DCT: Divide-and-Conquer Transformer Network with Knowledge Transfer for Query-driven HOI Detection

Changchang Sun
csun39@hawk.iit.edu
Illinois Institute of Technology
Chicago, IL, USA

Bin Duan
bduan2@hawk.iit.edu
Illinois Institute of Technology
Chicago, IL, USA

Hugo Latapie
hlatapie@cisco.com
Cisco Research
San Jose, CA, USA

Gaowen Liu
gaoliu@cisco.com
Cisco Research
San Jose, CA, USA

Yan Yan
yyan34@iit.edu
Illinois Institute of Technology
Chicago, IL, USA

ABSTRACT

Query-driven human-object interaction (HOI) detection methods have gained increasing attention benefiting from the advances of transformer architectures. Existing methods either perform the human-object association and interaction understanding simultaneously based on a single decoder or two cascade decoders. However, it is nontrivial to characterize task-aware distinct information and make an appropriate trade-off on multi-task learning using a shared transformer decoder. Following real-world decision-making process, both spatial and semantic information of the human-object pair are equally important to enhance the interaction understanding, and the prior knowledge stream can be passed between the object category and action type prediction branches of the HOI detector. Therefore, in this paper, we propose a novel divide-and-conquer transformer network for query-driven HOI detection with knowledge transfer, namely DCT, which consists of two-level decoders. In particular, for the first level, we divide the human-object association task by designing task-aware decoders and acquiring three independent query sets. Thereafter, to guide the interaction detection compatible with the human-object association, a spatial-semantic fusion module is introduced for the initialization of the second-level interaction decoder. Besides, we design a multi-task knowledge transfer module to further pass auxiliary information among different tasks. As a result, DCT outperforms the state-of-the-art frameworks on two HOI benchmark datasets.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Human-object Interaction Detection; Divide-and-Conquer; Query-driven Transformer

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06
https://doi.org/10.1145/3652583.3658028

ACM Reference Format:
Changchang Sun, Bin Duan, Hugo Latapie, Gaowen Liu, and Yan Yan. 2024. DCT: Divide-and-Conquer Transformer Network with Knowledge Transfer for Query-driven HOI Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652583.3658028

1 INTRODUCTION

Recently, due to the unprecedented interest in deeper scene understanding, human-object interaction (HOI) detection task in a static image has become a promising topic. It aims to accurately locate the human-object pairs within an image and understand their corresponding interaction, typically abstracted as a set of quadruples ⟨*human bounding box, object bounding box, object class, action class*⟩. Based on whether bounding box location and action category detection processes are contained in an end-to-end framework, existing HOI detection methods can be roughly classified into two lines: two-stage HOI detection methods [2, 7–9, 13, 20] and one-stage methods [5, 16, 18, 22, 34, 37, 47]. In general, HOI detection performance of two-stage methods are largely affected by the off-the-shelf object detector (e.g., Faster R-CNN [30]), increasing efforts have been dedicated to the one-stage HOI detection methods.

Inspired by the recently proposed object detector DETR [1], many researchers [5, 19, 34, 47] are more inclined to design a query-driven detector, where the target quadruples can be obtained in parallel by feeding the decoded query set into multiple feed-forward networks (FFN). However, it is non-trivial to complete human-object association and interaction understanding from single learned query set since the task-aware distinct information should be concerned. Thus, to mine the advantages of both one-stage and two-stage methods, some methods [23, 41, 42] construct a query-driven end-to-end HOI detector with cascaded interaction decoder to disentangle human-object detection and interaction classification. For example, as shown in Figure 1a, CDN [41] introduced a human-object pair decoder and generate a common query set for multiple prediction tasks. Meanwhile, the learned query set is utilized as the initialization of the following cascade decoder. To simplify the post-matching process, as presented in Figure 1b, GEN-VLKT [23] proposed a guided embedding association mechanism, where an instance decoder is first designed to detect humans and

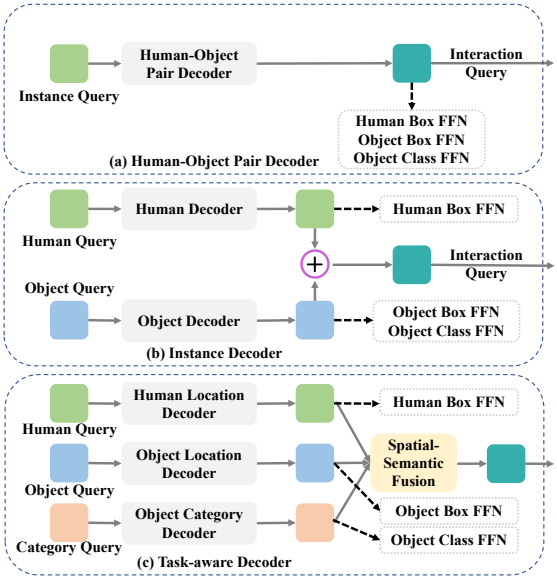
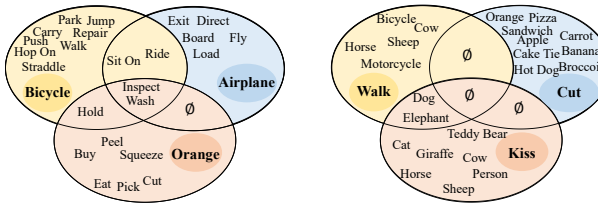


Figure 1: Comparison of the first-level decoder of query-driven Human-Object Interaction detectors that have two-level transformer-based decoders.

objects with two independent query sets, and then learned query sets are summed to initialize the following interaction decoder.

