Towards Hard-Positive Query Mining for DETR-based Human-Object Interaction Detection

Xubin Zhong[1], Changxing Ding[1,2*], Zijian Li[1], and Shaoli Huang[3]

[1] South China University of Technology , Guangzhou, China
[2] Pazhou Lab, Guangzhou, China
[3] Tencent AI-Lab, Shenzhen, China
eexubin@mail.scut.edu.cn, chxding@scut.edu.cn,
eezijianli@mail.scut.edu.cn, shaolihuang@tencent.com

Abstract. Human-Object Interaction (HOI) detection is a core task for high-level image understanding. Recently, Detection Transformer (DETR)-based HOI detectors have become popular due to their superior performance and efficient structure. However, these approaches typically adopt fixed HOI queries for all testing images, which is vulnerable to the location change of objects in one specific image. Accordingly, in this paper, we propose to enhance DETR's robustness by mining hard-positive queries, which are forced to make correct predictions using partial visual cues. First, we explicitly compose hard-positive queries according to the ground-truth (GT) position of labeled human-object pairs for each training image. Specifically, we shift the GT bounding boxes of each labeled human-object pair so that the shifted boxes cover only a certain portion of the GT ones. We encode the coordinates of the shifted boxes for each labeled human-object pair into an HOI query. Second, we implicitly construct another set of hard-positive queries by masking the top scores in cross-attention maps of the decoder layers. The masked attention maps then only cover partial important cues for HOI predictions. Finally, an alternate strategy is proposed that efficiently combines both types of hard queries. In each iteration, both DETR's learnable queries and one selected type of hard-positive queries are adopted for loss computation. Experimental results show that our proposed approach can be widely applied to existing DETR-based HOI detectors. Moreover, we consistently achieve state-of-the-art performance on three benchmarks: HICO-DET, V-COCO, and HOI-A. Code is available at https://github.com/MuchHair/HQM.

Keywords: Human-Object Interaction, Detection Transformer, Hard Example Mining

1 Introduction

Human-Object Interaction (HOI) detection is a fundamental task for human-centric scene understanding [2,3,51,53,54]. It aims to infer a set of HOI triplets

[*] Corresponding author.

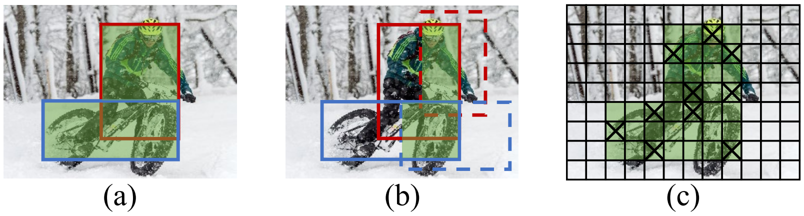


Fig. 1. Illustration of the hard-positive queries. (a) The green area represents important visual cues for HOI prediction of one human-object pair. (b) The dashed boxes are produced via Ground-truth Bounding-box Shifting (GBS), which only cover part of important image area and are then encoded into hard-positive queries. (c) Part of important visual cues are removed via Attention Map Masking (AMM), which increases the prediction difficulty of one positive query to infer HOI triplets. Best viewed in color.

$< human, interaction, object >$ from a given image [1,2]. In other words, it involves identifying not only the categories and locations of objects in an individual image, but also the interactions between each human–object pair. Recently, Detection Transformer (DETR)-based methods [4–7, 52] have become popular in the field of HOI detection due to their superior performance and efficient structure. These methods typically adopt a set of learnable queries, each of which employs the cross-attention mechanism [33] to aggregate image-wide context information in order to predict potential HOI triplets at specific locations.

However, the learnable queries are usually weight-fixed after training [4,52]. Since each query targets a specific location [5, 8], DETR-based methods are typically sensitive to changes in the object locations in testing images. Recent works improve DETR's robustness through the use of adaptive queries. For example, CDN [7] performs human-object pair detection before interaction classification occurs, generating adaptive interaction queries based on the output of the object detection part. However, its queries for object detection remain fixed. Moreover, two other object detection works [39, 43] opt to update each object query according to the output embedding of each decoder layer. An object query is typically formulated as a single reference point of one potential object. Notably, this strategy may not be easy to apply in the context of HOI detection, since the interaction area for one human-object pair is usually more complex to formulate [13, 28, 29]. Therefore, current DETR-based methods still suffer from poor-quality queries.

In this paper, we enhance the robustness of DETR-based HOI detection methods from a novel perspective, namely that of Hard-positive Query Mining (HQM). In our approach, the robustness refers to the ability of DETR model to correctly predict HOI instances even using poor-quality queries with limited visual cues (or say inaccurate position). Accordingly, a hard-positive query refers to a query that corresponds to one labeled human-object pair, but is restricted to employ limited visual cues to make correct HOI predictions. First, as illustrated in Fig. 1(b), we explicitly generate such queries via Ground-truth Bounding-box Shifting (GBS). In more detail, we shift the two ground-truth (GT) bounding boxes in one labeled human-object pair so that each shifted box covers only a certain portion of its GT box. We then encode the coordinates of the two shifted boxes into an HOI query. Accordingly, the resultant query contains only rough location information about the pair. This strategy models the extreme conditions caused by variations in object location for fixed HOI queries.

Second, as shown in Fig. 1(c), we increase the prediction difficulty of positive queries by means of Attention Map Masking (AMM). The positive queries in AMM are those DETR's learnable queries matched with ground-truth according to bipartite matching [35]. In more detail, for each positive query, a proportion of the top scores in the cross-attention maps are masked. In this way, the positive query employs only part of the visual cues for prediction purposes. Specially, as our goal is to enhance the robustness of learnable queries, we select the masked elements according to the value of their counterparts in the learnable queries. Queries generated via GBS and AMM are less vulnerable to overfitting and capable of producing valuable gradients for DETR-based models. Finally, the robustness of DETR-based models is enhanced for the test images.

During each iteration of the training stage, the DETR's learnable queries and our hard-positive queries are utilized sequentially for prediction with shared model parameters and loss functions. To promote efficiency, GBS and AMM are alternately selected in each iteration. This Alternate Joint Learning (AJL) strategy is more efficient and achieves better performance than other joint learning strategies. Moreover, during inference, both GBT and AMM are removed; therefore, our method does not increase the complexity of DETR-based models in the testing stage.

