Region Mining and Refined Query Improved HOI Detection in Transformer

Sixian Chan, *Member, IEEE*, Weixiang Wang, Zhanpeng Shao, *Member, IEEE*, Zheng Wang, and Cong Bai, *Member, IEEE*

***Abstract*—Compared with the object detection task, the Human-Object Interaction (HOI) detection is more complicated. There are interactions between humans and objects, but multiple pairs of interactions may overlap, causing the model to mine features in inappropriate regions and make incorrect predictions. Meanwhile, HOI detection is a multi-task learning task. It requires queries to provide dynamic and correct guided information, thereby better adapting to different images for detecting humans, objects, and interactions. However, the existing query design is fixed and lacks guiding information. To solve this problem, this paper proposes Region Minning and Refined Query (RMRQ). Firstly, the Ground Truth Mask Denoise (GTMD) adopts a denoise training strategy to simulate the overlapping and occlusion problem so that the model can mine more diverse features in the region, avoid mining wrong features, and have the characteristics of denoise to accelerate the training process. Secondly, the Dynamic Linguistic Query (DLQ) dynamically adapts to the input images based on the text information of the HOI triplet to generate queries with guidance information. Meanwhile, Multi-label Focal Loss (MFL) is proposed to constrain the generation of text information to ensure the quality of the query. Finally, extensive experiments on V-COCO and HICO-DET datasets demonstrate the excellent performance of our network and the effectiveness of the above modules.**

***Index Terms*—Human-object interaction detection, deep learning, action and behavior recognition, scene analysis and understanding.**

I. Introduction

HUMAN-OBJECT Interaction (HOI) detection performs multi-task learning [1], [2], which greatly helps for advanced image semantic understanding and human-centered scene understanding [3], [4], [5], [6]. For instance, to ensure the safety of worker behaviors on the shop floor, it is necessary to fully discover human-centered interactions and find as many interactions as possible to prevent accidents. Therefore, HOI detection aims to find as many HOI triplet$\langle human, object, interaction\rangle$ in the image as possible. Therefore, HOI detection aims to find the most HOI triplets $\langle human, object, interaction\rangle$ in the image. To achieve this, it is not only necessary to classify and locate people and objects but also to distinguish the interaction between people and objects.

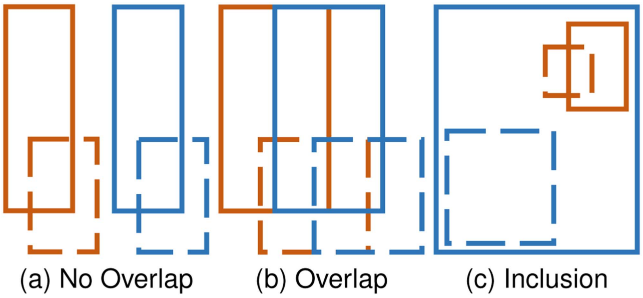


Fig. 1. Three types of regions under HOI detection. The representation of the same color should be recognized as a pair of human and object, and the region with a pair of human and object is called the interaction region. Solid and dashed lines indicate human and object, respectively.

Previous studies [7], [8], [9] have utilized convolutional neural networks (CNNs), but they have limited ability to extract long-distance interactions due to local features, such as $\langle human, kite, fly\rangle$ and $\langle human, ball, throw\rangle$. DETR [10], which is developed based on the attention mechanism, can naturally capture long-distance interactions. Also, DETR's set prediction method for object detection addresses the issue of unequal numbers of predictions and ground truth. These characteristics make DETR suitable for HOI detection, leading to the emergence of HOI detection models based on DETR.

Determining which part of the region is mined by HOI detection significantly influences the interaction prediction. In Fig. 1, currently, there are three main types of relationships regarding regions of interactions: no overlap, overlap, and inclusion. In the non-overlapping case, there is no overlapping of features, so the difficulty lies in the recognition of long interaction pairs. Fortunately, the mainstream model is based on the Transformer, which belongs to the global dependency. Hence, these kinds of cases are recognized very well. However, there are significant problems in recognizing overlapping regions and inclusion relationships.

Manuscript received 19 November 2023; revised 17 January 2024; accepted 16 March 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U20A20196, Grant 61906168 and Grant 61976191, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY23F020023, in part by the Construction of Hubei Provincial Key Laboratory for Intelligent Visual Monitoring of Hydropower Projects under Grant 2022SDSJ01, and in part by Hangzhou AI Major Scientific and Technological Innovation Project under Grant 2022A1ZD0061. *(Corresponding author: Zhanpeng Shao.)*

Sixian Chan is with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China, and also with the Hangzhou Xsuan Technology Company, Ltd., Hangzhou 310023, China.

Weixiang Wang, Zheng Wang, and Cong Bai are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China.

Zhanpeng Shao is with the College of Information Science and Engineering, Hunan Normal University, Changsha 410081, China (e-mail: zpshao@hunnu.edu.cn).

Recommended for acceptance by D. Tao.

Digital Object Identifier 10.1109/TETCI.2024.3392687

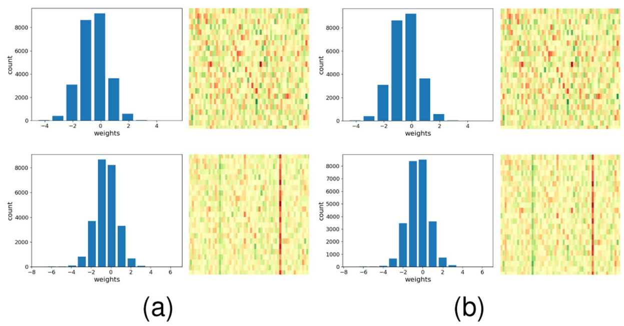


Fig. 2. (a) and (b) are two queries with different inputs. Their first and second rows show the visualizations of the query before and after crossing attention, respectively.

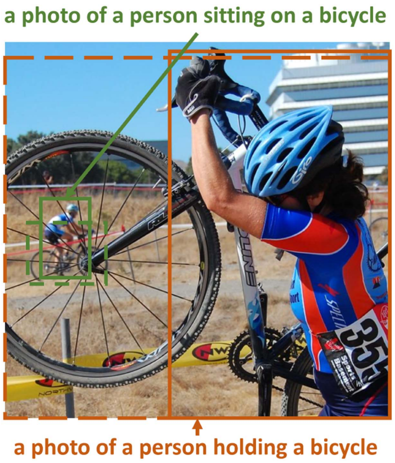

a photo of a person sitting on a bicycle
a photo of a person holding a bicycle

Fig. 3. Illustration of text-guided region selection. The same color indicates that the region and the text are matched. Solid and dashed lines indicate people and objects, respectively.

In Fig. 1(b)–(c), when there is an overlapping relationship with a relatively large overlap or even an inclusion relationship since the overlapped region is too large, the features of the two pairs are almost overlapped, and there is an occlusion problem. In this case, the object on the top will be prioritized to be recognized, matching the correct interaction action with the wrong human and object. Secondly, a model like the QPIC [11] requires a query to detect the three parts of human, object, and interaction at the same time, which is difficult to accomplish with a non-separable query. Though some previous studies [12], [13], [14] discover this problem and adopt separable queries, their query design is still the same as a non-separable query. In Fig. 2, the query of QPIC is not adapted to the input image, and there is no specific guidance information, which makes it a great challenge for HOI detection to perform multi-task learning. We can imagine that it is simple to become progressively more refined by the decoder on rough prediction results, but this is a difficult task on unknown prediction results [15]. In Fig. 3, if the query has pre-textual information $\langle human, object, interaction \rangle$, then the search area can be quickly reduced, with a consideration of only the area that matches the text.

In this regard, this paper proposes the Region Minning and Refined Query (RMRQ) to solve the above problems. Specifically, for the region overlapping problem, the Ground Truth Mask Denoise (GTMD) is proposed. Actually, ground truth is an excellent candidate region sample, but the original denoise strategy is proved invalid by our experiments. We believe that the region is changed because of the coordinate offset, and finally the features can no longer be mined in the correct region. Therefore, this paper adopts the idea of denoise to simulate the masking problem by randomly masking the region. Based on this, the model can solve the region overlapping problem, which results in a mismatching between interactions and pairs of humans and objects, by mining the features of other parts in the region. For the query design problem, inspired by the Grounding task that text can be used to guide the detection and recognition, this paper designs the Dynamic Linguistic Query (DLQ) to make the model adaptively generate text information based on the output of the encoder and CLIP [16] visual-language model. Then, the adaptive query is generated based on the text information to guide attention to the correct region and correctly match the region and interaction. Multi-label Focal Loss (MFL) adopts an overlap area of classification to supervised adaptive text information, which ensures that the generated features are relatively correct. The MFL supervises the generation of adaptive text messages.