However, existing query-driven HOI detectors with two-level transformer-based decoders still have some drawbacks. On the one hand, taking the representative methods CDN [41] and GEN-VLKT [23] for example, CDN utilized the learned query set from the first-level decoder to perform human-object location tasks as well as the object category classification task simultaneously. Meanwhile, although GEN-VLKT separately generated a query set for human and object location branches, the object category prediction task and object location task are still attached in a single branch. In fact, shared features may not make a good tradeoff for multi-task learning since each task has specific attention and features originally learned for the object box location task may not be optimally compatible with the object category task. Therefore, we divide human-object association decoder into three parts and design three task-aware decoders to study task-aware queries for each task, as shown in Figure 1c. On the other hand, in GEN-VLKT, the input of the cascade decoder is simply obtained by summing the learned human and object queries. Indeed, in accordance with the real-world HOI decision-making process, the interaction detection task is closely related to the human-object spatial features as well as the instance semantic features. To improve the performance of the interaction decoder, it is better to change the query initialization strategy of the cascade decoder and make it more interaction-aware. Last, it is worth noting that some underlying correlations between one object and actions that can be applied to it naturally exist. In other words, valid and possible actions will be narrowed down to a very small range set if the object is determined, and vice versa. As illustrated in Figure 2, we display some object-action combinations in the real-world HOI detection benchmark HICO-DET [2] from the perspectives of object and action, respectively. In Figure 2(a), the overlapped categories between two big circles stand for the actions



(a) Overlapped action category (b) Overlapped object category

Figure 2: Illustration of action-object combination in the real-world HOI detection benchmark HICO-DET from two perspectives, where overlapped area represents shared actions or objects. It is obvious that the distribution of action and object categories are correlated, delivering pivotal cues regarding complex human-object interaction detection.

that two objects can both be applied. For example, action "Ride" and "Sit On" only involve objects "Bicycle" and "Airplane" but can not apply on "Orange". Meanwhile, the action "Squeeze" is only owned by the object "Orange". In addition, as shown in Figure 2(b), there is no overlapped object between the action "Walk" and "Cut". Such relationships deliver pivotal cues and auxiliary information that can be transferred and passed among different tasks. In light of this, existing methods that optimize the object and action branches independently may fail to thoroughly explore the valuable information among multi-task.

To address the aforementioned issues, we propose a divide-and-conquer transformer network that consists of two decoding stages (see Figure 3). Specifically, we first construct three independent task-aware decoders and generate query sets for human and object box tasks as well as the object instance category prediction task, respectively. Then, we design a spatial-semantic fusion module to integrate the above three learned query sets whose output can be utilized to initialize the interaction decoder. In this way, both spatial features of human-object pair and the object semantic feature are all considered, guiding the interaction decoder more consistent with the real-word decision-making process. Furthermore, we design a multi-task knowledge transfer module to pass valuable message between object and action prediction branches, where the action category prediction branch can integrate auxiliary information from the object category prediction branch, and vice versa.

Our main contributions can be summarized in three-fold:

- To the best of our knowledge, we are the first to divide human-object association detection into three independent task-aware decoders. A spatial-semantic fusion module is designed to refine suitable initialization embeddings for the following interaction decoder from former task-aware decoders.
- We design a novel multi-task knowledge transfer module to further pass auxiliary information between object and action category prediction branches.
- Our proposed framework outperforms state-of-the-art models and achieves significant performance improvement on the large HICO-Det dataset. As a byproduct, we will release the datasets, codes, and involved parameters to benefit other researchers and the source codes are available at here now.

2 RELATED WORK

Conventional Two-stage HOI Detection. Conventional two-stage HOI detection methods [2, 7, 8, 15] generally detect all human and object targets and generate human-object pair proposals utilizing the state-of-the-art off-the-shelf object detectors (e.g., Faster R-CNN [30] or MASK R-CNN [11]) in the first stage, and then infer their HOI types by fusing scores of the designed multi-stream networks in the second stage. In a sense, to improve the final detection performance, more attention is paid to the second stage in this bottom-up pipeline. For example, HO-RCNN [2] introduced the first large benchmark for the HOI detection task and proposed a human-object region-based convolutional neural network to characterize the spatial relation between two bounding boxes. To dynamically highlight regions in an image, iCAN [8] proposed an instance-centric attention module to selectively aggregate relevant features used for recognizing HOIs. Noting that the spatial-semantic representation is invariant to complex appearance variations and enables to transmit knowledge among object classes, DRG [7] employed a dual relation graph to capture and aggregate contextual information. In addition, to solve the problem of open long-tailed interactions with objects, FCL [15] first established an object fabricator to generate effective object representations and then combined verbs and fabricated objects to compose new HOI samples. Overall, although existing two-stage HOI detection methods have achieved promising performance, they are time-consuming and badly suffer from the inferior performance of the detector in the first stage. Besides, individual appearance features and coarse spatial relation information are insufficient to predict the relations between human and object targets accurately.

One-stage HOI Detection. Compared with two-stage HOI detection methods, one-stage HOI detection methods [22, 37] directly locate the targeted human and object, and classify the interaction between them from scratch. For example, IP-Net [37] put forward a point-based framework, which regarded HOI detection as a key point detection and grouping problem. In addition, arguing that interaction points implicitly provide context and regularization for HOI detection, PPDM [22] proposed a parallel point detection and matching framework, which is more inclined to practical application. However, these one-stage methods are limited since they need to group or match the generated interaction points with the object detection results in the inference phase, which is time-consuming and complicated.