To the best of our knowledge, HQM is the first approach that promotes the robustness of DETR-based models from the perspective of hard example mining. Moreover, HQM is plug-and-play and can be readily applied to many DETR-based HOI detection methods. Exhaustive experiments are conducted on three HOI benchmarks, namely HICO-DET [2], V-COCO [1], and HOI-A [13]. Experimental results show that HQM not only achieves superior performance, but also significantly accelerates the training convergence speed.

2 Related Works

Human-Object Interaction Detection. Based on the model architecture adopted, existing HOI detection approaches can be divided into two categories: Convolutional Neural Networks (CNN)-based methods [11, 13, 25] and transformer-based methods [4–7, 30].

CNN-based methods can be further categorized into two-stage approaches [10, 11, 14, 17, 25, 26] and one-stage approaches [13, 28, 29]. In general terms, two-stage approaches first adopt a pre-trained object detector [9] to generate human and object proposals, after which they feed the features of human-object pairs into verb classifiers for interaction prediction. Various types of features can be utilized to improve interaction classification, including human pose [10, 11], human-object spatial information [15, 21, 24], and language features [14, 25, 26].

Although two-stage methods are flexible to include diverse features, they are usually time-consuming due to the cascaded steps. By contrast, one-stage methods are usually more efficient because they perform object detection and interaction prediction in parallel [13,28,29]. These methods typically depend on predefined interaction areas for interaction prediction. For example, UnionDet [29] used the union box of a human-object pair as interaction area, while PPDM [13] employed a single interaction point to represent interaction area. Recently, GGNet [28] utilized a set of dynamic points to cover larger interaction areas. However, the above predefined interaction areas may not fully explore the image-wide context information.

Recently, the transformer architecture has become popular for HOI detection. Most such methods are DETR-based [4–7,30,44,45]. These methods can be further divided into two categories: methods that employ one set of learnable queries for both object detection and interaction classification [4,6,44,45], and methods that utilize separate sets of queries for object detection and interaction prediction [5,7,30]. The above methods have achieved superior performance through their utilization of image-wide context information for HOI prediction. However, due to their use of weight-fixed queries, their performance is usually sensitive to location change of humans or objects.

DETR-based Object Detection. The DETR model realizes end-to-end object detection by formulating the task as a set prediction problem [8]. However, due to its use of weight-fixed and semantically obscure HOI queries queries, it suffers from slow training convergence [39–43]. To solve this problem, recent works have largely adopted one of two main strategies. The first of these is to impose spatial priors on attention maps in the decoder layers to reduce semantic ambiguity. For example, Dynamic DETR [41] estimates a Region of Interest (ROI) based on the embedding of each decoder layer, then constrains the cross-attention operation in the next decoder layer within the ROI region. The second strategy involves updating queries according to the output decoder embeddings from each decoder layer [39,43]. Each query in these works is typically formulated through a single reference point of the object instance. However, it may not be straightforward to make similar formulations in the context of HOI detection. This is because HOI detection is a more challenging task that involves not only the detection of a single object, but also detection of the human instance and interaction category.

In this paper, we enhance the robustness of DETR-based models from a novel perspective, namely that of hard-positive query mining. Compared with existing approaches, our method is easier to implement and does not increase model complexity during inference. In the experimentation section, we further demonstrate that our approach achieves better performance than existing methods.

Hard Example Mining. HQM can be regarded as a hard example mining (HEM) approach to transformer-based HOI detection. HEM has demonstrated its effectiveness in improving the inference accuracy of CNN-based object detection models [49,50]. However, this strategy has rarely been explored in HOI detection. Recently, Zhong et al. [28] devised a hard negative attentive loss to

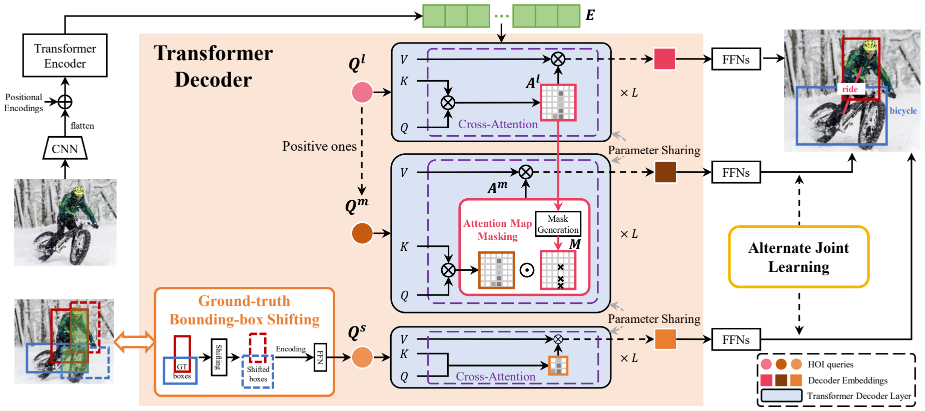


Fig. 2. Overview of HQM in the training stage based on QPIC [4]. In the interest of simplicity, only one learnable query $\mathbf{Q}^l$ and two hard-positive queries are illustrated. The hard-positive query $\mathbf{Q}^s$ is produced by GBS that encodes the coordinates of shifted bounding boxes of one human-object pair into a query. Another positive query $\mathbf{Q}^m$ is selected from the learnable queries according to bipartite matching with the ground-truth. The cross-attention maps of $\mathbf{Q}^m$ are partially masked to increase prediction difficulty. The two types of hard-positive queries are alternately selected in each iteration, and the chosen type of queries are utilized together with the learnble queries for loss computation. $\otimes$ and $\odot$ denote matrix multiplication and Hadamard product, respectively. In the inference stage, HQM is removed and therefore brings no extra computational cost. Best viewed in color.

overcome the problem of class imbalance between positive and negative samples for keypoint-based HOI detection models [12, 13]. In comparison, HQM modifies DETR's model architecture in the training stage and resolves the problem caused by its use of poor-quality queries.