Finally, experiments are conducted on V-COCO [17] and HICO-DET [18] datasets to demonstrate the effectiveness of our proposed modules and the superior performance of our model to the baseline and current state-of-the-art models. The contributions of our work are summarized as follows:

- The GTMD is proposed to allow the network to extract more diverse features from different parts of the human and object region to solve the region overlapping problem in the interaction area.
- The DLQ and MFL are proposed to generate dynamic queries with language information and ensure the quality of dynamic queries by adopting a novel idea of overlapping area supervision. They allow the decoder to learn progressively better on refined information.
- Extensive experiments demonstrate the effectiveness of our proposed modules, which also can provide insights with some key parameters to facilitate the further work.

II. Related Work

A. Human-Object Interaction Detection

HOI detection involves object classification and localization, human localization, and interaction classification. The recognition results are represented by HOI triplets in the form of $\langle human, object, interaction \rangle$. According to the model structure, existing HOI detection methods can be divided into two categories: CNN [19]-based methods and Transformer [20]-based methods.

CNN-based methods can be further divided into two-stage and one-stage methods. Two-stage methods often use excellent object detectors [21], [22], [23], [24], [25], [26] to pre-detect all objects in the image. Then, a pairing network is employed to establish the relationship between objects in the image and recognize the relationship. Lin et al. [8] utilized an attention [27] mechanism to construct the relationship between objects and recognize interactions. In-GraphNet [28] and SCG [29] use graph convolution [30], [31], [32] to establish the relationship between objects and recognize interactions. FCL [33] also uses graph convolution, but it generates more samples by using a fabricated sample method to address the issue of long-tailed data [34], [35]. One-stage methods pass the features extracted by the backbone through three streams: human stream, object stream, and interaction stream. After aligning the features of each stream, the detection of humans, objects, and interactions is completed simultaneously. PPDM [9] is a prototype of one-stage methods. In [7], it was discovered that the information mining of the interaction area was not enough based on PPDM, and a deep feature mining method called GGNet was proposed. In [36], a progressive feature extraction approach called RR-Net was proposed to gradually extract better features for HOI detection. DIRV [37] uses repeated humans and objects to achieve interaction classification by voting.

Transformer-based methods can also be divided into methods using non-separable queries and methods using separable queries. A non-separable query refers to using the same query for humans, objects, and interactions. This method requires the query to contain rich information or the features learned by the encoder to be very rich and unbiased. HOITrans [38] and QPIC [11] are prototypes of this method. MHOI [39] improves the features of the encoder from the perspective of multi-scale features, allowing the non-separable query to search for more features. QPIC-ODM [40] employs a memory bank to obtain features from different epochs to address the issue of long-tailed data. PharseHOI [41] generates more samples through text information to make the encoder capable of extracting multi-modal features. Separable query refers to humans, objects, and interactions not sharing the same query. Since the query is not shared, the requirements for the query are not as high as those for non-separable queries. Due to separation, humans, objects, and interactions must be aligned or fused. AS-Net [12] fuses the three. HOTR provides a simplified alignment method for humans, objects, and interactions. QPIC-CPC [42] introduces a training strategy that uses both separable and non-separable queries to learn features with generalization. MURE [14] concatenates three queries to obtain the intermediate query, and different queries from the intermediate query are aligned by cross-attention. However, our method aims to improve performance based on new training strategies and dynamic queries.

B. Denoise and Mine Diverse Feature

Denoise is a training strategy that solves the problems of learning ambiguity and inconsistency caused by the discreteness of bipartite matching, and it also accelerates convergence and improves performance. This strategy assumes that ground truth is a perfect sample that contains much correct information. However, such good samples are not available during inference, and to prevent the network from memorizing the ground truth, it is necessary to add a certain amount of noise to the ground truth. This enables the decoder to suppress the noise and restore the original sample. Moreover, since only a small amount of noise is added, the predicted results of the noisy sample and the ground truth can be directly matched without the need for bipartite matching. This study attempted to add a small offset to the boxes, similar to DN-DETR [43], to realize denoising. However, it was found that this implementation conflicts with HOI detection because the slight offset of the box may no longer correspond to the same ground truth, making it challenging to restore the original sample.

However, the regions of humans and objects are relatively large, implying that there is much potential information in these regions that can aid in recognizing interactions. Therefore, this study proposes adding noise to these regions to enable the network to extract more diverse features beyond the noise. Ground truth guidance for diverse features can enable the network to focus on more meaningful areas and address the position-sensitive issue of HOI. Additionally, this characteristic provides more clues to restore the sample and solve the learning problems caused by bipartite matching.

III. Method

In this section, QPIC is adopted as the baseline, and it is equipped with our RMRQ. The overall network structure is illustrated in Section III-A. Then, Section III-B-III-C introduces the GTMD module and DLQ module in detail, respectively. Finally, the MFL loss and the standard HOI loss are described in Section III-D.

A. Overview of Architecture

Back to QPIC: In Fig. 4, QPIC is divided into the CNN-based backbone, the transformer encoder, the transformer decoder, and the feed-forward network (FFN) for specific tasks. The QPIC first feeds an image $I \in \mathbb{R}^{3\times H\times W}$ to the CNN-based backbone to obtain the patch $P \in \mathbb{R}^{C\times H'\times W'}$. Then, it maps and flattens P to obtain $P' \in \mathbb{R}^{D\times(H'W')}$ and inputs P' into the transformer encoder to obtain memory $E \in \mathbb{R}^{D\times(H'W')}$. Next, the learnable queries $Q \in \mathbb{R}^{N_q\times D}$ and E with its position embedding $E_p \in \mathbb{R}^{D\times(H'W')}$ are fed into the transformer decoder for cross-attention [44], where N_q denotes the number of learnable queries. Finally, Q is fed into the FFN that performs the specific task to obtain the HOI triplet: $\langle human, object, interaction\rangle$. Noteworthy, the transformer decoder is stacked with multiple layers of decoders, and its cross-attention part is shown as follows:

$$Q = Q_c + Q_p \tag{1}$$

$$O_i = \begin{cases} CoAttn\left(O_{i-1}, E + E_p, E\right) & \text{, if } i > 0 \\ CoAttn\left(Q, E + E_p, E\right) & \text{, if } i = 0 \end{cases} \tag{2}$$

where $Q_c, Q_p \in R^{N_q\times D}$ refers to the query content and query position of Q; O_{i-1} denotes the output of the decoder at layer