Recently, inspired by the huge success of query-based objection detection DETR [1], the concept of end-to-end HOI detection has been noticed by many researchers [5, 16, 18, 34, 47]. Specifically, they directly extended the set prediction architecture for object detection to the task of HOI detection. For example, HOTR [18] presented a transformer encoder-decoder architecture where the instance decoder and interaction decoder run in parallel to handle the input from a shared encoder. Besides, QPIC [34] leveraged an attention mechanism to effectively aggregate features for detecting a wide variety of HOIs and realized four simple and intuitive detection heads. In general, the object query of the first decoder layer in these methods is simply initialized as zeros, causing the inferior capability. To solve this issue, CATN [5] proposed category-aware transformer network, where object query is initialized via category priors represented by an external object detection model. Besides, to select the most relevant object-action pairs within an image and refine queries' representation using rich semantic and spatial features, SSRT [16] designed a support feature generator using the output of the encoder. To boost end-to-end models with object-guided statistical priors, OCN [40] introduced a verb semantic model and use semantic aggregation to profit from the object-guided hierarchy. Moreover, OpenCat-L [44] reformulated HOI prediction as sequence generation using a language modeling framework, where the open-set vocabulary is exploited to predict novel interaction classes with a high degree of freedom.

In addition, increasing efforts have been dedicated to disentangling human-object association detection and interaction classification in a cascade [23, 41, 42] manner. For example, CDN [41] is the first transformer-based work to mine the benefits of one-stage and two-stage HOI detection methods by introducing an isolated interaction decoder. Besides, UPT [42] proposed the Unary–Pairwise Transformer to exploit unary and pairwise representations for HOIs. However, these methods still perform human and object detection as well as the object prediction based on a single query set. In fact, it is difficult to make a good trade-off on multi-task learning since the position location and category classification are totally different task. Besides, although GEN-VLKT [23] designed two independent query sets for human and object detection, the input of the following interaction decoder is only come from the output of the object decoder, lacking useful human spatial information. Notably, the information transfer and message passing between the object category prediction and the interaction detection is also totally ignored. To this end, in this paper, we propose a divide-and-conquer transformer network [38] for query-based HOI detection, where the interaction between different task branches can be activated.

3 METHODOLOGY

Overview. As illustrated in Figure 3, our proposed DCT is organized in a cascade manner with two-level decoders, where both task-aware decoder and interaction decoder have the same architecture in our work. Given an image, we first employ a CNN backbone (e.g., ResNet50) followed by a transformer encoder to learn a visual feature sequence, which can be utilized as the global memory for the following two-level decoders. Then, we introduce three independent task-aware decoders to predict the human spatial query set, object spatial query set, and object category query set, respectively. Notably, features at corresponding positions of three sets are treated as the result of a target human-object pair. Next, these three query sets are input to a spatial-semantic fusion module to generate the interaction-related query embeddings, which can be adopted to initialize the learnable query set of the second-level interaction decoder. Finally, a multi-task knowledge transfer module is designed to refine attention information mutually [32] and pass message between object category prediction and interaction type identification branches.

3.1 Divide-and-Conquer Transformer

Backbone. Given an image $\mathbf{Q} \in \mathbb{R}^{3\times H\times W}$, we employ a CNN backbone (e.g., ResNet50) to learn a feature map $\mathbf{Q}^{'} \in \mathbb{R}^{D_o\times H^{'}\times W^{'}}$, where D_o denotes the number of channels, and $H^{'}$ and $W^{'}$ are the

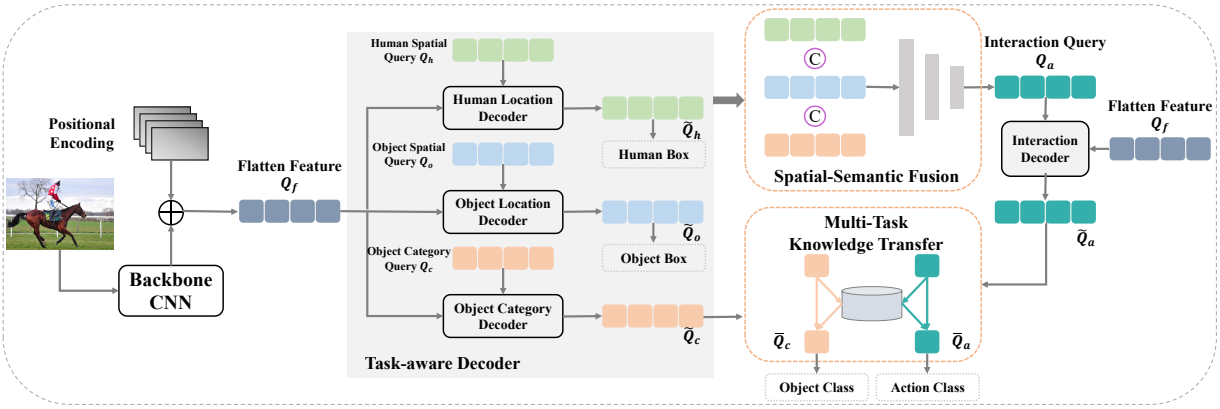


Figure 3: Illustration of the proposed scheme DCT. It consists of two-level decoder: task-aware decoder and cascaded interaction decoder. A spatial-semantic fusion module is introduced to integrate spatial information of human-object pair and semantic information of object. A multi-task knowledge transfer module is also designed to pass valuable message between object and action prediction branches.

dimensions of the obtained feature map. Generally, D_o is too large for the following operation. Thus, we feed the feature map $\mathbf{Q}^{'}$ into a 1×1 convolution layer to reduce the number of channels and derive the compressed visual feature $\mathbf{Q}^{'} \in \mathbb{R}^{D_c \times H^{'} \times W^{'}}$.

