

Category Query Learning for Human-Object Interaction Classification

Chi Xie^{1†} Fangao Zeng^{2‡} Yue Hu^{3‡} Shuang Liang^{1*} Yichen Wei^{2‡}
¹Tongji University ²MEGVII Technology ³Shanghai Jiao Tong University
¹{chixie, shuangliang}@tongji.edu.cn
²zfg472988436@163.com, wei_yi_chen@hotmail.com ³18671129361@sjtu.edu.cn

Abstract

Unlike most previous HOI methods that focus on learning better human-object features, we propose a novel and complementary approach called category query learning. Such queries are explicitly associated to interaction categories, converted to image specific category representation via a transformer decoder, and learnt via an auxiliary image-level classification task. This idea is motivated by an earlier multi-label image classification method, but is for the first time applied for the challenging human-object interaction classification task. Our method is simple, general and effective. It is validated on three representative HOI baselines and achieves new state-of-the-art results on two benchmarks. Code will be available at <https://github.com/charles-xie/CQL>.

1. Introduction

Human-Object Interaction (HOI) detection has attracted a lot of interests in recent years [3, 8–10, 21, 33]. The task consists of two sub-tasks. The first is human and object detection. It is usually performed by common object detection methods. The second is *interaction classification* of each human-object (HO) pair. This sub-task is very challenging due to the complex appearance variations in the interaction categories. See Fig. 1 for examples. It is the focus of most previous HOI methods, as well as this work.

Most previous HOI methods focus on *learning better human-object features*, including modeling relation and context via GNN [7, 29, 34, 37] or attention mechanism [8, 34, 46], decoupling localization and classification [22, 41, 48], leveraging vision-language knowledge [6, 22] and introducing multi-scale feature to transformer [16]. However, for interaction classification they all adopt the simple linear classifier that performs the dot product of the

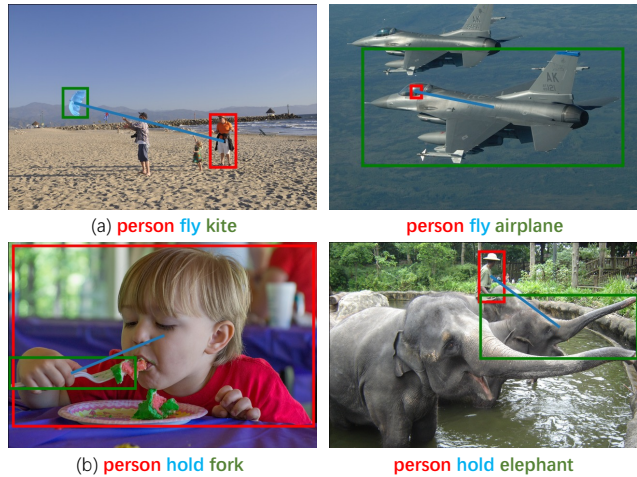


Figure 1. Interaction classification is inherently challenging. In (a), “fly” is semantically polysemic, resulting in different objects, poses and relative positions. In (b), when “hold” is associated with different objects, the appearance, scene background, and human poses are largely different.

human-object feature and a *static* weight vector, which represents an interaction category.

In this work, we propose a new approach that enhances the above paradigm and complements most previous HOI methods. It is motivated by the recent work Query2label [25], a transformer-based classification network. It proposes a new concept we call *category-specific query*. Unlike the queries in other transformer methods, each query is associated to a specific and fixed image category during training and inference. This one-to-one binding makes the query learn to model each category more effectively. The queries are converted to image specific category representations via a transformer decoder. This method achieves excellent performance on multi-label image classification task.

We extend this approach for human-object interaction classification. Essentially, our approach replaces traditional category representation as a static weight vector in previous HOI methods with category queries learnt as described above. The same linear classifier is adopted. Such cate-

*Corresponding author.

†Work done during internship at MEGVII technology.

‡Work done while worked at MEGVII.

gory queries are more effective, and *adaptive* for different images, giving rise to better modeling of the complex variations in each interaction category. This is the crucial difference between this work and a simple adaption of [25] to HOI. Notably, this work is the first to address the category weight representation problem in the HOI community.

Note that our proposed category specific query is different and not related to those queries in other transformer-based HOI methods [4, 15, 33, 49]. Specifically, category queries extract image-level features as the category representation. The queries in other methods are human-object instance-level features and category-agnostic.

Our method is simple, lightweight and general. The overview is in Fig. 2. It is complementary to any off-the-shelf HOI method that provides human-object features. The modification of both inference and training is small. The incurred additional cost is marginal.

In experiments, our approach is validated on three representative and strong HOI baseline methods, two transformer-based methods [22, 33] and a traditional two-stage method [42]. They are all significantly improved by our approach. New state-of-the-art results are obtained on two benchmarks. Specifically, we obtain 36.03 mAP on HICO-DET. Comprehensive ablation studies and in-depth discussion are also provided to verify the effectiveness of implementation details in our approach. It turns out that our method is more effective on challenging images that contain more human-object instances, a property that is rarely discussed by previous HOI methods.

2. Related Work

2.1. Instance Query Learning in HOI Detection

DETR [2] firstly proposes the concept of object instance query for object detection task. Such queries essentially learn the priors of both object appearance and spatial location. DETR leverages those queries to probe image features through a transformer [35] and localize unique objects in the image. Motivated by its great success, many works [15, 22, 33, 41, 48, 49] adapt such detection transformer framework to HOI detection by simply treating the HOI triplet [33, 49] or H-O pair [4, 15] as an object. A few of them [6, 30, 45] pay attention to adapting the plain query to this task. DOQ [30] proposes a knowledge distillation model using oracle queries to facilitate the representation learning of a transformer-based detector; HQM [45] explicitly constructs hard positive queries from ground truth to train the model to be less vulnerable to spacial variations; CATN [6] utilizes the object category prior generated from external object detector and language model for query initialization. In summary, these transformer-based methods use each query to aggregate context information not restricted to one interaction category, in order to predict

a potential HOI instance at a specific location.