3 Method

HQM is plug-and-play and can be applied to many existing DETR-based HOI detection models. [4–7,30,44,45]. In this section, we take the representative work QPIC [4] as an example. The overall framework of QPIC [4] equipped with HQM is shown in Fig. 2. In the following, we first give a brief review of QPIC, which is followed by descriptions of two novel hard-positive query mining methods, e.g., GBS (Section 3.2) and AMM (Section 3.3). Finally, we introduce an alternate joint learning strategy to apply GBS and AMM (Section 3.4) efficiently.

3.1 Overview of Our Method

Revisit to QPIC. As shown in Fig. 2, QPIC is constructed by a CNN-based backbone, a transformer encoder, a transformer decoder, and feed-forward networks (FFNs). Each input image $\mathbf{I} \in \mathbb{R}^{H_0 \times W_0 \times 3}$ is first fed into the CNN backbone and the transformer encoder to extract flattened visual features $\mathbf{E} \in \mathbb{R}^{(H \times W) \times D}$. QPIC then performs cross-attention between learnable queries $\mathbf{Q}^l \in \mathbb{R}^{N_q \times D}$ and $\mathbf{E}$ in the transformer decoder. N_q is the number of learnable queries fixed after training. $H \times W$ and D denote the number of the image patches and feature dimension for each patch, respectively. Besides, the transformer decoder is typically composed of multiple stacked layers. For clarity, we only present the cross-attention operation in one decoder layer. The output embedding $\mathbf{C}_i \in \mathbb{R}^{N_q \times D}$ of the i-th decoder layer can be formulated as follows:

$$\mathbf{C}_i = \text{Concat}([\mathbf{A}_h^l \mathbf{E} \mathbf{W}_h^V]_{h=1}^T), \tag{1}$$

$$\mathbf{A}_h^l = \text{Softmax}(Att_h(\mathbf{Q}^l, \mathbf{C}_{i-1}, \mathbf{E})), \tag{2}$$

where T is the number of cross-attention heads. $\mathbf{A}_h^l \in \mathbb{R}^{N_q \times (H \times W)}$ is the normalized cross-attention map for the h-th head. $\mathbf{W}_h^V$ is a linear projection matrix. $Att_h(\cdot)$ is a function for similarity computation. Finally, as illustrated in Fig. 2, each output decoder embedding is sent to detection heads based on FFNs to obtain object class scores, interaction category scores, and position of the human and object instances.

Hard-Positive Query Mining. Most DETR-based HOI detection methods, e.g., QPIC, adopt poor-quality queries after training. As analyzed in Section 1, performance of weight-fixed queries is sensitive to the location change of human and object instances in testing images. In the following, we propose to promote the robustness of DETR-based models via hard-positive query mining. One hard-positive query refers to a query that corresponds to one labeled human-object pair, but is restricted to employ limited visual cues to infer correct HOI triplets (shown in Fig. 1).

As illustrated in Fig. 2, two strategies are introduced to produce hard-positive queries, i.e., GBS and AMM. Their generated queries are denoted as $\mathbf{Q}^s \in \mathbb{R}^{N_g \times D}$ and $\mathbf{Q}^m \in \mathbb{R}^{N_g \times D}$, respectively. N_g is the number of labeled human-object pairs in one training image. Similar to $\mathbf{Q}^l$, $\mathbf{Q}^s$ and $\mathbf{Q}^m$ are sent to the transformer decoder and their output decoder embeddings are forced to infer correct HOI triplets. $\mathbf{Q}^l$, $\mathbf{Q}^s$ and $\mathbf{Q}^m$ share all model layers and loss functions.

3.2 Ground-truth Bounding-box Shifting

Previous works have shown that each query attends to specific locations in one image [5, 8]. To enhance DETR's robustness against the spatial variance of human-object pairs, we here explicitly compose hard-positive queries $\mathbf{Q}^s$ according to the GT position of labeled human-object pairs for each training image. As shown in Fig. 1(b), we shift the bounding boxes of labeled human-object pairs such that the shifted boxes only contain partial visual cues for HOI prediction.

Specifically, we encode a hard-positive query $\mathbf{q}^s \in \mathbf{Q}^s$ for one labeled human-object pair as follows:

$$\mathbf{q}^s = L_n(F_p(Shift(\mathbf{p}^s))), \tag{3}$$

Algorithm 1 Attention Map Masking for Each Attention Head

```
Input: attention maps $\mathbf{A}^m$, $\mathbf{A}^l \in \mathbb{R}^{H \times W}$ for a hard-positive query, $K$, $\gamma$
Get the indices $I_K$ of the top-$K$ elements in $\mathbf{A}^l$
Initialize a random binary mask $\mathbf{M} \in \mathbb{R}^{H \times W}$: $\mathbf{M}_{i,j} \sim Bernoulli(\gamma)$
for $\mathbf{M}_{i,j} \in \mathbf{M}$ do
    if $(i, j) \notin I_K$ then
        $\mathbf{M}_{i,j} = 1$
    end if
end for
Output: Masked attention map $\mathbf{A}^m = \mathbf{A}^m \odot \mathbf{M}$
```

where

$$\mathbf{p}^s = [x_h, y_h, w_h, h_h, x_o, y_o, w_o, h_o, x_h - x_o, y_h - y_o, w_h h_h, w_o h_o]^T. \tag{4}$$

The first eight elements in $\mathbf{p}^s$ are the center coordinates, width, and height of one GT human-object pair, respectively. $[x_h - x_o,\ y_h - y_o]$ denotes the relative position between the two boxes, respectively; while the last two elements are the areas of the two boxes. $Shift(\cdot)$ represents the shifting operation to GT bounding boxes (as illustrated in Fig. 2). $F_p(\cdot)$ is an FFN with two layers whose dimensions are both D. It projects $\mathbf{p}^s$ to another D-dimensional space. $L_n(\cdot)$ is a $tanh$ normalization function, which ensures the amplitudes of elements in $\mathbf{q}^s$ and the positional embeddings for $\mathbf{E}$ are consistent [52].

Compared with one concurrent work DN-DETR [55], GBS focuses on hard-positive query mining. To ensure the query is both positive and hard, we control the Intersection-over-Unions (IoUs) between each shifted box and its ground truth. We adopt low IoUs ranging from 0.4 to 0.6 in our experiment and find that GBS significantly improves the inference performance of DETR-based models.