Visual Encoder. Thanks to the self-attention mechanism, the transformer encoder can derive feature maps with richer contextual information. In our work, the encoder layer is built upon standard transformer architecture with a multi-head self-attention module and feed-forward network. As the encoder expects a sequence as the input, formally, we collapse the spatial dimensions of $\mathbf{Q}^{'}$ into one dimension and obtain a feature map $\mathbf{Q}_{f_0}$ of shape $[D_c, H^{'} \times W^{'}]$. Similar to CDN [41], we add a positional encoding $\mathbf{P} \in \mathbb{R}^{D_c \times (H^{'} \times W^{'})}$ due to the fact that the transformer architecture is permutation-invariant. Thereafter, we obtain a series of global memory features $\mathbf{Q}_f$ with same dimensions.

Task-aware Decoder. To comprehensively characterize the image-level contextual information for each prediction task, we introduce three task-aware decoders and generate three query sets for human location, object location, and object category prediction tasks, respectively. In particular, for three learnable query sets $(\mathbf{Q}_h, \mathbf{Q}_o, \mathbf{Q}_c) \in \mathbb{R}^{D_c \times N_q}$ where N_q is the number of query embeddings, we adopt a similar operation. Taking $\mathbf{Q}_h$ for example, we feed it together with the global visual feature $\mathbf{Q}_f$ and position encoding $\mathbf{P}$ to the predefined transformer decoder layers, and then we obtain the updated queries $\widetilde{\mathbf{Q}}_h$ as follows,

$$\widetilde{\mathbf{Q}}_h = s_h(\mathbf{Q}_h, \mathbf{Q}_f, \mathbf{P}), \tag{1}$$

where s_h refers to the human location decoder. Similarly, we obtain the updated queries $\widetilde{\mathbf{Q}}_o$ and $\widetilde{\mathbf{Q}}_c$ through object location decoder and object category decoder as follows,

$$\widetilde{\mathbf{Q}}_o = s_o(\mathbf{Q}_o, \mathbf{Q}_f, \mathbf{P}), \quad \widetilde{\mathbf{Q}}_c = s_c(\mathbf{Q}_c, \mathbf{Q}_f, \mathbf{P}), \tag{2}$$

where s_o and s_c denote the object location and category prediction decoders, respectively. Here, queries at the same position from three sets correspond to at most one human-object pair. In a sense, to comprehensively detect the interacted human-object pairs in the input image, the overlapped human-object pair, where one person has interaction with different objects simultaneously, should be represented as distinct query embedding. Toward this end, we set N_q into a relatively large number, such as 64 or 100, which is sufficient for multiple pairs that may appear in the image.

Spatial-Semantic Fusion Module. In real-world HOI detection behavior, the determination of action type in an image is essentially related to the spatial locations of humans and objects as well as the instance semantic features. In a sense, all these three elements can affect the final interaction detection result, which is also compatible with the conventional two-stage HOI detection methods. Therefore, the learnable interaction queries of the interaction decoder should not be initialized randomly. Instead, they should be guided to match the human and object pair spatially and semantically. However, in the context of query-driven HOI detectors, existing methods overlook the advantages of such observation since spatial and semantic information learning is not fine-grained to a specific task. Taking representative end-to-end transformer-based HOI models QPIC [34] and HOTR [18] for example, only one query set is learned to complete four independent tasks simultaneously. Besides, CDN [41] introduced instance decoder and interaction decoder in a cascade manner. The output of its first decoder is shared by the human location, object location, and object category prediction branches, and is also adopted to initialize the interaction queries for the following interaction decoder. In contrast, although GEN-VLKT [23] generates two query sets for human and object instances by concatenating and feeding the initial two query sets forward to the instance decoder, the object location and category prediction branches are also based on the same output, failing to fully learn the task-aware characteristics. Meanwhile, in its implementation, the interaction queries are only simply initialized by summing the two query sets learned by the instance decoders. Therefore, to solve their shortcomings, we introduce a spatial-semantic fusion module to handle the learned three task-aware query sets $(\mathbf{Q}_h, \mathbf{Q}_o, \mathbf{Q}_c)$. Instead of randomly initializing the learnable query embedding of the second-level interaction decoder, we consider the spatial features of humans and objects as well as the semantic feature of objects simultaneously. Intuitively, this is also in line with the real-world HOI detection process and can provide prior knowledge to the following interaction decoder. Specifically, for the input of

the interaction decoder, the learnable interaction query set $\mathbf{Q}_a$ is initialized as follows,

$$\mathbf{Q}_a = l_a(\widetilde{\mathbf{Q}}_h \copyright \widetilde{\mathbf{Q}}_o \copyright \widetilde{\mathbf{Q}}_c), \tag{3}$$

where © denotes the concatenation operation and l_a stands for the multi-layer perceptrons (MLP). In the real implementation, the number of layers is set as 3.

Cascaded Interaction Decoder. In our work, we cascade an interaction decoder that haves the same architecture as the first-level task-aware decoder, and the learnable interaction query set is initialized using the output of the spatial-semantic fusion module. In this way, more HOI detection-related prior knowledge is concerned to improve the performance of interaction decoder. In detail, together with the $\mathbf{Q}_f$ and positional embedding $\mathbf{P}$, we have the decoded interaction query set $\widetilde{\mathbf{Q}}_a$ as follows,

$$\widetilde{\mathbf{Q}}_a = s_a(\mathbf{Q}_a, \mathbf{Q}_f, \mathbf{P}), \tag{4}$$

where s_a represents the cascaded interaction decoder.