In DETR [2] and its variants [4, 15, 22, 33, 41, 49] in HOI, as the queries are category-agnostic, their association to object categories are dynamic and unstable during training. This could be problematic. For example, it is well known that the convergence of DETR training is slow. In contrast, our proposed query is category-specific. The learning is guided by image-level classification task and stable. Such queries learn category-specific priors and are good representation for interaction categories.

2.2. Feature Learning in HOI Detection

Early methods. Based on two-stage detection framework, early works make many efforts to help feature learning, including employing architectures effective in modeling relation and context like GNN [7, 29, 34, 37] and attention module [8, 34, 46], leveraging fine-grained visual features [11, 17, 19, 20, 36] like human pose and introducing language prior [1, 7, 17, 27, 46].

Transformer-based methods. Motivated by DETR [49], many methods [4, 15, 33, 49] leverage transformer architecture [35] and extend the object query in DETR to HOI query. With the help of HOI query and transformer’s built-in attention mechanism, those methods learn effective feature representation for HOI triplet or H-O pair.

Based on those pioneer transformer-based methods, recently, many methods are proposed to further help feature learning, by decoupling H-O pair localization and interaction classification [22, 41, 48] or exploiting multi-scale feature in transformer architecture [16]. Some works [6, 22] leverage vision-language knowledge in CLIP [31] or design a pretrained model [40] specifically for HOI; others utilize information like human poses [39] or spatial configurations [13] that has been used in early HOI detectors.

Relation to the proposed method. Previous methods learn the H-O feature while ours learns the category query as the category representation feature. Thus, they are complementary. The interaction classification is simply by measuring the similarity between the two types of features. The integration of our method to previous HOI methods is simple.

3. Our Method

The overview of our method is in Fig. 2. It consists of two components, the image-level category query learning (top block) and human-object interaction classification (bottom right block).

The first component is detailed in Sec. 3.1. It is briefly summarized here. A number of queries (embedding vectors) are associated to human-object interaction category, in a one-to-one manner. Such queries interact with image features (provided by a baseline HOI method) through a transformer decoder [35] and become image-specific queries.

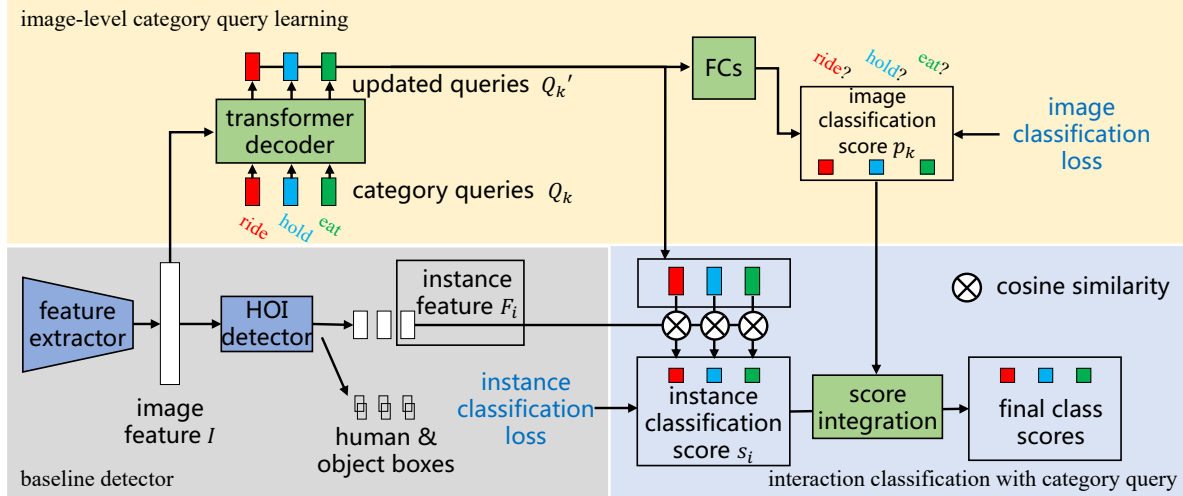


Figure 2. Overview of our method. It consists of two components (top and bottom right). It can be integrated with any baseline HOI method (bottom left) that provides image feature I and human-object instance features F_i . See Sec. 3 for details.

Learning of both the queries and decoder weights is supervised by an auxiliary image-level classification task. In this way, the queries are learnt to capture **category-specific** feature and become good feature representation for these categories. Besides some minor details, this step is the same as the previous work Query2label [25], which is for multi-label image classification task.

The second component is detailed in Sec. 3.2. For the first time, we adopt the category query learning method for human-object interaction classification tasks. The cosine similarity between the category query and human-object feature is used for interaction classification. Thus, it works with any HOI method that provides human-object features. Besides, the image-level classification results turn out moderately helpful in an score integration step, which is an extra technique that benefits the performance.

Overall, the proposed method is simple, effective, lightweight and general. It can be combined with most previous HOI detection methods (bottom left block in Fig. 2), with small modification, as elaborated in Sec. 4.

3.1. Image-level Category Query Learning

Similar as in Query2Label [25], for K human-object interaction categories, we define their one-to-one corresponding category queries, which are learnable embedding vectors, $\{Q_k\} \in \mathbb{R}^{K \times D}$, where D is the vector dimension.