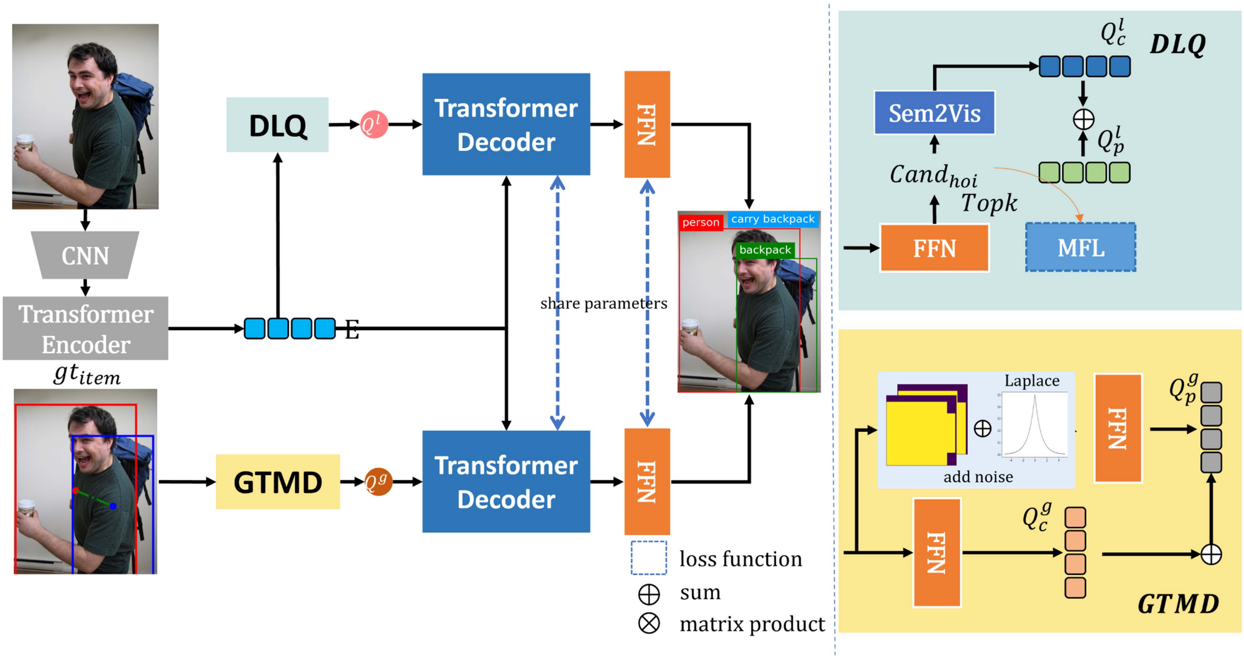


Fig. 4. Overview of RMRQ. During training, an encoder is used in DLQ to make a simple prediction supervised by MFL. Then, Q^l is generated based on the simple prediction. In GTMD, positional noise is added to the ground truth, and then Q^n is generated using gt_{item}. Subsequently, Q^l and Q^n are fed into a decoder with shared parameters for learning, and the HOI triple is predicted using the shared-parameter FFN. During inference, the DLQ branch is kept, while the GTMD is removed.

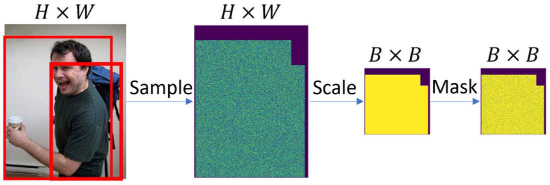
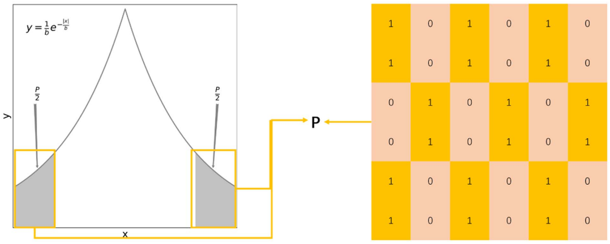


Fig. 5. Illustration of mask generation for GTMD.



1	0	1	0	1	0
1	0	1	0	1	0
0	1	0	1	0	1
0	1	0	1	0	1
1	0	1	0	1	0
1	0	1	0	1	0

Fig. 6. Independent and identically distributed sampling.

$i-1$,; E_p denotes the position embedding of E; $CoAttn(\cdot,\cdot,\cdot)$ denotes cross-attention.

Denoise Sample and Richer Query: Compared to QPIC, RMRQ has two queries: $Q^l \in \mathbb{R}^{N_l \times D}$ and $Q^n \in \mathbb{R}^{N_g \times D}$ during training. Specifically, Q^l and Q^n denote the linguistic query generated by DLQ and the noised query generated by GTMD, respectively. Especially, Q^l is constrained by our proposed MFL. Then, the two queries are fed to the shared parameters of the transformer decoder for feature learning. Finally, RMRQ feeds the learned features into the FFN to obtain two groups of HOI triplets, and each group is subject to the HOI loss. GTMD and DLQ are used throughout the training process. During inference, RMRQ only generates Q^l and one group of HOI triplets.

B. Ground Truth Mask Denoise

Region mining is crucial for HOI detection. Therefore, in Fig. 5, this study proposes masking the object location region: a noised spatial information binary mask is generated using ground truth, and it is used to build the query position.

Regarding the labeled location information, the following preprocessing operation is performed:

$$gt_{item} = [box_i, c_i, cdist, area_i], i \in \{h, o\} \tag{3}$$

where box, c, $cdist$, and $area$ represent the box coordinates, center coordinates, center distance, and area, respectively. h and o refer to humans and objects, respectively.

First, a sampling method is proposed for generating a binary mask of spatial information based on the mask proportion P of 1-valued as follows:

$$\mathcal{M}(x,b) = \frac{1}{2b} exp\left(-\frac{|x|}{b}\right) \tag{4}$$

$$1 - P = 2\int_0^T \mathcal{M}(x,b)dx \tag{5}$$

where T denotes the threshold when the mask proportion is P. The establishment of the relationship between P and T is explained in Fig. 6. In this figure, the value greater than the threshold T is set to 1, and the value less than the threshold T

is set to 0. Secondly, each block is independent and identically distributed and follows $Laplace(0, b)$, so we have the following derivation:

$$\begin{aligned} 1-P &= 2\int_0^T \frac{1}{2b} exp\left(-\frac{|x|}{b}\right) dx \\ \frac{1-P}{2} &= \int_0^T \frac{1}{2b} exp\left(-\frac{x}{b}\right) dx \\ 1-P &= -exp\left(-\frac{T}{b}\right)+1 \\ T &= -b\ln P \end{aligned} \tag{6}$$

Assume that the size of the generated spatial information binary mask is $B \times B$, so the box coordinates need to be first scaled down to the size of the mask, and then the probability matrix of the box region is generated using our proposed sampling method. Then, the regions are binarized according to the T to obtain the region mask $Mask_i$. Next, $Mask_i$ is placed at the corresponding position in the mask map to obtain the noisy spatial information binary mask SMark consisting of human $Mask_h$ and object $Mask_o$. Finally, it is fed into FFN and converted into a query position, as represented by the following equation:

$$\begin{aligned} Mask_i = Thres\left(Sample\left(\mathcal{M}, B, box_i\right), T\right), \\ i \in \{h, o\} \end{aligned} \tag{7}$$

$$SMask = \sum_{i=1}^{i\in\{h,o\}} Fill\left(Zero^B, Mask_i\right) \tag{8}$$

$$Q_c^n = FFN(SMask) \tag{9}$$

where $Sample(\cdot,\cdot,\cdot)$ denotes the probability matrix of the scaled region box. $Thres(\cdot,\cdot)$ means setting the value greater than T to 1 and setting the value less than T to 0. h and o denote the human and object, respectively. $Zero^B$ denotes the zero matrix of size B × B. $Fill(\cdot)$ means putting the binary mask $Mask_i$ into $Zero^B$ according to relative position. Here, the value of B is set to 16. Then, Q_c^n is initialized with gt_{item}, and Q_c^n and Q_p^n are added to obtain Q^n:

$$Q_c^n = FFN(gt_{item}) \tag{10}$$

$$Q^n = Q_c^n + Q_p^n \tag{11}$$

C. Dynamic Linguistic Query

Many excellent language-image models [16], [45] have been proposed recently, which can capture linguistic and contextual information very well. Previous studies [46] have shown that language information can guide image location. Thus, this study generates a dynamic query with linguistic information.

Before training the model, the HOI triplet is formatted so that the Language-Image model (CLIP) can present better linguistic information. First, the HOI triplet is divided into intact, missing interaction, and missing object. In the case of intact, the following conversion is performed: $\langle human, snowboard, lookat\rangle$ →a photo of a person looking at a snowboard. In the case of missing interaction, the following conversion is performed: $\langle human, couch, non\text{-}interaction\rangle$ →a photo of a person and a couch. In the case of a missing object, the following conversion is performed: $\langle human, no\text{-}object, smile\rangle$ →a photo of a person smiling. After formatting, these sentences are extracted by the language-image model to obtain the set of linguistic features S^l in advance. The output memory E is first input into an FFN to obtain the HOI triplet prediction set $pred_{hoi}$, and then the top k in the prediction set is selected as the interaction $Cand_{hoi}$ that appears in the image.

