

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

[8] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

[9] 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*, 2020.

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

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018.

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

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

[16] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

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

[18] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *ICLR*, 2020.

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