Each query Q_k aggregates image features $I \in \mathbb{R}^{H \times W \times D}$ through a transformer decoder and is updated to image specific query Q'_k ,

$$\{Q'_k\} = \text{decoder}(\{Q_k\}, I). \quad (1)$$

Note that the decoder structure has several variants, which are studied in [5]. Our experiments show that the

structure is of minor importance, as discussed in Sec. 5.4. Specifically, our decoder consists of two layers, each of which consisting of a cross-attention layer, a self-attention layer and a FFN layer, in order.

Then, each image-level classification probabilities p_k is computed by applying a category-specific fully-connected layer and a sigmoid activation on the updated query Q'_k .

$$p_k = \text{sigmoid}(\text{FC}(Q'_k)) \quad (2)$$

Learning of the category query $\{Q_k\}$ and the decoder weights is supervised by common image classification losses. To deal with the label imbalance problem, focal loss [23] and asymmetric loss (ASL) [32] are used. Asymmetric loss is a variant of focal loss. It is more robust for high label imbalance and noises. Our experiments (see Sec. 5.4) show that it is slightly better.

Specifically, with the classification probability p_k and the shifted probability $p'_k = \max(p_k - m, 0)$, the asymmetric loss is

$$\mathcal{L}_{img} = \frac{1}{K} \sum_{k=1}^K \begin{cases} (1 - p_k)^{\gamma+} \log(p_k), & y_k = 1, \\ (p'_k)^{\gamma-} \log(1 - p'_k), & y_k = 0, \end{cases} \quad (3)$$

where the binary label y_k indicates the existence of category k in the image, and $\gamma+$, $\gamma-$ as well as m are hyper-parameters. We use the default values in ASL [32], $\gamma+ = 0$, $\gamma- = 4$ and $m = 0.05$.

In this way, the category queries are learnt to encode the category priors. Figure 3 is the visualization of the heatmaps in the cross-attention layer of the decoder. Each category query learns to locate the human body parts related to discriminative feature of its corresponding interaction category, e.g., in Fig. 3c, the query of “hold” highlights

the hand region, while in the same image the query of “ride” highlights the foot region. It qualitatively demonstrates that the query learning is effective in encoding **category-specific** information.

The updated queries adaptively extract category-related features for each image, with the help of the transformer decoder’s built-in multi-head cross-attention layer.

3.2. Interaction Classification with Category Query

In this step, we apply the updated category queries $\{Q'_k\}$ as the weights for interaction classification. Given the i -th human-object instance, its classification probability score for k -th interaction category is simply the cosine similarity between its feature F_i and the category query feature Q'_k

$$s_{i,k} = \text{sigmoid}\left(\frac{Q'_k \cdot F_i}{|Q'_k| \times |F_i|}\right). \quad (4)$$

There are no restrictions or assumptions on the human-object feature F_i . Most previous HOI methods should be applicable here.

In this step, the traditional classification weight from static parameters is replaced with the category query adaptive on each image. This behavior is essentially different from [25], which uses category queries as image feature.

By “adaptive”, we mean that the queries are updated dynamically according to the image contents. As exemplified in Fig. 3a and Fig. 3c, the queries of “hold” learn to highlight different interactive areas in different images and update themselves with features from these areas via attention. This shows the query learning is **adaptive to image**.

As discussed in Sec. 5.4, the classification step in Eq. (4) is crucial to make the category query learning effective on human-object interaction classification. Without this step, the image-level category query learning using only image-level classification from Query2label [25] is of little use.

Score integration step. The segmentation method in [12] discards certain categories during pixel classification that have low image classification scores. Motivated by this method, we take a similar score integration step. The image-level classification score $\{p_i\}$ is used to enhance the human-object instance classification. The idea is that, the instance score should be higher if the image-level probability is higher. Our implementation is similar as in [12]. During both training and testing, for each image, the top- κ categories ($\kappa = 70$ in this work) with higher image classification scores $\{p_i\}$ are selected. The instance score $s_{i,k}$ is slightly modulated such that it becomes higher if the rank of category k is higher. This strategy gives rise to moderate improvement, as verified in Tab. 4. We left the implementation details in the *supplementary materials*.

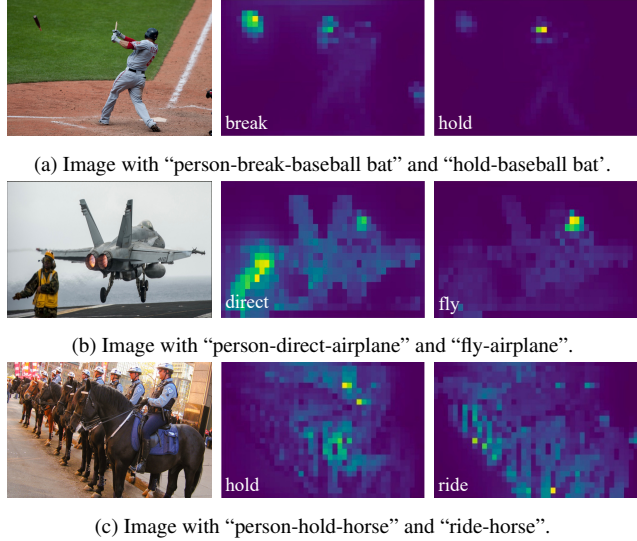


Figure 3. From left to right: image, attention maps of the cross-attention layer in the decoder for different interaction categories.

4. Integration to Off-the-shelf HOI Detectors

As shown in Fig. 2, our method is ready to integrate with any baseline HOI method that provides image feature I and human-object instance feature $\{F_i\}$. The integration is simple. During inference, the human-object instance interaction classification part is replaced by our method in Sec. 3.2, the top and bottom right block in Fig. 2.