Multi-task Knowledge Transfer Module. In a sense, the types of actions that one object can be applied are specific, and vice versa. Once the action is settled, corresponding objects that it can act on will also be narrowed down to a very small range set. Such underlying relationships deliver pivotal cues for both the object prediction and interaction identification tasks. Hence, we introduce the multi-task knowledge transfer module to pass useful and complementary prior knowledge between the object prediction and action prediction branches, where an attention-guided message passing mechanism for information fusion is utilized.

Specifically, similar to PAD-Net[39], we employ the attention mechanism to separately extract knowledge from two different prediction branches, where the attention can act as a gate function to automatically filter useful information. For example, regarding the object category prediction branch, we first obtain an attention map as follows,

$$\mathbf{M}_c = \sigma(\mathbf{W}_c \otimes \widetilde{\mathbf{Q}}_c), \tag{5}$$

where σ is the Sigmoid function, $\mathbf{W}_c$ denotes the convolution parameter, and $\otimes$ refers to the matrix multiplication. Here, in the real implementation of our work, the learned query sets are formally represented as matrices. Then, based on the learned attention map $\mathbf{W}_c$, we can filter useful information from the action prediction branch and obtain the final category features used for object category classification as follows,

$$\overline{\mathbf{Q}}_c = \widetilde{\mathbf{Q}}_c + \mathbf{M}_c \odot (\mathbf{W}_{ac} \otimes \widetilde{\mathbf{Q}}_a), \tag{6}$$

where $\mathbf{W}_{ac}$ denotes the convolution parameter, $\odot$ is the element-wise multiplication. In a similar manner, we can derive the final action features used for action category classification as follows,

$$\begin{aligned} \mathbf{M}_a &= \sigma(\mathbf{W}_a \otimes \widetilde{\mathbf{Q}}_a), \\ \overline{\mathbf{Q}}_a &= \widetilde{\mathbf{Q}}_a + \mathbf{M}_a \odot (\mathbf{W}_{ca} \otimes \widetilde{\mathbf{Q}}_c). \end{aligned} \tag{7}$$

Ultimately, well-designed action and object features can be adopted to predict corresponding classification tasks.

Feed-forward Networks for Prediction. For each learned query set, we introduce a feed-forward network to perform corresponding bounding box prediction or category classification. For the human and object bounding box prediction branches, namely, F_h and F_o, we design a three-layer MLP whose output is the coordinates of the predicted bounding box set, represented as $\mathbf{Q}_h^b \in \mathbb{R}^{N_q \times 4}$ and $\mathbf{Q}_o^b \in \mathbb{R}^{N_q \times 4}$. Moreover, we obtain a vector set $\mathbf{Q}_c^l \in [0,1]^{N_q \times (C+1)}$ from the object category prediction branch F^c, and the output of the action category prediction branch F^a can also be denoted as $\mathbf{Q}_a^l \in [0,1]^{N_q \times M}$.

3.2 Loss Function

Following most of existing query-driven transformer detector QPIC [34] and CDN [41], we first adopt the Hungarian algorithm [1] to pair predicted HOIs with ground-truth HOIs using bipartite matching loss. Then a similar loss function involving the bounding box regression cost and object category classification cost as well as the action prediction cost is calculated to train the whole network. Specifically, for each image, suppose that we have a set of HOI ground-truth T=($\mathbf{T}_h^b \in \mathbb{R}^{K\times 4}, \mathbf{T}_o^b \in \mathbb{R}^{K\times 4}, \mathbf{T}_c^l \in \{0,1\}^{K\times(C+1)}, \mathbf{T}_a^l \in \{0,1\}^{K\times M}$), where K stands for the number of HOIs in one image. It is noticed that N_q is generally set as a relatively large number and greater than the number of ground-truth HOIs K in one image. Therefore, to impose the cardinality of HOI ground-truth set equals to N_q, we first pad ground-truth HOIs T using a set Φ filled with $\varnothing$. Formally, T_p=$T \cup \Phi$. Then it is desirable to minimize the matching loss between sets T_p and Q=($\mathbf{Q}_h^b, \mathbf{Q}_o^b, \mathbf{Q}_c^l, \mathbf{Q}_a^l$) to reach the optimal assignment. Formally, $\hat{\tau}$= argmin $\sum_{k=1}^{N_q} \mathcal{L}_{k,\tau_k}$ reflects the mapping mode, where τ_k denotes the index of predicted HOI for the k-th ground-truth HOI, $\mathcal{L}_{k,\tau_k}$ is the matching cost between them. Overall, similar to QPIC [34] and CDN [41], $\mathcal{L}$ is comprised of five types of loss functions as follows,

$$\mathcal{L} = \gamma^b \mathcal{L}^b + \gamma^{GIoU} \mathcal{L}^{GIoU} + \gamma^p \mathcal{L}^p + \gamma^c \mathcal{L}^c + \gamma^a \mathcal{L}^a, \tag{8}$$

where γ^b, γ^{GIoU}, γ^p, γ^c, γ^a are nonnegative tradeoff parameters for bounding box cost, intersection-over-union loss [31], interactive score cost, object classification cost, and action prediction cost, respectively. The detailed definition of these four losses can be easily found in QPIC [34].