3.3 Attention Map Masking

One popular way to enhance model robustness is Dropout [34]. However, applying Dropout directly to features or attention maps of $\mathbf{Q}^l$ may cause interference to bipartite matching [35], since the feature quality of the query is artificially degraded. To solve this problem, we implicitly construct another set of hard-positive queries $\mathbf{Q}^m$ via AMM after the bipartite matching of $\mathbf{Q}^l$. Queries in $\mathbf{Q}^m$ are copied from the positive queries in $\mathbf{Q}^l$ according to results by bipartite matching. As shown in Fig. 2, to increase the prediction difficulty of $\mathbf{Q}^m$, some elements in the cross-attention maps for each query in $\mathbf{Q}^m$ are masked. In this way, each query in $\mathbf{Q}^m$ is forced to capture more visual cues from the non-masked regions.

Detailed operations of AMM are presented in Algorithm 1. For clarity, only one hard-positive query $\mathbf{q}^m \in \mathbf{Q}^m$ is taken as an example, whose attention maps are denoted as $\mathbf{A}_h^m$ $(1 \leq h \leq T)$. For simplicity, we drop the subscripts of both $\mathbf{A}_h^m$ and $\mathbf{A}_h^l$ in the following.

AMM has two parameters, i.e., K and γ. Since our ultimate goal is to enhance the robustness of $\mathbf{Q}^l$ rather than $\mathbf{Q}^m$, we select dropped elements in $\mathbf{A}^m$ according to the value of their counterparts in $\mathbf{A}^l$. Specifically, we first select the top K elements according to the value in $\mathbf{A}^l$. Then, we randomly mask the selected K elements in $\mathbf{A}^m$ with a ratio of γ.

Discussion. AMM is related but different from Dropout and its variants [34,36]. Their main difference lies in the way to select dropped elements. First, AMM drops elements with high values, while Dropout drops elements randomly. Second, AMM requires a reference, i.e., $\mathbf{A}^l$, for dropped element selection in $\mathbf{A}^m$. Dropout requires no reference. In the experimentation section, we show that AMM achieves notably better performance than the naive Dropout.

3.4 Alternate Joint Learning

The two hard-positive query mining methods, i.e., GBS and AMM, can be applied jointly to generate diverse hard queries. However, as DETR-based HOI detection methods typically require large number of training epochs to converge, it is inefficient to adopt both methods together in each iteration. We here present the Alternate Joint Learning (AJL) strategy, in which GBS and AMM are applied alternately for each training iteration. Concretely, the learnable queries of DETR and our hard queries are fed into the transformer decoder sequentially. The main reason for this lies in the design of AMM. The masked attention scores for hard queries are selected according to those of learnable queries (in Section 3.3). Therefore, learnable queries should pass the model first to provide attention scores. Compared to applying GBS and AMM together for each iteration, AJL is more efficient and achieves better performance in our experiments.

Overall Loss Function. We adopt the same loss functions for object detection and interaction prediction as those in QPIC [4]. The overall loss function in the training phase can be represented as follows:

$$\mathcal{L} = \alpha \mathcal{L}_l + \beta \mathcal{L}_h, \tag{5}$$

where

$$\mathcal{L}_l = \lambda_b \mathcal{L}_{l_b} + \lambda_u \mathcal{L}_{l_u} + \lambda_c \mathcal{L}_{l_c} + \lambda_a \mathcal{L}_{l_a}, \tag{6}$$

$$\mathcal{L}_h = \lambda_b \mathcal{L}_{h_b} + \lambda_u \mathcal{L}_{h_u} + \lambda_c \mathcal{L}_{h_c} + \lambda_a \mathcal{L}_{h_a}. \tag{7}$$

$\mathcal{L}_l$ and $\mathcal{L}_h$ denote the loss for learnable queries and hard-positive queries, respectively. $\mathcal{L}_{k_b}$, $\mathcal{L}_{k_u}$, $\mathcal{L}_{k_c}$, and $\mathcal{L}_{k_a}$ $(k \in \{l, h\})$ denote the L1 loss, GIOU loss [47] for bounding box regression, cross-entropy loss for object classification, and focal loss [48] for interaction prediction, respectively. These loss functions are realized in the same way as in [4]. Moreover, both α and β are set as 1 for simplicity; while λ_b, λ_u, λ_c and λ_a are set as 2.5, 1, 1, 1, which are the same as those in [4].

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

HICO-DET. HICO-DET [2] is the most popular large-scale HOI detection dataset, which provides more than 150,000 annotated instances. It consists of 38,118 and 9,658 images for training and testing, respectively. There are 80 object categories, 117 verb categories, and 600 HOI categories in total.

V-COCO. V-COCO [1] was constructed based on the MS-COCO database [31]. The training and validation sets contain 5,400 images in total, while its testing set includes 4,946 images. It covers 80 object categories, 26 interaction categories, and 234 HOI categories. The mean average precision of Scenario 1 role (mAP_{role}) [1] is commonly used for evaluation.

HOI-A. HOI-A was recently proposed in [13]. The images are collected from wild; it is composed of 38,629 images, with 29,842 used for training and 8,787 for testing. HOI-A contains 11 object categories and 10 interaction categories.

4.2 Implementation Details

We adopt ResNet-50 [32] as our backbone model. Following QPIC [4], we initialize parameters of our models using those of DETR that was pre-trained on the MS-COCO database as an object detection task. We adopt the AdamW [46] optimizer and conduct experiments with a batch size of 16 on 8 GPUs. The initial learning rate is set as 1e-4 and then decayed to 1e-5 after 50 epochs; the total number of training epochs is 80. N_q and D are set as 100 and 256, respectively. For GBS, the IoUs between the shifted and ground-truth bounding boxes range from 0.4 to 0.6; while for AMM, K and γ are set as 100 and 0.4, respectively.

4.3 Ablation Studies

We perform ablation studies on HICO-DET, V-COCO, and HOI-A datasets to demonstrate the effectiveness of each proposed component. We adopt QPIC [4] as the baseline and all experiments are performed using ResNet-50 as the backbone. Experimental results are tabulated in Table 1.