During training, the original loss in the baseline HOI method \mathcal{L}_{base} is added to our image classification loss in Eq. (3). The final loss \mathcal{L} for training is

$$\mathcal{L} = \mathcal{L}_{base} + \lambda * \mathcal{L}_{img}, \quad (5)$$

where the weight λ is 1.0 by default. All other hyper parameters and details during training remain the same as in the baseline HOI method.

Thus, our method is general and applicable to most existing HOI methods. In this work, we select three representative yet different baseline methods to verify the effectiveness of our approach, as described below.

QPIC [33] is the first to introduce transformer method into HOI task. It is also the baseline for many recent works [13, 28, 41, 45, 48]. Its performance is much better than early one-stage [14, 21, 38] and two-stage [7, 11, 17] methods while keeping a simple and end-to-end architecture. It consists of a CNN backbone as well as a transformer encoder and decoder.

During our integration, the feature map in its transformer encoder is used as the image feature I . The human-object feature $\{F_i\}$ is the query feature in its decoder.

SCG [42] is a traditional two-stage method and the best in this category. It is also one of the best method that

Method	Pipeline	E2E	HICO-DET			V-COCO		Efficiency	
			Full	Rare	Non-Rare	S1	S2	#Params	FPS
QPIC [33]	transformer	✓	28.93	21.62	31.12	61.39	63.65	41M	19.5
+ <i>Ours</i>	transformer	✓	31.08(+2.15)	23.90	33.22	63.67(+2.28)	65.49	46M(+5M)	18.3(-6.2%)
SCG [42]	two-stage	✗	31.28	24.16	33.40	56.93	62.51	57M	4.5
+ <i>Ours</i>	two-stage	✗	32.74(+1.46)	26.25	34.68	59.14(+2.21)	65.61	64M(+7M)	4.1(-8.9%)
GEN-VLKT [22]	transformer	✗	33.69	29.94	34.81	64.89	66.74	42M	21.7
+ <i>Ours</i>	transformer	✗	35.36(+1.67)	32.97	36.07	66.40(+1.51)	69.17	47M(+5M)	20.6(-5.1%)

Table 1. The performance numbers of three different baseline HOI methods with and without integration of our method, on two datasets. “E2E” denotes whether a HOI detector is end-to-end. All models are tested on Tesla V100.

does not use transformer. It uses a multi-stream graph neural network(GNN) for interaction classification. In our experiment, the detection boxes are from a fine-tuned detector provided by DRG [7] for HICO-DET and a fine-tuned DETR for V-COCO.

During our integration, the CNN feature map in the backbone of SCG is used as image feature I . The human-object feature $\{F_i\}$ is generated through ROI pooling with detected human and object boxes and fused with the GNN.

GEN-VLKT [22] is also transformer-based, but not end-to-end as pairwise NMS [41] is used for post-processing. It is the current state-of-the-art method. It uses two parallel decoders for object detection and interaction classification, namely instance decoder and interaction decoder.

During our integration, the feature map in its transformer encoder is used as image feature I . The query feature in the interaction decoder is used as the human-object feature $\{F_i\}$. Note that, unlike the majority of HOI detection methods, the original GEN-VLKT uses HOI categories rather than interaction categories during interaction classification. Our experiments still use interaction categories, in order to be consistent with most other methods.

5. Experiments

In this section, we verify the applicability and effectiveness of the proposed method through experiments. In Sec. 5.1, we introduce the experimental settings. Then we demonstrate the effectiveness of the proposed method over 3 baselines in Sec. 5.2, and show it achieves SOTA results on major benchmarks in Sec. 5.3. Next, in Sec. 5.4 we conduct comprehensive ablation studies on the key components as well as detailed technical designs. Lastly, we provide some analysis and visualization in Sec. 5.5.

5.1. Datasets

HICO-DET [3] and V-COCO [10] are two widely-used HOI benchmarks. HICO-DET contains 47,776 images, with 38,118 for training and 9,658 for testing. There are 600 HOI categories in HICO-DET, consisting of 117 interaction classes and 80 object classes. Each HOI category is composed of an interaction and an object. V-COCO is

a subset of MS-COCO [24] with HOI annotations, including 10,346 images (2,533 for training, 2,867 for validation and 4,946 for testing). It has 80 object categories same with HICO-DET and 29 interaction categories.

Evaluation metrics. For HICO-DET, we adopt the commonly used mAP metric [3]. Each prediction is a \langle human, interaction, object \rangle triplet. A prediction is a true positive only when the human and object bounding boxes both have IoU >0.5 w.r.t. ground truth and the interaction classification result is correct. We evaluate the performance in two different settings following [3]. In the *known object* setting, for each HOI category, we evaluate the prediction only on the images containing the target object category. In *default* setting, the detection result of each category is evaluated on the full test set. In each setting, we report the mAP over (1) all 600 HOI categories (Full), (2) 138 categories with less than 10 training samples (Rare), and (3) the remaining 462 categories (Non-rare). For V-COCO, we use the role mAP following [10], under both scenario #1 (including objects) and #2 (ignoring objects). The performance is evaluated using its official evaluation toolkit.

5.2. Improvement on Three Different Baselines

Tab. 1 summarizes the performance of the three baseline HOI methods before and after integration of our method. Backbone is ResNet50. All these methods are significantly improved. Specifically, QPIC [33] is improved by 2.15 mAP, making it competitive with those more recent works [16, 41, 43]. SCG [42] is improved by 1.46 mAP, demonstrating that our method is not limited to transformer-based baselines. The current SOTA method GEN-VLKT [22] is improved by 1.67 mAP, producing the new SOTA result (also refer to Tab. 2).