Inference for Interaction Detection. In our work, the designed first-level task-aware decoders and second-level cascaded interaction decoder are all followed by a FFN network to generate N_q final predictions in the corresponding position. Specifically, for the i-th prediction result corresponding with the j-th action, it is defined as $< \mathbf{Q}_{h_i}^b, \mathbf{Q}_{o_i}^b, \text{argmax}_k\, \mathbf{Q}_{c_i}^l(k), \mathbf{Q}_{a_i}^l(j) >$, where k is the index where vector $\mathbf{Q}_{c_i}^l(k)$ reaches the maximal prediction score. Finally, the HOI triplet score is defined as $\text{argmax}_k\, \mathbf{Q}_{c_i}^l(k) * \mathbf{Q}_{a_i}^l(j) * c_i^p$, where c_i^p is the interactive score from the interactive FFN head for the potential human-object pair [41].

4 EXPERIMENT

4.1 Experimental Settings

Datasets. We utilize two datasets for the evaluation: HICO-DET [2] and VCOCO [10]. In the training and testing phases, for fairness, we follow the standard evaluation scheme of baseline methods.

HICO-DET. This is a large-scale HOI detection benchmark by augmenting the HICO [3] classification benchmark with additional instance annotations, including 47,776 images associated with more than 150K human-object pairs. It contains 600 HOI categories over

Table 1: The role mAP performance comparison between our proposed model and state-of-the-art methods on HICO-DET. And the best results are highlighted in bold.

Method	Backbone	Full↑	Rare↑	Non-rare↑
Two-stage HOI detection methods				
Shen et al. [33]	VGG-19	6.46	4.24	7.12
HO-RCNN [2]	CaffeNet	7.81	5.370	8.54
InteractNet [9]	ResNet-50-FPN	9.84	7.16	10.77
GPNN [29]	ResNet-101	13.11	9.34	14.23
iCAN [8]	ResNet-50	12.80	8.53	14.07
PMFNet [36]	ResNet-50-FPN	17.46	15.65	18.00
VSGNet [35]	ResNet-152	19.80	16.05	20.91
FCMNet [25]	ResNet-50	20.41	17.34	21.56
PD-Net [45]	ResNet-152	20.81	15.90	22.28
ConsNet [26]	ResNet-50-FPN	22.15	17.12	23.65
PastaNet [20]	ResNet-50	22.65	21.17	23.09
VCL [13]	ResNet-101	23.63	17.21	25.55
DRG [7]	ResNet-50-FPN	24.53	19.47	26.04
One-stage HOI detection methods				
Wang et al. [37]	ResNet-50	19.56	12.79	21.58
PPDM [22]	ResNet-50	21.73	13.78	24.10
DIRV [6]	ResNet-50	21.81	16.32	23.44
HOTR [18]	ResNet-50	25.10	17.34	27.42
PhraseHOI [21]	ResNet-50	29.29	22.03	31.46
CPC [28]	ResNet-50	29.63	23.14	31.57
HoiTransformer [47]	ResNet-50	23.46	16.91	25.41
AS-Net [4]	ResNet-50	28.87	24.25	30.25
QPIC [34]	ResNet-50	29.07	21.85	31.23
SSRT [16]	ResNet-50	30.36	25.42	31.83
OCN [40]	ResNet-50	30.91	25.56	32.51
MSTR [19]	ResNet-50	31.17	25.31	32.92
UPT [42]	ResNet-50	31.66	25.94	33.36
STIP [43]	ResNet-50	31.60	**27.75**	32.75
CATN [5]	ResNet-50	31.86	25.15	33.84
Zhou et.al [46]	ResNet-50	31.75	27.45	33.03
CDN-S base [41]	ResNet-50	30.96	27.02	31.24
GEN-VLKT-S [23]	ResNet-50	31.88	26.24	33.57
DCT-S(Ours)	ResNet-50	**32.72**	27.48	**34.35**

80 common object categories and 117 actions. Following iCAN [8] and similar to existing HOD detection works, we form a testing set of 9,658 images and a training set of 38,118 images, respectively. Moreover, we separate the HOI category of HICO-DET into two groups: 138 HOI categories with less than 10 training instances (Rare) and 462 HOI categories with 10 or more training instances (Non-Rare), and report results on these two situations.

VCOCO. This dataset is a subset of the COCO dataset [24], containing 2,533 images for training, 2,867 images for validation, and 4,946 images for testing. Each image is manually annotated with binary labels for 29 action categories over 80 common object classes. In our work, we combine images originally for training and validation to form the training set of 4,946 images and keep the remaining as the testing set.

Evaluation Protocols. We utilize the commonly used role mean average precision (mAP) to measure the detection performance on two datasets. Specifically, a detected quadruple is considered true positive if it has correct object and action classes, while human and object bounding boxes have Interaction-over-Union (IOU) ratio greater than 0.5 concerning ground-truth annotations. Besides, for HICO-DET, we report our method in the context of three dataset partitioning: Full, Rare, and Non-Rare. Regarding VCOCO, we notice that some HOIs are defined with no object labels. Thus, similarly to HOTR [18], we introduce two scenarios for role mAP evaluation. For scenario 1, the object bounding box without an object is considered as [0, 0, 0, 0], and the model should correctly predict it. Notably, this scenario fits the missing role due to occlusion. For scenario 2, the occluded object is ignored in the testing phase.