Effectiveness of GBS. GBS is devised to explicitly generate hard-positive queries leveraging the bounding box coordinates of labeled human-object pairs. When GBS is incorporated, performance of QPIC is promoted by 1.50%, 1.39% and 1.13% mAP on HICO-DET, V-COCO, and HOI-A datasets, respectively. Moreover, as shown in Fig. 3(a), GBS also significantly accelerates the training convergence of QPIC. It justifies the superiority of GBS in improving DETR-based HOI detectors. We further evaluate the optimal values of IoUs and provide experimental results in the supplementary material.

Effectiveness of AMM. AMM is proposed to implicitly construct hard-positive queries using masking to cross-attention maps. As illustrated in Table 1, the performance of QPIC is notably improved by 1.51%, 1.48%, and 1.20% mAP on HICO-DET, V-COCO, and HOI-A datasets, respectively. Furthermore,

Table 1. Ablation studies on each key component of HQM. For HICO-DET, the DT mode is adopted for evaluation.

	Components					mAP			# Epochs
Method	GBS	AMM	CJL	PJL	AJL	HICO-DET	V-COCO	HOI-A	HICO-DET
Baseline	-	-	-	-	-	29.07	61.80	74.10	150
Incremental	✓	-	-	-	-	30.57	63.19	75.23	80
	-	✓	-	-	-	30.58	63.28	75.30	80
	✓	✓	✓	-	-	30.11	63.03	75.01	80
	✓	✓	-	✓	-	30.81	63.39	75.59	80
Our Method	✓	✓	-	-	✓	**31.34**	**63.60**	**76.13**	80

as shown in Fig. 3(b), AMM also significantly reduces the number of training epochs required on the HICO-DET dataset. We also provide a detailed analysis for K and γ values in the supplementary material.

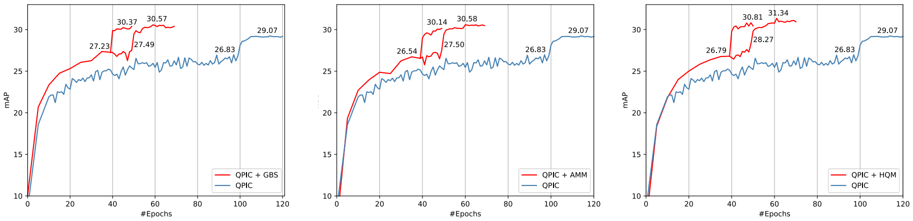
Combination of GBS and AMM. We here investigate three possible strategies that combine GBS and AMM for more effective DETR training, namely, Cascaded Joint Learning (CJL), Parallel Joint Learning (PJL) and Alternate Joint Learning (AJL).

Cascaded Joint Learning. In this strategy, we formulate GBS and AMM as two successive steps to produce one single set of hard-positive queries. In more details, we first apply GBS to produce one set of hard-positive queries. Then, we apply AMM to cross-attention maps of queries generated by GBS. As shown in Table 1, CJL achieves worse performance than the model using GBS or AMM alone. This may be because queries generated by CJL contain rare cues for HOI prediction, thereby introducing difficulties in optimizing DETR-based models.

Parallel Joint Learning. In this strategy, GBS and AMM are employed to generate one set of hard-positive queries, respectively. Then, both sets of hard-positive queries are employed for HOI prediction. To strike a balance between the loss of learnable queries and hard-positive queries, the loss weight for each type of hard-positive queries is reduced by one-half. Moreover, they are independent, which means there is no interaction between these two types of queries. As shown in Table 1, PJL achieves better performance than the model using GBS or AMM alone. Moreover, it outperforms QPIC by 1.74%, 1.59%, and 1.49% mAP on HICO-DET, V-COCO, and HOI-A datasets, respectively. However, PJL lowers the computational efficiency due to the increased number of hard-positive queries.

Alternate Joint Learning. In this strategy, GBS and AMM are applied alternately for each training iteration. The learnable queries of DETR and our hard-positive queries are fed into the transformer decoder sequentially, meaning there is no interference between each other. As tabulated in Table 1, AJL outperforms other joint learning strategies. AJL also has clear advantages in efficiency compared with PJL. Moreover, it significantly promotes the performance of QPIC by 2.27%, 1.80%, and 2.03% mAP on the three datasets, respectively. The above experimental results justify the effectiveness of AJL.

Application to Other DETR-based Models. Both GBS and AMM are plug-and-play methods that can be readily applied to other DETR-based HOI



(a) (b) (c)

Fig. 3. The mAP and training convergence curves for QPIC and our method on HICO-DET. Our method significantly improves QPIC in both mAP accuracy and convergence rate.

Table 2. Effectiveness of GBS and AMM on HOTR and CDN in the DT Mode of HICO-DET.

Incremental Components				mAP		
Baseline	GBS	AMM	AJL	Full	Rare	Non-rare
HOTR	-	-	-	23.46	16.21	25.62
	✓	-	-	24.67	23.29	25.34
	-	✓	-	24.73	23.52	25.09
	✓	✓	✓	**25.69**	**24.70**	**25.98**
CDN	-	-	-	31.44	27.39	32.64
	✓	-	-	32.07	27.52	33.43
	-	✓	-	32.05	27.15	33.51
	✓	✓	✓	**32.47**	**28.15**	**33.76**

Table 3. Comparisons with variants of GBS on HICO-DET.

	Full	Rare	Non-rare
QPIC [4]	29.07	21.85	31.23
w/o $Shift(\cdot)$	29.61	22.67	31.68
w Gaussian noise	30.05	24.08	31.82
QPIC + GBS	**30.57**	**24.64**	**32.34**

detection models, e.g., HOTR [5] and CDN [7]. The main difference between HOTR [5] and QPIC is that HOTR performs object detection and interaction prediction in parallel branches with independent queries. Here, we mainly apply HQM to its interaction detection branch. As presented in Table 2, HOTR+GBS (AMM) outperform HOTR by 1.21% (1.27%) mAP in DT mode for the full HOI categories. When AJL is adopted, the performance of HOTR is considerably improved by 2.23%, 8.49% and 0.36% mAP in DT mode for the full, rare and non-rare HOI categories, respectively. Besides, significant improvements can also be observed by applying our methods to CDN. Impressively, when incorporated with our method, performance of CDN is promoted by 1.03% mAP for the full HOI categories.