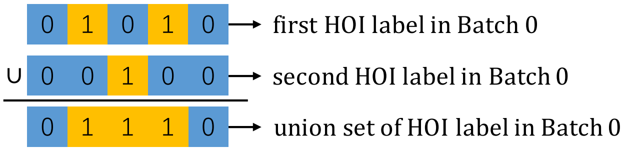

0 1 0 1 0 → first HOI label in Batch 0
∪ 0 0 1 0 0 → second HOI label in Batch 0
0 1 1 1 0 → union set of HOI label in Batch 0

Fig. 7. The generation of $Label_{batch}$.

$$pred_{hoi} = FFN(avgpool(E)) \tag{12}$$

$$Cand_{hoi} = \left\{S_i^l | i \in Topk(pred_{hoi})\right\} \tag{13}$$

where $avgpool(\cdot)$ denotes the average pooling, and $Topk(\cdot)$ denotes taking the HOI triplet with the top k high prediction scores in $pred_{hoi}$. Then converting it into its linguistic features S^l. Then it is projected to the image space to obtain the query content with a priori classification information:

$$Q_c^l = Sem2Vis\left(Cand_{hoi}\right) \tag{14}$$

$$Q^l = Q_c^l + Q_p^l \tag{15}$$

where $Sem2Vis(\cdot)$ represents the project that transforms the linguistic features into the image space, where the simplest FFN is used. The design of Q_p^l is consistent with that of QPIC.

D. Loss Function

Multi-label Focal Loss: This study believes that the higher the accuracy of $Cand_{hoi}$, the better adaptability of the generated Q_c^l. To determine which ground truth corresponds to a supervised predicted interaction, it needs to use bipartite matching. However, as mentioned previously, bipartite matching can have certain negative effects. Therefore, this study proposes supervising whether the predicted interaction appears in the ground truth set. As shown Fig. 7, it is relatively easy to supervise the categories of interactions that appear in the image, so the HOI labels of a batch are merged to obtain $Label_{batch}$.

The union of ground truth labels $Label_{batch}$ and intermediate prediction results $pred_{hoi}$ can be represented by an area overlap calculation. Therefore, L_{MFL} is introduced as follows:

$$\begin{aligned} L_{MFL} = &-\frac{1}{num_{pos}}\left(inds_{pos}\left(1-pred_{hoi}\right)^{\gamma}\log\left(pred_{hoi}\right)\right. \\ &\left.+ inds_{neg}pred_{hoi}^{\gamma}\log\left(1-pred_{hoi}\right)\right) \end{aligned} \tag{16}$$

where num_{pos} denotes the number of positive samples, pos and neg denote the positive sample set and negative sample set, respectively. $inds(\cdot,\cdot)$ denotes the index of the specific sample in the total sample. γ is the hyperparameter, and its value is set to 2.

Overall Loss: We follow the HOI loss of QPIC [11], which we write as L_{hoi}:

$$L_{hoi} = \lambda_b L_b + \lambda_u L_u + \lambda_c L_c + \lambda_a L_a \tag{17}$$

where L_b, L_u, L_c, and L_a denote the box loss, IOU loss, object classification loss, and interaction classification loss, respectively. Specifically, the box loss includes L1 Loss for humans and objects, the IOU loss includes GIOU for humans and objects, and the classification loss includes object and interaction classification loss. λ_b, λ_u, λ_c, and λ_a take the values of 2.5, 1, 1, and 1, respective. Because of the introduction of GTMD and DLQ, there will be L_{hoi}^{DLQ} and L_{hoi}^{GTMD}, and then the total loss L_{total} is obtained based on the combination of L_{MFL}, where the value of α is set to 1:

$$L_{total} = L_{hoi}^{GTMD} + L_{hoi}^{DLQ} + \alpha L_{MLF} \tag{18}$$

IV. Experiment

A. Datasets and Evaluation Metrics

HICO-DET: HICO-DET [18] is the most widely used large-scale HOI detection dataset. It provides 38,118 images for training and 9,658 images for testing. Meanwhile, it contains 150,000 annotated instances. The dataset is annotated with 80 objects, 117 verbs, and 600 HOI categories. The training images are all used for training without dividing the validation set, and the test set is all used for testing. HICO-DET will label the categories and locations of all the humans and objects appearing in an image and then label the categories of interactions based on the pairing of humans and objects. For the evaluation on HICO-DET, only those predicted human and object boxes that have IOUs equal to or greater than 0.5 with the ground truth are considered true positives. Then, interactions in the corresponding image are recognized, and the classification results of objects and interactions are counted. This study adopts mAP (mean Average Precision) for performance evaluation. In the case of a known object setting, the HOI triplets with correctly identified objects are selected first. Full denotes all 600 HOI categories. Rare denotes HOI categories that occur less than 10 times in the training set. Non-Rare denotes HOI categories that occur more than 10 times in the training set. Under the Know Object setting, the HOI triplets with correctly identified objects are selected first. Then, the mAP is calculated based on the frequency of interaction occurrence. In general, we prioritize the three cases under the default setting. Preference is given to full in the default setting.

V-COCO: V-COCO [17] is an HOI detection dataset constructed based on the MS-COCO [47] dataset. It contains 2,533 images for training, 2,867 images for validation, and 4,946 images for testing. It is annotated with 80 objects, 29 verbs (4 are body actions without interactive objects), and 263 HOI categories. The dataset partitioning, use, and labeling are the same as those of HICO-DET. This study takes scenarios 1 and 2 in this dataset for performance evaluation. Specifically, scenario 1 is the most commonly used, and it is often used when objects are occluded. In this scenario, a prediction is considered correct if the predicted interaction is correct and the predicted human box has an IOU value of greater than 0.5 with the ground truth, while the predicted object is empty. Scenario 2 is often used when the detected object does not belong to the COCO category. In this case, a prediction is considered correct if the predicted interaction is correct, and both the predicted human and object boxes have an IOU value of greater than 0.5 with the ground truth.

B. Implementation Details

We using Pytorch framework to build source code. Following previous works, this study takes ResNet-50 [19] as our backbone. According to the QPIC setting, our network is initialized with a pre-trained DETR. AdamW is taken as our optimizer, and all experiments are performed on three RTX3090TI graphic cards with a batch size of 12. A total of 80 training epochs are used. Besides, the initial learning rate is set to $1e-4$, and it decays to $1e-5$ at the $50th$ epoch. Additionally, the values of N_l and D are set to 100 and 256, respectively, and the value of B is set to 64. During the training process, our proposed GTMD training strategy, DLQ, and the MFL that supervises DLQ are used. Therefore, there are two types of queries for the decoder to learn, and accordingly, two sets of HOI triplet predictions are obtained. During the inference process, as GTMD or loss calculation is not used, only one query is generated by DLQ, and thus only one set of HOI triplets is obtained as the final prediction. Also, the preparation of the dataset is the same as QPIC.

C. Comparisons With State-of-The-Art

Comparisons on HICO-DET: The one-stage method is not as accurate as the two-stage method, but our method is close to the two-stage methods such as UPT and SQA. As listed in Table I, our approach achieves better performance than the state-of-the-art. It is noteworthy that when our proposed modules are added to QPIC-R50 separately, the performance is increased by 2.04, 3.31, and 1.65 in the full, rare, and non-rare cases, respectively. Meanwhile, compared to QPIC-R101, our model not only takes less time to train but also performs better. Compared to the same method using linguistic information, PhraseHOI, our model has 1.82, 3.13, and 1.42 higher performance in these three cases, respectively. Besides, compared with the CPC of the consistent learning approach, the performance improvement of our model is 0.78, 1.62, and 0.53, respectively. Compared to SQAB, our model is higher by 0.29, 0.25, and 0.3, respectively.

Comparisons on V-COCO: In Table I, our model has an outstanding performance compared to the state-of-the-art approach. After our proposed modules are added to QPIC-R50 separately, the performance is improved by 2.25 and 3.18, and 2.27 and 2.58 on scenarios 1 and 2, respectively. In scenario 1, our model outperforms PhraseHOI by 3.36, and it outperforms MHOI by 2.35.

On the HICO-DET and V-COCO datasets, our model achieves significant performance improvements compared to QPIC and existing excellent training strategies such as CPC and feature enhancement methods like MHOI.