Implementation Details. We utilize ResNet-50 and ResNet-101 [12] as the backbone to extract features from original images, respectively. Following DETR [1], the transformer encoder contains 6 layers with multi-head attention of 8 heads. By changing the number of decoder layers, we have two architectures: DCT-S and DCT-L. The decoder has 3 layers for DCT-S, while the decoder has 6 layers for DCT-L. Besides, both backbone and transformer encoder-decoder are initialized with parameters of DETR pre-trained on the COCO dataset [24], while all other parameters in the neural networks are initialized randomly. In addition, for the training of the VCOCO dataset, we exclude COCO's training images that are contained in the VCOCO testing set when pre-trained the DETR. Pertaining to the optimization, we adopt AdamW [27] as the optimizer with the initial transformer's learning rate to 10^{-4}, the backbone's learning rate to 10^{-5}, and the weight decay to 10^{-4}. Besides, the number of query embedding is set to 64 for HICO-Det and 100 for VCOCO, and its dimension is set to 256. Furthermore, exactly the same with QPIC and CDN, the hyper-parameters of γ^b, γ^{GIoU}, γ^p, γ^c, γ^a are set to 2.5, 1, 1, 1, and 1, respectively. In the real training phase, compared with the performance improvement, the additional training time is acceptable. Under the same experimental settings where 4 GPUs are utilized and batch size is set as 6, the training time of each epoch of vanilla CDN model is 12 minutes and that of our method is 13 minutes on average.

4.2 State-of-the-Art Comparisons

To comprehensively evaluate the proposed framework DCT, we first reported the mAP results of different methods on the HICO-Det dataset in Table 1, consisting of conventional two-stage HOI detection methods and one-stage HOI detection ones. Besides, existing one-stage HOI detection methods are sequentially divided into three groups: conventional one-stage methods and query-driven HOI detectors as well as the query-driven HOI detectors with cascade decoder. From Table 1, we can draw the following observations: 1) Overall, our DCT with ResNet-50 backbone consistently outperforms all the other baselines on full data of HICO-Det. In particular, with the best baseline of query-driven detector CATN [5], DCT achieves an improvement of 0.86. In fact, such an improvement is significant since CATN [5] introduces extra auxiliary information by initializing the object query via category priors obtained from pre-trained word2vector models. In contrast, the performance of our DCT is obtained without visual and linguistic prior knowledge or other auxiliary models. The performance superiority can be attributed to the design of our divide-and-conquer transformer network with internal knowledge transfer. 2) CDN and GEN-VLKT are two representative query-driven detectors with cascaded decoders. For fairness and only to verify the quality of the basic framework, we choose the result of the CDN-S base reported in the original

Table 2: The role mAP performance comparison between our proposed model and state-of-the-art methods on VCOCO.

Method	Backbone	Scenario 1	Scenario 2
Two-stage HOI detection methods			
iCAN [8]	ResNet-50	45.30	–
VCL [14]	ResNet101	48.30	–
DRG [7]	ResNet-50-FPN	51.00	52.40
VSGNet [35]	ResNet-152	51.80	57.00
PD-Net [45]	ResNet-152	52.60	–
FCMNet [25]	ResNet-50	53.10	–
ConsNet [26]	ResNet-50-FPN	53.20	–
One-stage HOI detection methods			
UnionNet [17]	ResNet-50-FPN	47.50	56.20
Wang et al. [37]	ResNet-50-FPN	52.30	–
DIRV [6]	ResNet-50	56.10	–
PhraseHOI [21]	ResNet-50	57.40	–
HoiTransformer [47]	ResNet-50	52.90	–
AS-Net [4]	ResNet-50	53.90	–
QPIC [34]	ResNet-50	58.80	61.00
CATN [5]	ResNet-50	60.10	–
SSRT [16]	ResNet-50	63.70	65.90
Zhou et.al [46]	ResNet-50	66.20	68.50
MSTR [19]	ResNet-50	62.00	65.20
CDN-S [41]	ResNet-50	61.68	63.77
DCT-S(Ours)	ResNet-50	63.00	64.70

paper, where re-weighting and PNMS strategies are removed. Similarly, for GEN-VLKT, we select the results reported in the original paper where a verb classifier with 117 categories for GEN is adopted and large-scale visual-linguistic pre-trained model CLIP is removed, named GEN-VLKT-S (Base verb). As shown in Table 1, compared to CDN-S base using ResNet 50 as the backbone, our DCT promoted mAP from 30.96% to 32.72% on full HICO-Det. Meanwhile, compared to the GEN-VLKT-S (Base verb) with ResNet 50 as the backbone, our DCT promoted mAP from 31.88% to 32.72%. 3) Notably, for the Rare group of HICO-Det, our DCT promoted mAP from 27.02% to 27.48% compared to CDN-S base, while attaining an improvement from 26.24% to 27.48% compered to GEN-VLKT-S (Base verb). Furthermore, for the Non-Rare group of HICO-Det, our DCT promoted mAP from 31.24% to 34.35% compared to the CDN-S base, while attaining an improvement from 33.57% to 34.35% compered to GEN-VLKT-S (Base verb). Notably, such performance improvement is only attributed to the superiority of the designed divide-and-conquer Transformer Network. Notably, compared with some baseline methods like OCN [40] which utilizes the word embeddings, our proposed method outperforms them in the situation that no additional prior knowledge like embeddings of prior categories obtained from pre-trained word2vector and CLIP models are introduced, where only the visual information is utilized.

Table 2 shows the result comparison on VCOCO dataset. As can be seen, compared with the classical end-to-end query-driven HOI detection method QPIC, our proposed DCT has an improvement of 4.2% and 3.7% on two scenarios, respectively. Meanwhile, compared to CDN-S which has the most similar pipeline to ours, DCT attains an improvement of 1.32% on scenario 1. In fact, for fairness and only to verify the quality of the basic framework, the performance of QPIC and CDN-S are more convincing to verify the superiority of our designed basic framework. Compared with SSRT and Zhou et.al, the performance of DCT on small V-COCO dataset is comparable and acceptable. Besides, our proposed DCT achieves 1% improvement on Scenario 1 but obtains a comparable result compared with MSTR.