4.4 Comparisons with Variants of GBS and AMM

Comparisons with Variants of GBS. We compare the performance of GBS with its two possible variants. Experimental results are tabulated in Table 3.

First, 'w/o $Shift(\cdot)$' means removing the box-shifting operation in Eq. (4). This indicates that the ground-truth position of one human-object pair is leveraged for query encoding. Therefore, the obtained queries in this setting are easy-positives rather than hard-positives. It is shown that the performance of this variant is lower than our GBS by 0.96%, 1.97% and 0.66% mAP in DT mode for the full, rare and non-rare HOI categories respectively. This experimental result provides direct evidence for the effectiveness of hard-positive queries.

Table 4. Comparisons with variants of AMM on HICO-DET.

	Full	Rare	Non-rare
QPIC [4]	29.07	21.85	31.23
w/o top-K	30.06	24.10	31.84
w/o $\mathbf{A}^l$	30.11	24.28	31.85
w/o $\mathbf{Q}^m$	28.75	21.97	30.78
QPIC [4] + AMM	**30.58**	**25.48**	**32.10**

Table 5. Performance Comparisons on HOI-A. D-based is short for DETR-based.

	Methods	Backbone	mAP
CNN-based	iCAN [25]	ResNet-50	44.23
	TIN [10]	ResNet-50	48.64
	GMVM [38]	ResNet-50	60.26
	C-HOI [37]	ResNet-50	66.04
	PPDM [13]	Hourglass-104	71.23
D-based	AS-Net [30]	ResNet-50	72.19
	QPIC [4]	ResNet-50	74.10
	QPIC [4] + **HQM**	ResNet-50	**76.13**

Second, 'w Gaussian noise' represents that we remove $Shift(\cdot)$ and add Gaussian noise to $\mathbf{q}^s$ in Eq. 3. This variant provides another strategy to generate hard-positive queries. Table 3 shows that GBS outperforms this variant by 0.52% mAP for the full HOI categories. The main reason is that operations in GBS are more explainable and physically meaningful than adding random Gaussian noise.This experiment justifies the superiority of GBS for producing hard-positive queries.

Comparisons with Variants of AMM. We here compare the performance of AMM with some possible variants, namely, 'w/o top-K', 'w/o $\mathbf{A}^l$', and 'w/o $\mathbf{Q}^l$'. Experimental results are tabulated in Table 4.

First, 'w/o top-K' is a variant that randomly masks elements rather than the large-value elements in an attention map with the same γ ratio. We can observe that the performance of this variant is lower than AMM by 0.52% in terms of DT mAP for the full HOI categories. Compared with this variant, AMM is more challenging since visual cues are partially removed. Therefore, AMM forces each query to explore more visual cues in unmasked regions, which avoids overfitting. This experiment further demonstrates the necessity of mining hard queries.

Second, 'w/o $\mathbf{A}^l$' means that we select the masked elements according to $\mathbf{A}^m$ rather than $\mathbf{A}^l$ in Algorithm 1. Compared with AMM, the mAP of this variant drops by 0.47%, 1.20%, 0.25% for the full, rare and non-rare HOI categories. This may be because the learnable queries rather than the hard-positive queries are employed during inference. Therefore, masking according to $\mathbf{A}^l$ can push the hard-positive queries to explore complementary features to those attended by the learnable ones. In this way, more visual cues can be mined and the inference power of learnable queries can be enhanced during inference.

Finally, 'w/o $\mathbf{Q}^m$' indicates that we apply the same masking operations as AMM to the attention maps of $\mathbf{Q}^l$ rather than those of $\mathbf{Q}^m$. In this variant, $\mathbf{Q}^m$ are removed and only $\mathbf{Q}^l$ are adopted as queries. It is shown that the performance of this setting is significantly lower than those of AMM. As analyzed in Section 3.3, applying dropout directly to attention maps of $\mathbf{Q}^l$ may degrade the quality of their decoder embeddings, bringing in interference to bipartite matching and therefore causing difficulties in optimizing the entire model.

4.5 Comparisons with State-of-the-art Methods

Comparisons on HICO-DET. As shown in Table 6, our method outperforms all state-of-the-art methods by considerable margins. Impressively, QPIC + HQM outperforms QPIC by 2.27%, 4.69%, and 1.55% in mAP on the full, rare

Table 6. Performance comparisons on HICO-DET.

	Methods	Backbone	DT Mode		
			Full	Rare	Non-Rare
CNN-based	InteractNet [27]	ResNet-50-FPN	9.94	7.16	10.77
	UnionDet [29]	ResNet-50-FPN	17.58	11.72	19.33
	PD-Net [16]	ResNet-152	20.81	15.90	22.28
	DJ-RN [15]	ResNet-50	21.34	18.53	22.18
	PPDM [13]	Hourglass-104	21.73	13.78	24.10
	GGNet [28]	Hourglass-104	23.47	16.48	25.60
	VCL [19]	ResNet-101	23.63	17.21	25.55
DETR-based	HOTR [5]	ResNet-50	23.46	16.21	25.62
	HOI-Trans [6]	ResNet-50	23.46	16.91	25.41
	AS-Net [29]	ResNet-50	28.87	24.25	30.25
	QPIC [4]	ResNet-50	29.07	21.85	31.23
	ConditionDETR [39]	ResNet-50	29.65	22.64	31.75
	CDN-S [7]	ResNet-50	31.44	27.39	32.64
	HOTR [5] + **HQM**	ResNet-50	**25.69**	**24.70**	**25.98**
	QPIC [4] + **HQM**	ResNet-50	**31.34**	**26.54**	**32.78**
	CDN-S [7] + **HQM**	ResNet-50	**32.47**	**28.15**	**33.76**

Table 7. Performance comparisons on V-COCO.