TABLE I
COMPARISON ON HICO-DET AND V-COCO

Type	Methods	HICO-DET						V-COCO	
		Full(D)	Rare(D)	Non-Rare(D)	Full(K)	Rare(K)	Non-Rare(K)	Scenario 1	Scenario 2
CNN-based	Lin et. al [8]	16.63	11.30	18.22	19.22	14.56	20.61	48.10	/
	In-GraphNet [28]	17.72	12.93	19.31	/	/	/	48.90	/
	MLCNet [48]	17.95	16.62	18.35	/	/	/	55.2	/
	ACP [49]	20.59	15.92	21.98	/	/	/	52.98	/
	PD-Net [50]	20.81	15.90	22.28	24.78	18.88	26.54	52.60	/
	RR-Net [36]	20.96	13.43	23.21	/	/	/	55.30	/
	DIRV [37]	21.78	16.38	23.39	25.52	20.84	26.92	56.10	/
	PPDM [9]	21.94	13.97	24.32	24.81	17.09	27.12	/	/
	GGNet [7]	23.47	16.48	25.60	27.36	20.23	29.48	54.70	/
	DSSF [51]	25.23	18.72	27.17	28.53	21.68	30.57	57.64	/
	OC-Immunity [52]	25.44	23.03	26.16	/	/	/	/	/
	SCG [29]	29.26	24.61	30.65	/	/	/	54.20	60.90
	FCL [33]	29.12	23.67	30.75	31.31	25.62	33.02	52.35	/
Transformer-based	UPT(Two-stage) [53]	31.66	25.94	33.36	35.05	29.27	36.77	59.00	64.50
	SQA(Two-stage) [54]	31.99	29.88	32.62	35.12	32.74	35.84	65.42	70.56
	HOI-Trans [38]	26.61	19.15	28.84	29.13	20.98	31.57	52.90	/
	HOTR [13]	25.10	17.34	27.42	/	/	/	55.20	/
	AS-Net [12]	28.87	24.25	30.25	31.74	27.07	33.14	53.90	/
	QPIC-R50 [11]	29.07	21.85	31.23	31.41	24.00	33.63	58.80	60.10
	QPIC-R101 [11]	29.90	23.92	31.69	32.32	26.21	34.15	58.30	60.70
	MHOI [39]	29.67	24.37	31.25	31.87	27.28	33.24	58.70	64.50
	QPIC-ODM [40]	29.26	22.07	31.41	/	/	/	/	/
	PharseHOI [41]	29.29	22.03	31.46	31.97	23.99	34.36	57.40	/
	QPIC-CPC [42]	29.63	23.14	31.57	/	/	/	/	/
	SQAB [55]	30.82	24.92	32.58	33.58	27.19	35.49	/	/
	CDT [56]	30.48	**25.48**	32.37	/	/	/	61.43	**65.37**
	SSRT [57]	30.36	24.31	31.83	/	/	/	/	/
	Ours	**31.11**	25.16	**32.88**	**33.89**	**27.78**	**35.72**	61.05	63.28

Bold denotes the best results. Underline denotes the secondary results. D and K represent default and known objects, respectively. The higher the value, the better.

TABLE II
ABLATION STUDIES WITH GTMD, DLQ AND MFL IN HICO-DET AND V-COCO

GTMD	DLQ	MFL	V-COCO		HICO-DET		
			Scenario 1	Scenario 2	Full(D)	Rare(D)	Non-Rare(D)
			58.80	60.10	29.07	21.85	31.23
	✓		59.68	62.01	29.38	23.35	31.18
✓			59.87	62.00	29.56	23.54	31.35
	✓	✓	60.40	62.63	29.51	23.53	31.30
✓	✓		59.75	61.86	30.10	24.05	31.91
✓	✓	✓	**61.05**	**63.28**	**31.11**	**25.16**	**32.88**

D. Ablation Studies

The influence of GTMD: In Table II, relative to the baseline, the use of GTMD leads to a performance improvement of 1.07 and 1.90 on V-COCO, and 0.31, 1.5, and 0.12 on HICO-DET, respectively. It shows that the GTMD module has performance improvements in all three categories of HICO-DET, with the most significant performance improvement in the rare category. This indicates that, with the assistance of our training strategy, the GTMD can explore more diverse features to better identify the type of interaction, especially for rare interactions. Meanwhile, it proves that the proposed GTMD module is suitable for HOI detection.

The influence of DLQ and MFL: In Table II, the DLQ has a specific search region and effective appearance cues, which enables the decoder to fine-tune quickly to find objects and recognize interactions. Relative to the baseline, the DLQ improves performance by 0.88 and 1.91 on V-COCO and 0.31 and 1.5 in full and rare settings on HICO-DET, respectively. This indicates that when the query dynamically adapts to the input and provides more information, it can better assist in multi-task learning such as HOI detection. Meanwhile, DLQ is more adaptive with the introduction of MFL. Compared to the baseline, DLQ+MFL achieves a performance improvement of 1.6 and 2.53 on V-COCO and 0.44, 1.68, and 0.07 on HICO-DET, respectively. This also confirms that MFL can effectively supervise the query generated by DLQ, enabling the query to better adapt to the input and provide correct and suitable information for different inputs. Additionally, both DLQ and the combination of DLQ and MFL can achieve significant performance improvements in the rare category of HICO-DET, demonstrating that the dynamic

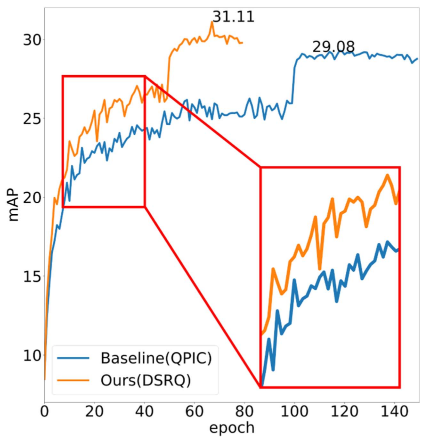


Fig. 8. Convergence in HICO-DET.

TABLE III
THE WAY TO ADD NOISE

Denoise Method	scenario 1	scenario 2
none	59.78	61.94
coord offset	46.07	48.28
mask scale	60.05	62.05
mask with P	**61.05**	**63.28**

query based on language information can alleviate the sample imbalance problem.

When GTMD, DLQ, and MFL are all added to the baseline, the best performance is achieved. This leads to a performance improvement of 2.25 and 3.18 on V-COCO, and a performance improvement of 2.04, 3.31, and 1.65 on HICO-DET. This proves that the proposed modules benefit each other mutually. DLQ and MFL provide adaptive input queries with rich information, allowing the decoder to learn better. Meanwhile, with the assistance of the GTMD training strategy, high-quality samples are generated based on the ground truth, enabling the decoder to not only mine more diverse features but also avoid the problems inherent in binary matching. Therefore, DLQ and MFL jointly improve the baseline's generalization ability from the query and feature mining perspectives, respectively. Since the calculation in scenarios 1 and 2 of V-COCO needs to be remeasured according to the official code, it is impossible to show the mAP of each epoch, so this study chooses to visualize the change of mAP on HICO-DET. As shown in Fig. 8, compared with QPIC, RMRQ requires fewer training iterations with a more significant training curve, and it achieves higher accuracy. This indicates that RMRQ not only improves accuracy but also accelerates convergence.

E. Exploring With Variants of GTMD, DLQ and MFL

Exploring with variants of GTMD: This study investigates how to add noise to the ground truth to improve HOI detection. In Table III , four denoise methods in HOI detection are compared. The none indicates a denoise process without adding any noise, the coord offset indicates a denoise process that randomly offsets the coordinates of the boxes of humans and objects, and the

TABLE IV
THE MASK PROPORTION CHOICES

P	scenario 1	scenario 2
0.0%	59.78	61.94
1.0%	59.58	62.02
5.0%	60.5	62.56
8.0%	**61.05**	**63.28**
9.0%	60.09	62.1
10.0%	60.06	62.1
20.0%	60.15	62.28
50.0%	59.73	61.95
70.0%	59.66	61.72
100.0%	59.72	61.85

TABLE V
THE IMPACT OF INFORMATION RICHNESS ON THE MODEL PERFORMANCE

Query Partition	scenario 1	scenario 2
100/0	**61.05**	**63.28**
75/25	60.11	62.16
50/50	60.05	62.2
25/75	60.06	62.25
0/100	59.87	62.00

The total number of queries is 100. The query partition represents the ratio of the number of queries generated by DLQ to the number of original queries. The 100/0 means 100 queries produced by DLQ and 0 original query.

TABLE VI
THE CHOICE OF LOSS AND THE COMBINATION APPROACH

DLQ Supervision	scenario 1	scenario 2
Focal [58](match)	60.64	62.93
Focal(match)+Dice	60.24	62.46
Bce	60.00	62.64
Dice [59]	60.42	62.55
MFL+Dice	60.54	62.59
MFL	**61.05**	**63.28**

The match means pair prediction with ground truth using bipartite matching.