On the V-COCO dataset [10], the performance improvement is similar, which is 2.28, 2.21 and 1.51 mAP on QPIC, SCG and GEN-VLKT, respectively.

Notably, for SCG [42], as the object detector is fixed during the training of its interaction classification network, the improvement by our method is purely due to better interaction classification, not a better fine-tuned CNN backbone or a better object detector. This further consolidates that the category query learning is effective.

Method	Detector	Backbone	Default			Known Object		
			Full	Rare	Non-rare	Full	Rare	Non-rare
DRG [7]	HICO-DET	ResNet50-FPN	24.53	19.47	26.04	27.98	23.11	29.43
GG-Net [47]	HICO-DET	Hourglass104	23.47	16.48	25.60	27.36	20.23	29.48
IDN [18]	HICO-DET	ResNet50	26.29	22.61	27.39	28.24	24.47	29.37
QPIC [33]	HICO-DET	ResNet50	29.07	21.85	31.23	31.68	24.14	33.93
SCG [42]	HICO-DET	ResNet50-FPN	31.33	24.72	33.31	34.37	27.18	36.52
CDN [41]	HICO-DET	ResNet50	31.78	27.55	33.05	34.53	29.73	35.96
DT [48]	HICO-DET	ResNet50	31.75	27.45	33.03	34.50	30.13	35.81
STIP [44]	HICO-DET	ResNet50	31.60	27.75	32.75	34.41	30.12	35.69
HQM [45]	HICO-DET	ResNet50	32.47	28.15	33.76	-	-	-
MSTR [16]	HICO-DET	ResNet50	31.17	25.31	32.92	34.02	28.83	35.57
RLIP [40]	COCO+VG	ResNet50	32.84	26.85	34.63	-	-	-
IF [26]	HICO-DET	ResNet50	33.51	30.30	34.46	36.28	33.16	37.21
GEN-VLKT-B [22]	HICO-DET	ResNet50	33.75	29.25	35.10	36.78	32.75	37.99
GEN-VLKT-M [22]	HICO-DET	ResNet101	34.78	31.50	35.77	38.07	34.94	39.01
GEN-VLKT-L [22]	HICO-DET	ResNet101	34.95	31.18	36.08	38.22	34.36	39.37
BodyPartMap [39]	HICO-DET	ResNet50	35.15	33.71	35.58	37.56	35.87	38.06
GEN-VLKT-B + <i>Ours</i>	HICO-DET	ResNet50	35.36	32.97	36.07	38.43	34.85	39.50
GEN-VLKT-M + <i>Ours</i>	HICO-DET	ResNet101	35.83	32.91	36.70	38.79	35.28	39.84
GEN-VLKT-L + <i>Ours</i>	HICO-DET	ResNet101	36.03	33.16	36.89	38.82	35.51	39.81

Table 2. The proposed method achieves state-of-the-art on HICO-DET [3]. The best results are marked in **bold**.

Method	Backbone	Scenario #1	Scenario #2
DRG [7]	R50FPN	51.0	-
SCG [42]	R50	54.2	60.9
GG-Net [47]	HG104	54.7	-
QPIC [33]	R50	58.8	61.0
HQM [45]	R50	63.6	-
CDN [41]	R50	61.7	63.8
GEN-VLKT-B [22]	R50	62.4	64.5
GEN-VLKT-M [22]	R101	63.3	65.6
GEN-VLKT-L [22]	R101	63.6	65.9
MSTR [16]	R50	62.0	65.2
BodyPartMap [39]	R50	63.0	65.1
IF [26]	R50	63.0	65.2
DT [48]	R50	66.2	68.5
STIP [44]	R50	65.1	69.7
GEN-VLKT-B + <i>Ours</i>	R50	66.4	69.2
GEN-VLKT-M + <i>Ours</i>	R101	66.8	69.8
GEN-VLKT-L + <i>Ours</i>	R101	66.5	69.9

Table 3. Comparison with state-of-the-art methods on V-COCO [10] dataset. The best results are marked in **bold**.

To verify that the performance improvement is not due to a larger model, we also compare the model size and running speed. Our method increases the model parameters by a few millions, which is small compared to the original model size. The running speed, measured by FPS, is only decreased by a few percent. The marginal additional cost shows that our method is quite lightweight.

5.3. Comparison with State-of-the-art

Tab. 2 and Tab. 3 compare our method with many previous methods for HICO-DET and V-COCO datasets, respec-

	C1	C2	C3	Full	Rare	Non-Rare
a	-	-	-	33.69	29.94	34.81
b	✓	-	-	33.86 (+0.17)	31.12	34.68
c	✓	✓	-	34.98 (+1.29)	31.73	35.95
d	✓	✓	✓	35.36 (+1.67)	32.97	36.07

Table 4. Ablation study of several variants of our method, starting from the baseline (a) to our approach (d). **C1**, **C2** and **C3** are described in Sec. 5.4. The best results are marked in **bold**.

tively. GEN-VLKT [22] is used as our baseline.

On HICO-DET, our result with ResNet50 backbone already outperforms all previous methods under both *default* and *known object* settings. With the stronger ResNet101 backbone, our method achieves the new state-of-the-art 36.03 full mAP under *default* settings and 38.82 under *known object* settings.

On V-COCO dataset, our method achieves the new state-of-the-art performance on Scenario 1, with an AP of 66.4 for ResNet50, surpassing [48]. For scenario 2, it is comparable with the state-of-the-art [44].

5.4. Ablation Experiments

We perform various ablation experiments to validate the effectiveness of different components in our method. HICO-DET dataset and GEN-VLKT [22] baseline are used.