	Methods	Backbone	AP_{role}
CNN-based	DRG [18]	ResNet-50-FPN	51.0
	PMFNet [24]	ResNet-50-FPN	52.0
	PD-Net [16]	ResNet-152	52.6
	ACP [20]	ResNet-152	52.9
	FCMNet [22]	ResNet-50	53.1
	ConsNet [23]	ResNet-50-FPN	53.2
	InteractNet [27]	ResNet-50-FPN	40.0
	UnionDet [29]	ResNet-50-FPN	47.5
	IP-Net [12]	Hourglass-104	51.0
	GGNet [28]	Hourglass-104	54.7
DETR-based	HOI-Trans [6]	ResNet-101	52.9
	AS-Net [30]	ResNet-50	53.9
	HOTR [5]	ResNet-50	55.2
	QPIC [4]	ResNet-50	58.8
	CDN-S [7]	ResNet-50	61.7
	QPIC [4] + **HQM**	ResNet-50	**63.6**

and non-rare HOI categories of the DT mode, respectively. When our method is applied to HOTR and CDN-S with ResNet-50 as the backbone, HOTR (CDN-S) + HQM achieves a 2.23% (1.03%) mAP performance gain in DT mode for the full HOI categories over the HOTR (CDN-S) baseline. These experiments justify the effectiveness of HQM in enhancing DETR's robustness. Comparisons under the KO mode are presented in the supplementary material.

Moreover, we compare the performance of HQM with Conditional DETR [39]. Conditional DETR relieves the weight-fixed query problem via updating queries according to decoder embeddings in each decoder layer. We extend this approach to HOI detection by using an interaction point to represent one potential human-object pair. To facilitate fair comparison, all the other settings are kept the same for HQM and Conditional DETR. Table 6 shows that HQM achieves better performance. The above experiments justify the superiority of HQM for improving the robustness of DETR-based models in HOI detection.

Comparisons on HOI-A. Comparison results on the HOI-A database are summarized in Table 5. The same as results on HICO-DET, our approach outperforms all state-of-the-art methods. In particular, QPIC + HQM significantly outperforms QPIC by 2.03% in mAP when the same backbone adopted.

Comparisons on V-COCO. Comparisons on V-COCO are tabulated in Table 7. It is observed that our method still outperforms all other approaches, achieving 63.6% in terms of AP_{role}. These experiments demonstrate that HQM can consistently improve the robustness of DETR-based models on HICO-DET, VCOCO, and HOI-A datasets.

4.6 Qualitative Visualization Results

As presented in Fig. 4, we visualize some HOI detection results and cross-attention maps from QPIC (the first row) and QPIC + HQM (the second row). It can be observed that QPIC + HQM captures richer visual cues. One main reason for this may be that QPIC + HQM is trained using hard-positive queries. Therefore, QPIC + HQM is forced to mine more visual cues to improve the model prediction accuracy during inference. More qualitative comparisons results are provided in the supplementary material.

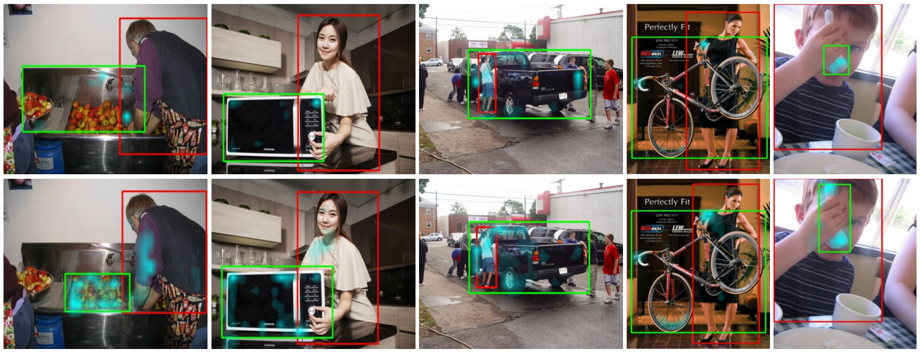

wash apple operate microwave wash truck carry bicycle lick spoon

Fig. 4. Visualization of HOI detection results and the cross-attention maps in one decoder layer on HICO-DET. Images in the first and second rows represent results for QPIC and QPIC+HQM, respectively. Best viewed in color.

5 Conclusions

This paper promotes the robustness of existing DETR-based HOI detection models. We creatively propose Hard-positive Queries Mining (HQM) that enhances the robustness of DETR-based models from the perspective of hard example mining. HQM is composed of three key components: Ground-truth Bounding-box Shifting (GBS), Attention Map Masking (AMM), and Alternate Joint Learning (AJL). GBS explicitly encodes hard-positive queries leveraging coordinates of shifted bounding boxes of labeled human-object pairs. At the same time, AMM implicitly constructs hard-positive queries by masking high-value elements in the cross-attention scores. Finally, AJL is adopted to alternately select one type of the hard positive queries during each iteration for efficiency training. Exhaustive ablation studies on three HOI datasets are performed to demonstrate the effectiveness of each proposed component. Experimental results show that our proposed approach can be widely applied to existing DETR-based HOI detectors. Moreover, we consistently achieve state-of-the-art performance on three benchmarks: HICO-DET, V-COCO, and HOI-A.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62076101 and 61702193, the Program for Guangdong Introducing Innovative and Entrepreneurial Teams under Grant 2017ZT07X183, Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011549, and Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004.