TABLE VII
THE NUMBER OF CANDIDATE HOI SETS SELECTED

topk	scenario 1	scenario 2
0	59.51	61.49
1	59.4	61.59
2	60.16	62.25
4	**61.05**	**63.28**
8	59.91	62.12
12	60.3	62.5

TABLE VIII
PERFORMANCE COMPARISON

Method	Params(M)	FPS	Train HICO(H)	EPOCH	Full(D)
AS-Net	52.75	10	134	100	28.87
HOTR	51.41	/	/	/	25.10
HOI-Trans	60.62	24	187	250	26.61
QPIC-R50	**41.46**	**37**	231	150	29.07
QPIC-R101	60.38	32	251	150	29.90
MHOI	86.5	/	/	130	29.76
QPIC-CPC	42.47	31	135	90	29.63
SQAB	101.77	/	/	**25**	30.82
Ours	85.46	21	**64**	80	**31.11**

FPS refers to how many images can be processed in one second. Train HICO refers to the number of hours it takes to train the model using HICO-DET on our device. Epoch refers to rounds of this model needs to be trained. And full(D) refers to the value of mAP for the full category under the default setting under HICO-DET.

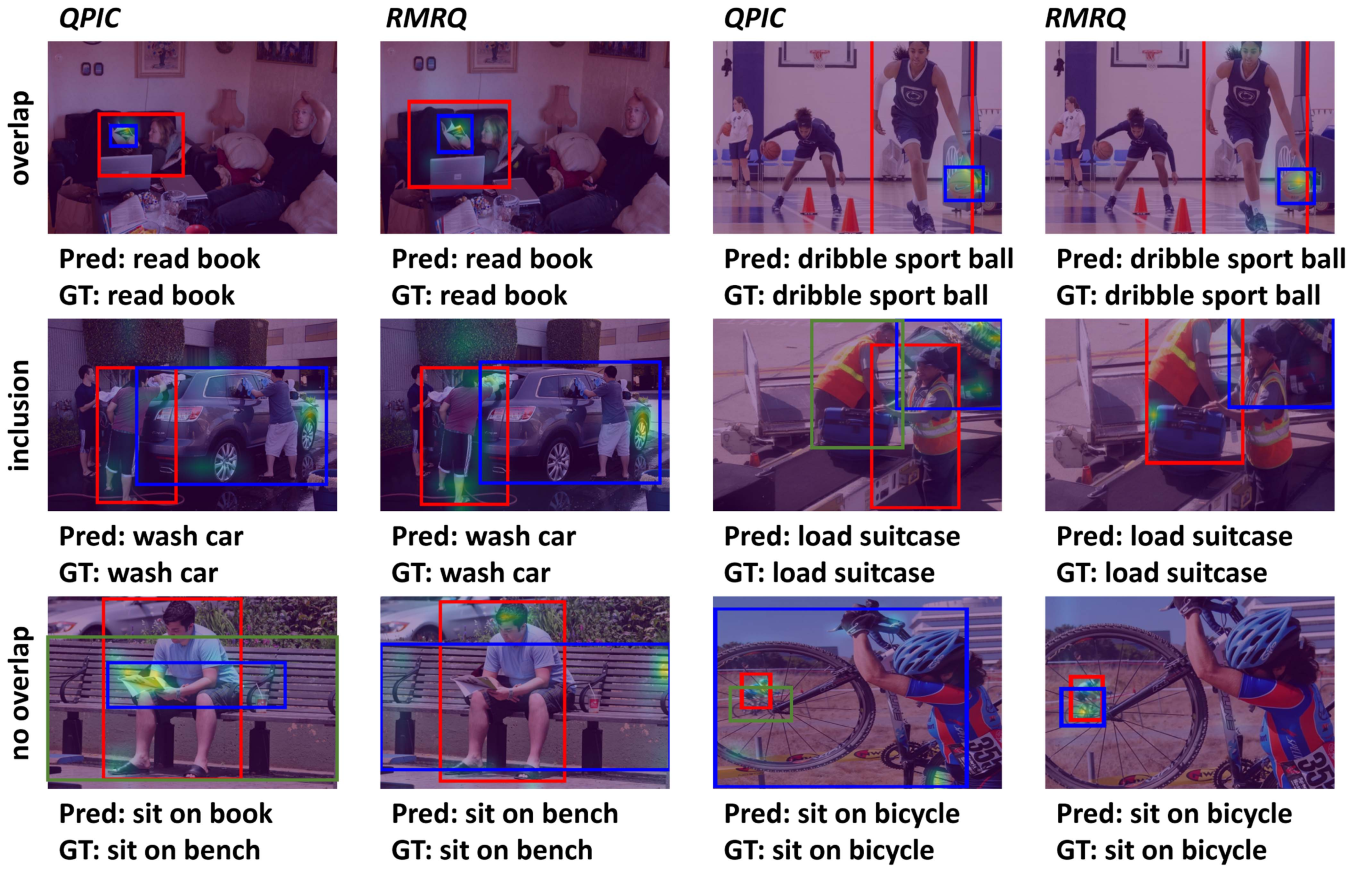


Fig. 9. The visualization results of HOI detection and the heatmap of cross-attention. The text below the image indicates the interaction corresponding to the image and whether the QPIC and RMRQ obtain correct recognization results, respectively, where ✓ indicates correct, and × indicates incorrect. More bright areas indicate richer features and brighter areas indicate more salient features.

mask scale indicates randomly scaling up the masks of the regions of humans and objects. The mask with p represents the denoise process adopted by our GTMD. The coord offset only randomly offsets the box by 0-10 pixels, and the mask scale only enlarges the box by 0-10 pixels. The effect of not adding noise to ground truth (none) is tested, and relative to the baseline, it leads to a performance improvement of 0.98 and 1.84. That indicates that ground truth can help network mine features. However, the coord offset is even lower than that of the baseline and the none approach, indicating that the position of humans and objects is very sensitive to interaction identification. This makes it difficult for the network to restore the original samples based on the noisy samples and prevents the network from correctly mining features. This leads to the failure of the denoising idea and worse learning ability of the network. When a mask scale is used, compared with the none approach, the accuracy is improved by 0.27 and 0.11, respectively. This indicates that enlarging the region without reducing the original feature exploration area allows the network to have a larger field of view to identify interactions. Experimentally, it is found that the third method outperforms the previous, contributing to a performance improvement by 14.98, 15, and 1, 1.23, respectively. Also, the third method still improves performance by 1.27 and 1.34 compared to the none approach. This demonstrates that the mask with the P method allows the network to truly extract diverse features from different parts of the region by masking the regions. Therefore, the network has more clues to complete the HOI detection task. From Table IV, it can be found that as P increases, the performance generally first increases and then decreases. When the value of P is set to 0.08, the best performance is obtained. This indicates that if the value of P is too small, the network will still focus on almost the same area, and if the value of P is set too large, it will cause too few areas to be attended to, which is approximately equivalent to no masking. Therefore, only when the value of P is set appropriately, can the network focus on multiple parts as much as possible and mine diverse features.

Exploring with variants of DLQ and MFL: In Table V, the impact of information richness on the model performance is investigated. The query partition represents the ratio of the number of queries generated by DLQ to the number of original queries. It can be seen that the higher the proportion of DLQ queries to the total number of queries, the better the effect. This demonstrates that given richer query information, the decoder can more easily find relevant features to complete the HOI detection task. To produce dynamic queries with detailed information, this study investigates various loss functions in the research of the segmentation field. In Table VI, three types of losses are used, and considering the way to use the focal loss, there are six combinations in total. It is found that MFL works

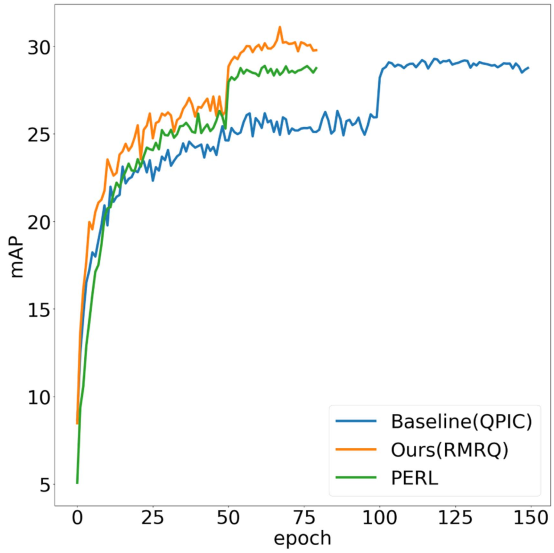


Fig. 10. The query position exploration experiment. PERL denotes the query position of DLQ in RMRQ using the positional features of PERL instead. This experiment was performed on the HICO-DET.