The proposed method can be divided into 3 components: **C1** means applying Query2Label [25] to the baseline detector as a multi-task learning (with feature extractor shared). In detail, it adds the queries, the decoder and image classification loss (Eq. (1), Eq. (2) and Eq. (5)). **C2** means using the learned query in **C1** as adaptive interaction classification

loss type	λ	Full	Rare	Non-Rare
-	0	34.21	30.15	35.42
focal loss [23]	0.5	34.43	31.06	35.44
	1.0	34.51	31.08	35.53
	1.5	34.35	31.54	35.19
	2.0	34.29	31.18	35.22
ASL [32]	0.5	34.57	31.91	35.36
	1.0	34.98	31.73	35.95
	1.5	34.77	32.08	35.57
	2.0	34.41	31.92	35.15

Table 5. Ablation on the type and weight λ of the image classification loss. The best results are marked in **bold**.

Layer structure	Full	Rare	Non-Rare
S \rightarrow C \rightarrow F	34.73	32.09	35.52
C \rightarrow S \rightarrow F	34.98	31.73	35.95
C \rightarrow F	34.63	31.34	35.61

Table 6. Ablation on structure of each layer in the category decoder. Here ‘‘S’’, ‘‘C’’ and ‘‘F’’ stands for the self-attention, cross-attention and FFN in a standard transformer [35].

L	Full	Rare	Non-Rare
1	34.66	31.47	35.61
2	34.98	31.73	35.95
3	34.86	31.56	35.85

Table 7. Ablation on the number of layers in the category decoder.

weight (Eq. (4)). This is the key component of the proposed method, which makes a distinction between the proposed method and a simple adaption of Query2label [25] to HOI task. **C3** denotes the score integration step in Sec. 3.2.

To validate our approach in Fig. 2, several variants with these components in our approach are experimented and summarized in Tab. 4. First, (b) is a simple combination of Query2label [25] and the baseline detector (a) in a multi-task setting. (b) is only slightly better than (a), showing that simply applying Query2Label to HOI is barely helpful.

Second, variant (c) significantly boosts (b), indicating using the queries as adaptive classification weights is the key to the performance improvement. This shows the effectiveness of our most crucial technical design in **C2**: applying the learned category queries as adaptive interaction classification weights.

Finally, the complete approach is (d). It further adds the integration step of image classification and instance classification score on (c). This technique produces moderate improvement over (c), i.e., 1.67 mAP vs. 1.29 mAP.

Image classification loss. Tab. 5 compares different loss functions and weights. First, when $\lambda = 0$, which means no image-level supervision is applied, the improvement over the baseline (33.69 mAP) drops to only 0.52 mAP. This demonstrates the image classification supervision is essen-

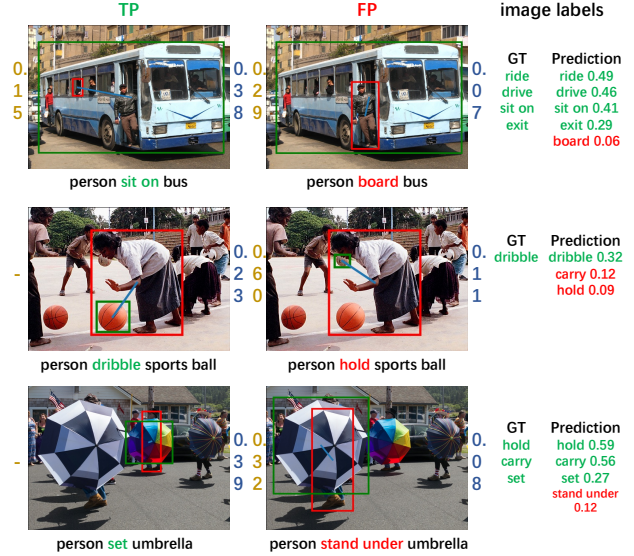


Figure 4. Some qualitative comparison between the baseline and the proposed method on HICO-DET. From left to right, column 1: true positive (TP) detection results, whose interaction score is increased by our method; column 2: false positive (FP) detection results, whose interaction score is decreased by our method; column 3: corresponding image-level GT and predictions by our method. Scores on the left and right of an image are the interaction classification scores of the visualized instance from the baseline and our method. Best viewed in color. More in *supplementary materials*.

tial to make the category query learning effective, with either focal loss or ASL. Additionally, ASL is slightly better than focal loss. By default, $\lambda = 1.0$ is adopted.

Is asymmetric loss the key? In the ablation above, we can see that ASL does help our image-level query learning. However, it is not the major reason for the performance improvement. To figure out this, we replace focal loss in the plain baseline GEN-VLKT with ASL and the result is only slightly better by 0.08 mAP.

Decoder structure. Tab. 6 compares several structures of the decoder. Compared to the standard decoder [2, 35] (S \rightarrow C \rightarrow F in the table), putting cross-attention first (C \rightarrow S \rightarrow F) is slightly better (by 0.25 mAP) without extra computation. If we remove self-attention (C \rightarrow F), the performance drops by 0.35 mAP compared with the S \rightarrow C \rightarrow F setting. This is probably because self-attention helps to learn the dependencies between different category queries.

Tab. 7 compares different numbers of decoder layers, denoted as L . We find that $L = 2$ is sufficient. More layers do not help the performance.

5.5. Discussions and Analysis

To understand why the category query learning is effective for human-object interaction classification, we provide some analysis and qualitative results.