References

1. S. Gupta, and J. Malik. Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015) 2, 3, 9
2. Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng. Learning to detect human-object interactions. In: WACV (2018) 1, 2, 3, 9
3. J. Ji, R. Krishna, L. Fei-Fei, and J. Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In: CVPR (2020) 1
4. M. Tamura, H. Ohashi, and T. Yoshinaga. QPIC: Query-Based Pairwise Human-Object Interaction Detection with Image-Wide Contextual Information. In: CVPR (2021) 2, 3, 4, 5, 8, 9, 11, 12, 13
5. B. Kim, J. Lee, J. Kang, E. Kim, and H. Kim. HOTR: End-to-End Human-Object Interaction Detection with Transformers. In: CVPR (2021) 2, 3, 4, 5, 6, 11, 13
6. C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei, End-to-end human object interaction detection with hoi transformer. In: CVPR (2021) 2, 3, 4, 5, 13
7. 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: NeurIPS (2021) 2, 3, 4, 5, 11, 13
8. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In: ECCV (2020) 2, 4, 6
9. S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015) 3
10. Y. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H. Fang, Y. Wang, and C. Lu. Transferable Interactiveness Knowledge for Human-Object Interaction Detection. In: CVPR (2019) 3, 12
11. T. Gupta, A. Schwing, and D. Hoiem. No-Frills Human-Object Interaction Detection: Factorization, Layout Encodings, and Training Techniques. In: ICCV (2019) 3
12. T. Wang, T. Yang, M. Danelljan, F. Khan, X. Zhang, and J. Sun. Learning Human-Object Interaction Detection using Interaction Points. In: CVPR (2020) 5, 13
13. 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: CVPR (2020) 2, 3, 4, 5, 9, 12, 13
14. O. Ulutan, A. Iftekhar, and B. Manjunath. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In: CVPR (2020) 3
15. Y. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, and C. Lu. Detailed 2D-3D Joint Representation for Human-Object Interaction. In: CVPR (2020) 3, 13
16. X. Zhong, C. Ding, X. Qu, and D. Tao. Polysemy deciphering network for human-object interaction detection. In: ECCV (2020) 13
17. X. Zhong, C. Ding, X. Qu, and D. Tao. Polysemy deciphering network for robust human-object interaction detection. In: IJCV (2021) 3
18. C. Gao, J. Xu, Y. Zou, and J. Huang. DRG: Dual Relation Graph for Human-Object Interaction Detection. In: ECCV (2020) 13
19. Z. Hou, X. Peng, Y. Qiao, and D. Tao. Visual Compositional Learning for Human-Object Interaction Detection. In: ECCV (2020) 13
20. D. Kim, X. Sun, J. Choi, S. Lin, and I. Kweon. Detecting Human-Object Interactions with Action Co-occurrence Priors. In: ECCV (2020) 13
21. P. Zhou, and M. Chi. Relation parsing neural network for human-object interaction detection. In: ICCV (2019) 3
22. Y. Liu, Q. Chen, and A. Zisserman. Amplifying key cues for human-object-interaction detection. In: ECCV (2020) 13

23. Y. Liu, J. Yuan, and C. Chen. ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection. In: ACM MM (2020) 13
24. B. Wan, D. Zhou, Y. Liu, R. Li, and X. He. Pose-aware Multi-level Feature Network for Human Object Interaction Detection. In: ICCV (2019) 3, 13
25. C. Gao, Y. Zou, and J. Huang. ican: Instance-centric attention network for human-object interaction detection. In: BMVC (2018) 3, 12
26. T. Wang, R. Anwer, M. Khan, F. Khan, Y. Pang, L. Shao, and J. Laaksonen. Deep Contextual Attention for Human-Object Interaction Detection. In: ICCV (2019) 3
27. G. Gkioxari, R. Girshick, Detecting and recognizing human-object interactions. In: CVPR (2018) 13
28. X. Zhong, X. Qu, C. Ding, and D. Tao. Glance and Gaze: Inferring Action-aware Points for One-Stage Human-Object Interaction Detection. In: CVPR (2021) 2, 3, 4, 13
29. B. Kim, T. Choi, J. Kang, and H. Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In: ECCV (2020) 2, 3, 4, 13
30. M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian. Reformulating hoi detection as adaptive set prediction. In: CVPR (2021) 3, 4, 5, 12, 13
31. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan. Microsoft coco: Common objects in context. In: ECCV (2014) 9
32. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In: CVPR (2016) 9
33. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Attention is all you need. In: NeurIPS (2017) 2
34. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. In: The journal of machine learning research (2014) 7, 8
35. and H. Kuhn. The Hungarian method for the assignment problem. In: Naval research logistics quarterly (2020) 3, 7
36. G. Ghiasi, T. Lin, and Q. Le. Dropblock: A regularization method for convolutional networks. In: Wiley Online Library (1955) 8
37. T. Zhou, W. Wang, S. Qi, H. Ling, J. Shen. Cascaded human-object interaction recognition. In: CVPR (2020) 12
38. Pic leaderboard. http://www.picdataset.com/challenge/leaderboard/hoi2019, 2019. 12
39. D. Meng, X. Chen, Z. Fan, G. Zeng, H. Li, Y. Yuan, L. Sun, J. Wang. Conditional DETR for Fast Training Convergence. In: ICCV (2021) 2, 4, 13
40. P. Gao, M. Zheng, X. Wang, J. Dai, H. Li. Fast Convergence of DETR With Spatially Modulated CoAttention. In ICCV (2021) 4
41. X. Dai, Y. Chen, J. Yang, P. Zhang, L. Yuan, L. Zhang. Dynamic DETR: End-to-End Object Detection With Dynamic Attention. In: ICCV (2021) 4
42. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai. Deformable DETR: Deformable Transformers for End- to-End Object Detection. In: ICLR (2020) 4
43. S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In: ICLR (2022) 2, 4
44. H. Yuan, M. Wang, D. Ni, and L. Xu. Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics. In: AAAI (2022) 4, 5
45. Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong. Improving Human-Object Interaction Detection via Phrase Learning and Label Composition. In: AAAI (2022) 4, 5
46. I. Loshchilov, and F. Hutter. Decoupled Weight Decay Regularization. In: ICLR (2018) 9
47. H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR (2019) 8

48. T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar. Focal loss for dense object detection. In: ICCV (2017) 8
49. X. Wang, A. Shrivastava, and A. Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. arXiv preprint arXiv:2201.12329 (2022) 4
50. A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In: CVPR (2017) 4
51. K. Wang, P. Wang, C. Ding, and D. Tao. Batch coherence-driven network for part-aware person re-identification. In: TIP (2021) 1
52. X. Qu, C. Ding, X. Li, X. Zhong, and D. Tao. Distillation Using Oracle Queries for Transformer-Based Human-Object Interaction Detection. In: CVPR (2022) 2, 7
53. X. Lin, C. Ding, J. Zhang, Y. Zhan, and D. Tao. RU-Net: Regularized Unrolling Network for Scene Graph Generation. In: CVPR (2022) 1
54. X. Lin, C. Ding, Y. Zhan, Z. Li, and D. Tao. HL-Net: Heterophily Learning Network for Scene Graph Generation. In: CVPR (2022) 1
55. F. Li, H. Zhang, S. Liu, J. Guo, L. Ni, and L. Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In: CVPR (2022) 7