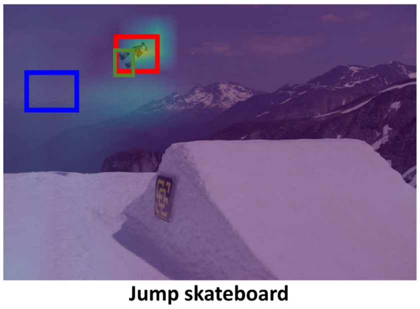


Fig. 11. The flaw in our model. The red and blue boxes are the human and object predicted by the model, respectively. The green box is the correct object.

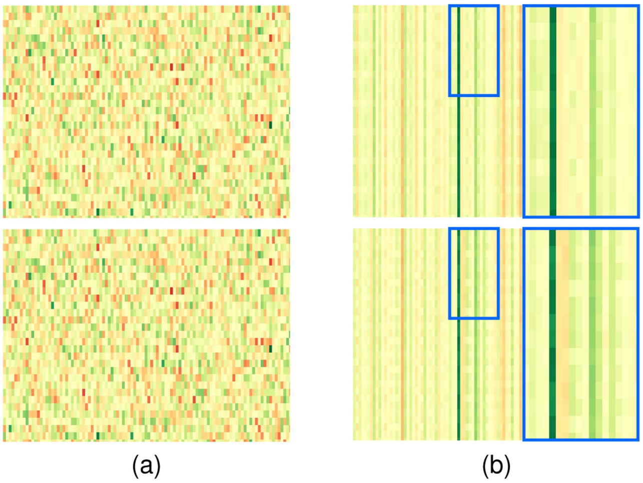


Fig. 12. The visualization of the query. (a) and (b) mean query of QPIC and RMRQ. The same rows of (a) and (b) denote the queries for the same input.

TABLE IX
INFLUENCE OF α

alpha	scenario 1	scenario 2
0.25	60.05	62.23
0.5	60.38	62.57
0.75	60.71	62.89
1	61.05	63.28

best. Meanwhile, this study investigates how many predictions should be selected as the candidate set for generating dynamic queries because there are not many types of interactions in an image. As listed in Table VII, the best results are obtained when the value of K is chosen as 4. Also, different ratios of MFL and HOI loss are studied, and considering that MFL and HOI loss have similar loss values, the upper limit of the ratio is set to 1:1. Table IX shows that the best performance is achieved when the ratio of MFL and HOI loss is 1:1.

F. Performance Results

In Table VIII, among the comparative models we listed, first of all our model has the highest accuracy with only 80 training rounds. Although our number of parameters is in the mid-ranking, the FPS is still 21 in the evaluated state and the training time spent is the least in the reproducible condition. Overall, our model still achieves tradeoff.

G. Visualization Results

In Fig. 9, the attention of QPIC and RMRQ and the human and object boxes are visualized. Based on the three overlapping region divisions, it can be found that in the case of multiple pairs of interactions with no overlap, both the baseline and our model perform very well, but our attention map is better, indicating that our model can pay attention to more features to infer the interactions, humans, and objects. When there are inclusion relations, the baseline makes a very serious mistake, e.g., when the sit-on action occurs in the center of the picture and the book is also in the center, the baseline thinks that the sit-on action occurs on the book. Also, in the sample of sitting on a bicycle, since the larger bicycle covers the small bicycle, the baseline mistakenly thinks that the sit on action occurs on the larger bicycle, but our model still recognizes the human and object and makes a correct match. In Fig. 12, attributed to the guidance of random masking regions and linguistic information, our model can find the correct human and object and mine diverse features in the regions of this pair to avoid the above problems.

H. Limitation and Future

Currently, the query position in our DLQ branch is still self-learning and does not like query position of GTMD that has better guided-information. Also, we tried PERL [60] to generate query postion of DLQ. However, in Fig. 10, we found the results are worse relative to the original. It is even inferior to QPIC in the later stages of training. We suspect that the PERL targets temporally related position features, while the images in our task are unrelated to each other, leading to problems with the position features generated by PERL. In the future, we need to explore a query position that can generate prior information under the DLQ branch and do better alignment with GTMD. Also, in Fig. 11, because the original region is small, GTMD results in no features to be mined after masking so much regions. That leads to some recognition errors. This problem need to be solved in future. Finally, GTMD currently masks areas including the background, but the background should always be masked, and then areas of human and objects are masked randomly. In the future, we will refer to BPDA [61] to further improve the masking approach.

V. Conclusion

This paper proposes a novel method for HOI detection, where three strategies including GTMD, DLQ, and MFL are formulated. GTMD mines features from different parts of the region to avoid problems caused by overlapping interactions; also, it explores a denoise method suitable for HOI detection, which reduces the number of training epochs and improves the model's generalization performance. Meanwhile, DLQ is introduced to dynamically adapt the input and generate specific queries to replace the original fixed queries, thereby providing guidance information. Besides, MFL is proposed to supervise the generation of DLQ by adopting the idea of overlapping areas to ensure the quality of DLQ generation. Abundant ablation experiments have demonstrated the effectiveness of the proposed modules. Also, comparative experiments have shown that the proposed model performs well on the V-COCO and HICO-DET datasets. Currently, the query position in our DLQ branch is still self-learning and does not align well with the query position in GTMD. In the future, we need to explore a query position that can generate prior information under the DLQ branch and do better alignment with GTMD. Also, there is a risk that objects that are too small in the image may fail to extract any features due to GTMD, and this will need to be addressed in the future as well.

References

[1] Z. Liu, T. Li, Y. Chen, K. Wei, M. Xu, and H. Qi, "Deep multi-task learning based fast intra-mode decision for versatile video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 6101–6116, Oct. 2023.

[2] J. Li, Y. Tian, T. Huang, and W. Gao, "Multi-task rank learning for visual saliency estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 623–636, May 2011.

[3] J. Shao, C. C. Loy, K. Kang, and X. Wang, "Crowded scene understanding by deeply learned volumetric slices," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 613–623, Mar. 2017.

[4] L. Brun, A. Saggese, and M. Vento, "Dynamic scene understanding for behavior analysis based on string kernels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1669–1681, Oct. 2014.

[5] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "RGB-T semantic segmentation with location, activation, and sharpening," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1223–1235, Mar. 2023.

[6] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. Van Gool, "stagNet: An attentive semantic RNN for group activity and individual action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 549–565, Feb. 2020.

[7] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13234–13243.

[8] X. Lin, Q. Zou, and X. Xu, "Action-guided attention mining and relation reasoning network for human-object interaction detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1104–1110.

[9] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 479–487.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.,* 2020, pp. 213–229.

[11] M. Tamura, H. Ohashi, and T. Yoshinaga, "QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10410–10419.

[12] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating HOI detection as adaptive set prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9004–9013.

[13] B. Kim, J. Lee, J. Kang, E. S. Kim, and H. J. Kim, "HOTR: End-to-end human-object interaction detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 74–83.

[14] S. Kim, D. Jung, and M. Cho, "Relational context learning for human-object interaction detection," in *Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognit.*, 2023, pp. 2925–2934, doi: 10.1109/CVPR52729.2023.00286.

[15] S. Li et al., "Logical relation inference and multiview information interaction for domain adaptation person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 12, 2023, doi: 10.1109/TNNLS.2023.3281504.

[16] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[17] S. Gupta and J. Malik, "Visual semantic role labeling," 2015, *arXiv:1505.04474*.

[18] Y. -W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 381–389.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016 pp. 770–778.

[20] A. Dosovitskiy et al., "An image is worth 16 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Representations*, 2021.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[22] C.-H. Chuang, J.-W. Hsieh, L.-W. Tsai, S.-Y. Chen, and K.-C. Fan, "Carried object detection using ratio histogram and its application to suspicious event analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 6, pp. 911–916, Jun. 2009.

[23] J. Liu, Y. Huang, J. Peng, J. Yao, and L. Wang, "Fast object detection at constrained energy," *IEEE Trans. Emerg. Topics Comput.*, vol. 6, no. 3, pp. 409–416, Jul.-Sep. 2018.

[24] A. Pramanik, S. K. Pal, J. Maiti, and P. Mitra, "Granulated RCNN and multi-class deep SORT for multi-object detection and tracking," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 6, no. 1, pp. 171–181, Feb. 2022.