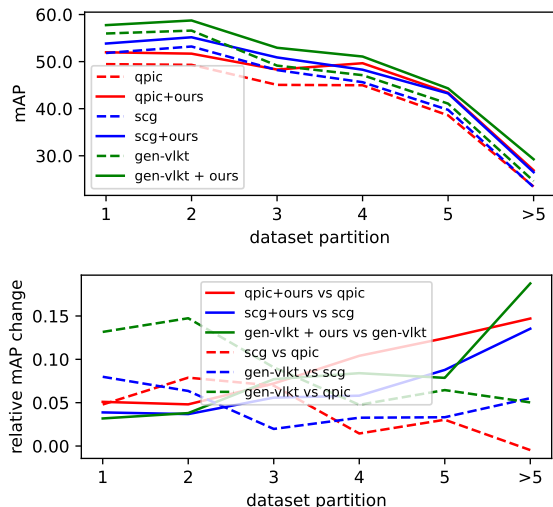


Figure 5. Performance evaluated on image partitions with different $n = 1, 2, 3, 4, 5, > 5$ instances for an interaction category. Top: mAP numbers for three baselines (dashed) as well as our integration (solid). Bottom: relative mAP ratios for improvement produced by our method (solid) and between two arbitrary previous methods (dashed).

The attention maps in the cross-attention layer of the decoder is visualized in Fig. 3. For different category queries, the corresponding attention maps show they learn to capture the semantics of the category, while being adaptive to different images. For example, in Fig. 3a, the broken part of the bat in the air is highlighted for “break” while the left part in the hand is highlighted for “hold”. This is similar for Fig. 3b. In Fig. 3c, the attention map highlights many instances with corresponding action “hold” and “ride”.

To qualitatively demonstrate how our method helps, we visualize some cases of the baseline and the proposed method in Fig. 4. In the first case, the baseline predicts a TP of “person sit on bus” with a low score 0.15, and a FP of “person board bus” with a high score 0.29. Our model also predicts the same TP and FP, but lowers the FP score to 0.07 and lifts the TP score to 0.38. Besides, our model correctly predicts image-level scores: the wrong category “board” is given a low score of 0.06 while the four correct categories are given high scores. In the second case, our model successfully predicts the TP “person dribble” missed by the baseline and suppress the FP “person hold sports” from 0.6 to 0.11 with the help of correct image-level classification result. This is similar for the third case.

Last, as our category query learning is performed on image level, we conjecture that it is more helpful for images with dense human-object interactions. In such images, a human-object instance is relatively small and hard to learn good feature on its own. However, it may benefit more from the global image level category query feature, which aggregates more information from other similar instances in this

image. To validate this conjecture, we partition the images according to their “interaction density” and check whether our method produces larger improvement on images that are “denser”. Specifically, for each interaction category, its mAP is evaluated separately on six different image partition subsets, where each image contains different numbers ($n=1, 2, 3, 4, 5$ and > 5) of human-object instances of this category. The mAP results of the three baseline methods and their integrated versions (as in Tab. 1) are shown in Fig. 5 (top, dashed vs. solid lines). It shows that: 1) the mAP is lower for larger n , indicating the “denser” images are more challenging; 2) our method improves the baselines consistently on all different partitions.

To check whether the proposed method is more effective on “denser” images, we use the relative mAP improvement, which is a ratio, $\frac{mAP_{ours} - mAP_{baseline}}{mAP_{baseline}}$, for analysis. The ratio curves of the three baselines are shown in Fig. 5 (bottom, solid lines). It is clear that the relative improvement becomes larger for larger n . This indicates that the image-level category query learning is more effective on these challenging dense images.

To further verify that this behavior is not commonly true, we also compute the relative ratio of three more comparisons, “gen-vlkt vs. qpic”, “gen-vlkt vs. scg” and “scg vs. qpic”, in which the former outperforms the latter. These curves are also shown in Fig. 5 (bottom, dashed lines). There is no clear pattern in these curves, indicating that the performance gap between two arbitrary HOI methods are in general not related to the image “density”.

6. Conclusion

This work proposes a novel approach for the human-object interaction classification sub-task in HOI detection. We study the problem of interaction category modeling, in contrast to most previous methods focusing on human-object feature learning. We adopt the concept of category query in a previous method [25] for HOI, for the first time, and show that it is simple, general and highly effective.

Clearly, this idea of category query modeling is not limited to multi-label image classification and HOI detection. We hope it is useful for other vision tasks.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 62076183, 61936014 and 61976159, in part by the Natural Science Foundation of Shanghai under Grant 20ZR1473500, in part by the Shanghai Science and Technology Innovation Action Project of under Grant 20511100700 and 22511105300, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100, and in part by the Fundamental Research Funds for the Central Universities. The authors would also like to thank the anonymous reviewers for their careful work and valuable suggestions.