[25] G. T. Tchendjou and E. Simeu, "Detection, location and concealment of defective pixels in image sensors," *IEEE Trans. Emerg. Topics Comput.*, vol. 9, no. 2, pp. 664–679, Apr.-Jun. 2021.

[26] J. Leng, Y. Liu, X. Gao, and Z. Wang, "CRNet: Context-guided reasoning network for detecting hard objects," *IEEE Trans. Multimedia*, vol. 26, pp. 3765–3777, 2024.

[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[28] D. Yang and Y. Zou, "A graph-based interactive reasoning for human-object interaction detection," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 1111–1117.

[29] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13299–13307, doi: 10.1109/ICCV48922.2021.01307.

[30] S. Miao, Y. Hou, Z. Gao, M. Xu, and W. Li, "A central difference graph convolutional operator for skeleton-based action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4893–4899, Jul. 2022.

[31] Y. Tang, Y. Wei, X. Yu, J. Lu, and J. Zhou, "Graph interaction networks for relation transfer in human activity videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 9, pp. 2872–2886, Sep. 2020.

[32] Z. Tang, H. Wang, X. Yi, Y. Zhang, S. Kwong, and C.-C. J. Kuo, "Joint graph attention and asymmetric convolutional neural network for deep image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 421–433, Jan. 2023.

[33] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Detecting human-object interaction via fabricated compositional learning," in *Proc. IEEE /CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14646–14655.

[34] W. Wang, Z. Zhao, P. Wang, F. Su, and H. Meng, "Attentive feature augmentation for long-tailed visual recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5803–5816, Sep. 2022.

[35] Y. Hu, J. Gao, and C. Xu, "Learning multi-expert distribution calibration for long-tailed video classification," *IEEE Trans. Multimedia*, vol. 26, pp. 555–567, 2024.

[36] D. Yang, Y. Zou, C. Zhang, M. Cao, and J. Chen, "RR-Net: Relation reasoning for end-to-end human-object interaction detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3853–3865, Jun. 2022.

[37] H. Fang, Y. Xie, D. Shao, and C. Lu, "DIRV: Dense interaction region voting for end-to-end human-object interaction detection," in *Proc. 35h AAAI Conf. Artif. Intell., 33rd Conf. Innov. Appl. Artif. Intell., 11th Symp. Educ. Adv. Artif. Intell.*, 2021, pp. 1291–1299.

[38] C. Zou et al., "End-to-end human object interaction detection with HOI transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11825–11834.

[39] Y. Cheng, Z. Wang, W. Zhan, and H. Duan, "Multi-scale human-object interaction detector," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1827–1838, Apr. 2023.

[40] G. Wang, Y. Guo, Y. Wong, and M. S. Kankanhalli, "Chairs can be stood on: Overcoming object bias in human-object interaction detection," in *Proc. Eur. Conf. Comput, Vis.* 2022, pp. 654–672.

[41] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," in *Proc. 36th AAAI Conf. Artif. Intelligence, AAAI 2022, Thirty-Fourth Conf. Innov. Appl. Artif. Intell., Twelveth Symp. Educ. Adv. Artif. Intell.*, 2022, pp. 1509–1517.

[42] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1009–1018.

[43] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 13619–13627.

[44] H. Lin, X. Cheng, X. Wu, and D. Shen, "CAT: Cross attention in vision transformer," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859720.

[45] W. Kim, B. Son, and I. Kim, "VILT: Vision-and-language transformer without convolution or region supervision," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[46] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Visual grounding with transformers," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859880.

[47] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[48] X. Sun, X. Hu, T. Ren, and G. Wu, "Human object interaction detection via multi-level conditioned network," in *Proc. Int. Conf. Multimedia Retrieval*, 2020, pp. 26–34, doi: 10.1145/3372278.3390671.

[49] D. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 718–736, doi: 10.1007/978-3-030-58589-1_43.

[50] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis.,.*, 2020, pp. 69–85, doi: 10.1007/978-3-030-58565-5_5.

[51] D. Gu, S. Ma, and S. Cai, "DSSF: Dynamic semantic sampling and fusion for one-stage human–object interaction detection," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5012913.

[52] X. Liu, Y. Li, and C. Lu, "Highlighting object category immunity for the generalization of human-object interaction detection," in *Proc. 36th AAAI Conf. Artif. Intell., 34th Conf. Innov. Appl. Artif. Intell., 12th Symp. Educ. Adv. Artif. Intell.*, 2022, pp. 1819–1827.

[53] F. Z. Zhang, D. Campbell, and S. Gould, "Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer," in *Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognit.*, 2022, pp. 20072–20080, doi: 10.1109/CVPR52688.2022.01947.

[54] F. Zhang, L. Sheng, B. Guo, R. Chen, and J. Chen, "SQA: Strong guidance query with self-selected attention for human-object interaction detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096029.

[55] J. Li, H. Lai, G. Gao, J. Ma, H. Quan, and D. Chen, "SQAB: Specific query anchor boxes for human–object interaction detection," *J. Displays*, vol. 80, 2023, Art. no. 102570.

[56] D. Zong and S. Sun, "Zero-shot human–object interaction detection via similarity propagation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 06, 2023, doi: 10.1109/TNNLS.2023.3309104.

[57] A. S. M. Iftekhar, 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 *Proc. IEEE/CVF Conf. Computer Vis. Pattern Recognit.*, 2022, pp. 5343–5353, doi: 10.1109/CVPR52688.2022.00528.

[58] T.-Y. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

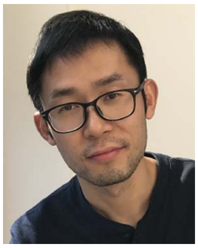
[59] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[60] Y. Tu, L. Li, L. Su, J. Du, K. Lu, and Q. Huang, "Viewpoint-adaptive representation disentanglement network for change captioning," *IEEE Trans. Image Process.*, vol. 32, pp. 2620–2635, 2023.

[61] Y. Wang, G. Qi, S. Li, Y. Chai, and H. Li, "Body part-level domain alignment for domain-adaptive person re-identification with transformer framework," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 3321–3334, 2022.

Sixian Chan (Member, IEEE) was born in Anqing, China, in 1990. He received the Ph.D. degree from the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China, in 2018. He is currently a Lecturer of computer science and technology with the Zhejiang University of Technology. He is the author of more than 40 articles. His research interests include image processing, machine learning, object detection, and video tracking.

Weixiang Wang was born in Wenzhou, China, in 1999. He received the bachelor's degree in software engineering from the Zhejiang College, Zhejiang University of Technology, Shaoxing, China, in 2020. He is currently working toward the master's degree with the School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. His research interests include computer vision and deep learning.



Zhanpeng Shao (Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering from the Xi'an University of Technology, Xi'an, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer vision from the City University of Hong Kong, Hong Kong, in 2015. From 2015 to 2016, he was a Senior Research Associate with Shenzhen Research Institute, City University of Hong Kong. From 2018 to 2019, he was a Postdoctoral Research Fellow with the Department of Computing Science, University of Alberta, Edmonton, AB, Canada. From 2016 to 2022, he was an Associate Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. He is currently an Associate Professor with the College of Information Science and Engineering, Hunan Normal University, Changsha, China. His research interests include computer vision, pattern recognition, machine learning, and robot sensing. He was the recipient of the Best Conference Paper Award on ICMA2014 and ICMA2016.

Zheng Wang received the B.S. degree in computer science from the Zhejiang University of Technology, Hangzhou, China, in 2017, and the Ph.D. degree from the School of Computer Science, Fudan University, Shanghai, China, in 2022. He is currently a Lecturer with the Zhejiang University of Technology. He has authored or coauthored eight papers and two authorized invention patents. His research interests include vision and language tasks, such as cross-modal video retrieval, multimedia content analysis, long-tail content recognition, and anomaly detection. He was the reviewer for several conferences, such as CVPR, ACL, ACMMM, AAAI, ICME, and journals, such as TOMM, *Neurocomputing*, and MVAP.

Cong Bai (Member, IEEE) received the B.E. degree from Shandong University, Jinan, China, in 2003, the M.E. degree from Shanghai University, Shanghai, China, in 2009, and the Ph.D. degree from the National Institute of Applied Sciences, Rennes, France, in 2013. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China. His research interests include computer vision and multimedia processing.