References

- [1] A Bansal, S. S Rambhatla, A Shrivastava, and R Chellappa. Detecting human-object interactions via functional generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10460–10469, 2020. [2](#)
- [2] N Carion, F Massa, G Synnaeve, N Usunier, A Kirillov, and S Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. [2](#), [7](#)
- [3] Y.-W Chao, Y Liu, X Liu, H Zeng, and J Deng. Learning to detect human-object interactions. In *2018 IEEE winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. [1](#), [5](#), [6](#)
- [4] M Chen, Y Liao, S Liu, Z Chen, F Wang, and C Qian. Reformulating hoi detection as adaptive set prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9004–9013, 2021. [2](#)
- [5] B Cheng, I Misra, A. G Schwing, A Kirillov, and R Girshick. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. [3](#)
- [6] L Dong, Z Li, K Xu, Z Zhang, L Yan, S Zhong, and X Zou. Category-aware transformer network for better human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19538–19547, 2022. [1](#), [2](#)
- [7] C Gao, J Xu, Y Zou, and J.-B Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. [1](#), [2](#), [4](#), [5](#), [6](#)
- [8] C Gao, Y Zou, and J.-B Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018. [1](#), [2](#)
- [9] G Gkioxari, R Girshick, P Dollár, and K He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018. [1](#)
- [10] S Gupta and J Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. [1](#), [5](#), [6](#)
- [11] T Gupta, A Schwing, and D Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9677–9685, 2019. [2](#), [4](#)
- [12] H He, Y Yuan, X Yue, and H Hu. Rankseg: Adaptive pixel classification with image category ranking for segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 682–700. Springer Nature Switzerland Cham, 2022. [4](#)
- [13] A Iftexhar, H Chen, K Kundu, X Li, J Tighe, and D Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5353–5363, 2022. [2](#), [4](#)
- [14] B Kim, T Choi, J Kang, and H. J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020. [4](#)
- [15] B Kim, J Lee, J Kang, E.-S Kim, and H. J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 74–83, 2021. [2](#)
- [16] B Kim, J Mun, K.-W On, M Shin, J Lee, and E.-S Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022. [1](#), [2](#), [5](#), [6](#)
- [17] D.-J Kim, X Sun, J Choi, S Lin, and I. S Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020. [2](#), [4](#)
- [18] Y.-L Li, X Liu, X Wu, Y Li, and C Lu. Hoi analysis: Integrating and decomposing human-object interaction. *Advances in Neural Information Processing Systems*, 33:5011–5022, 2020. [6](#)
- [19] Y.-L Li, L Xu, X Liu, X Huang, Y Xu, S Wang, H.-S Fang, Z Ma, M Chen, and C Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020. [2](#)
- [20] Y.-L Li, S Zhou, X Huang, L Xu, Z Ma, H.-S Fang, Y Wang, and C Lu. Transferable interactiveness knowledge for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3594, 2019. [2](#)
- [21] Y Liao, S Liu, F Wang, Y Chen, C Qian, and J Feng. 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*, pages 482–490, 2020. [1](#), [4](#)
- [22] Y Liao, A Zhang, M Lu, Y Wang, X Li, and S Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. [1](#), [2](#), [5](#), [6](#)
- [23] T.-Y Lin, P Goyal, R Girshick, K He, and P Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. [3](#), [7](#)
- [24] T.-Y Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, and C. L Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. [5](#)
- [25] S Liu, L Zhang, X Yang, H Su, and J Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#), [8](#)
- [26] X Liu, Y.-L Li, X Wu, Y.-W Tai, C Lu, and C.-K Tang. Interactiveness field in human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20113–20122, 2022. [6](#)
- [27] Y Liu, Q Chen, and A Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020. [2](#)

- [28] J Park, S Lee, H Heo, H. K Choi, and H. J Kim. Consistency learning via decoding path augmentation for transformers in human object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1019–1028, 2022. 4
- [29] S Qi, W Wang, B Jia, J Shen, and S.-C Zhu. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 401–417, 2018. 1, 2
- [30] X Qu, C Ding, X Li, X Zhong, and D Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19558–19567, 2022. 2
- [31] A Radford, J. W Kim, C Hallacy, A Ramesh, G Goh, S Agarwal, G Sastry, A Askell, P Mishkin, J Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2
- [32] T Ridnik, E Ben-Baruch, N Zamir, A Noy, I Friedman, M Protter, and L Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 3, 7
- [33] M Tamura, H Ohashi, and T Yoshinaga. 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*, pages 10410–10419, 2021. 1, 2, 4, 5, 6
- [34] O Ulutan, A Iftekhar, and B. S Manjunath. 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*, pages 13617–13626, 2020. 1, 2
- [35] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A. N Gomez, Ł Kaiser, and I Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 7
- [36] B Wan, D Zhou, Y Liu, R Li, and X He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 2
- [37] H Wang, W.-s Zheng, and L Yingbiao. Contextual heterogeneous graph network for human-object interaction detection. In *European Conference on Computer Vision*, pages 248–264. Springer, 2020. 1, 2
- [38] T Wang, T Yang, M Danelljan, F. S Khan, X Zhang, and J Sun. Learning human-object interaction detection using interaction points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4116–4125, 2020. 4
- [39] X Wu, Y.-L Li, X Liu, J Zhang, Y Wu, and C Lu. Mining cross-person cues for body-part interactiveness learning in hoi detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 121–136. Springer, 2022. 2, 6
- [40] H Yuan, J Jiang, S Albanie, T Feng, Z Huang, D Ni, and M Tang. Rlip: Relational language-image pre-training for human-object interaction detection. In *Advances in Neural Information Processing Systems*, 2022. 2, 6
- [41] A Zhang, Y Liao, S Liu, M Lu, Y Wang, C Gao, and X Li. Mining the benefits of two-stage and one-stage hoi detection. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2, 4, 5, 6
- [42] F. Z Zhang, D Campbell, and S Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13319–13327, 2021. 2, 4, 5, 6
- [43] F. Z Zhang, D Campbell, and S Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20104–20112, 2022. 5
- [44] Y Zhang, Y Pan, T Yao, R Huang, T Mei, and C.-W Chen. Exploring structure-aware transformer over interaction proposals for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19548–19557, 2022. 6
- [45] X Zhong, C Ding, Z Li, and S Huang. Towards hard-positive query mining for detr-based human-object interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 444–460. Springer, 2022. 2, 4, 6
- [46] X Zhong, C Ding, X Qu, and D Tao. Polysemy deciphering network for robust human-object interaction detection. *International Journal of Computer Vision*, 129(6):1910–1929, 2021. 1, 2
- [47] X Zhong, X Qu, C Ding, and D Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13234–13243, 2021. 6
- [48] D Zhou, Z Liu, J Wang, L Wang, T Hu, E Ding, and J Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022. 1, 2, 4, 6
- [49] C Zou, B Wang, Y Hu, J Liu, Q Wu, Y Zhao, B Li, C Zhang, C Zhang, Y Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11825–11834, 2021. 2