Table 3: The role mAP performance comparison using ResNet-101 backbone on HICO-DET.

Method	Full	Rare	Non-Rare
QPIC-L [34]	29.90	23.92	31.69
CDN-L base [41]	32.54	24.79	34.04
GEN-VLKT-L [23]	32.55	28.12	33.87
OpenCat-L [44]	32.68	28.42	33.75
DCT-L (Ours)	**33.49**	**29.27**	**34.85**

Table 4: The role mAP performance comparison using ResNet-101 backbone on VCOCO.

Method	Backbone	Scenario 1	Scenario 2
QPIC-L [34]	ResNet-101	58.30	61.70
CDN-L base [41]	ResNet-101	61.70	63.80
GEN-VLKT-L [23]	ResNet-101	63.58	65.93
OpenCat-L [44]	ResNet-101	61.90	63.20
DCT-L(Ours)	ResNet-101	**64.50**	**67.30**

Table 5: The component analysis of our DCT.

Method	Full	Rare	Non-Rare
CDN-S base	30.96	27.02	31.24
DCT-S w/t MTKT	32.35	27.99	33.65
DCT-S w/t SSF	31.31	26.95	32.61
DCT-S (Ours)	**32.72**	27.48	**34.35**

4.3 Backbone Analysis

To gain more deep insights, apart from the ResNet-50 backbone, we further investigated the performance of the proposed DCT adopting ResNet-101 backbone. Here, we selected classic HOI detector QPIC and newest OpenCat. Meanwhile, we chose the state-of-the-art HOI detector with cascade decoder: CDN-L base and GEN-VLKT-L (Base verb) as baselines. As can be seen from Tables 3 and 4, our DCT consistently shows superiority over the representative baseline methods on HICO-DET and VCOCO datasets adopting the ResNet-101 as the backbone, proving the superiority of our model on HOI detection once again. Specifically, compared with the newest work OpenCat, our proposed method increased mAP from 32.68%, 28.42%, and 33.75% to 33.49%, 29.27%, and 34.85% on HICO-DET dataset, respectively. As for the VCOCO dataset, when adopting the ResNet-101 as the backbone, our proposed method improved mAP from 61.90%, 63.20% to 64.50%, 67.30% for two scenarios, validating the superiority of DCT on relatively small dataset.

4.4 Component Analysis

To verify the effectiveness of each key component in our model, namely, spatial-semantic fusion module (SSF) and multi-task knowledge transfer (MTKT) module, we reported the role mAP performance comparison between our proposed model and two derivatives of our model: DCT-S w/t MTKT and DCT-S w/t SSF, where

Table 6: The role mAP performance comparison between DCT-S and DCT-S-single on HICO-DET.

Method	Full	Rare	Non-Rare
CDN-S base	30.96	27.02	31.24
GEN-VLKT-S	31.88	26.24	33.57
DCT-S-single	32.06	28.52	33.12
DCT-S (Ours)	**32.72**	27.48	**34.35**

Table 7: The comparison of parameters and FLOPs.

Methods	FLOPs	Parameters	Training	Testing
CDN-S	89.46G	36.92M	15m 09s	3m 38s
GEN-VLKT-S	89.77G	37.48M	19m 24s	4m 04s
DCT	89.98G	40.67M	15m 24s	3m 40s

MTKT and SSF modules were removed respectively. Meanwhile, the baseline method CDN-S base can be regarded as a version of our DCT where both MTKT and SSF were removed simultaneously. Table 5 showed the role mAP performance on HICO-Det. As can be seen, DCT-S consistently outperforms DCT-S w/t MTKT and DCT-S w/t SSF. Meanwhile, the performance of CDN has been improved by adding any one of our proposed modules. Specifically, on the one hand, when adding the SSF to the baseline CDN, its performance can be improved from 30.96%, 27.02%, and 31.24% to 32.35%, 27.99%, and 33.65%. On the other hand, when adding MTKT module to the baseline CDN, its performance can be improved from 30.96%, 27.02%, and 31.24% to 31.31%, 27.48%, and 32.61%. Above all, we can draw the conclusion that it is necessary to take into account proposed two modules at the same time.

4.5 Bidirectional Knowledge Transfer

In our method, we assume that the knowledge transfer and message passing between object category and interaction prediction branches is bidirectional. To prove this viewpoint, we take the knowledge flow from object category prediction branch to interaction prediction branch for example and investigate its effectiveness in our model. We conduct comparative experiments with one derivative of our model using ResNet-50 termed as DCT-S-single, where knowledge flow from interaction prediction branch to object category prediction branch is removed. The result is displayed in Table 6. Obviously, the performance of DCT-S-single is inferior to DCT which has bidirectional knowledge transfer but also better than CDN-S base and GEN-VLKT-S on the full dataset of HICO-DET. Such observation indicates that the introduction of prior knowledge from object branch to action branch can bring benefits, and there is still improvement room by passing prior knowledge reversely, indicating that both bidirectional knowledge transfers are necessary to complete the information transfer between two tasks.

4.6 Parameter and FLOPs

We also report the FLOPs and parameters of the CDN-S base, GEN-VLKT-S and DCT in Table 7. As can be seen, compared with baselines, our model has experienced a slight increase in its parameter count due to the design of task-aware decoders. However, the model's training time per iteration and testing time are relatively comparable with CDN-S.

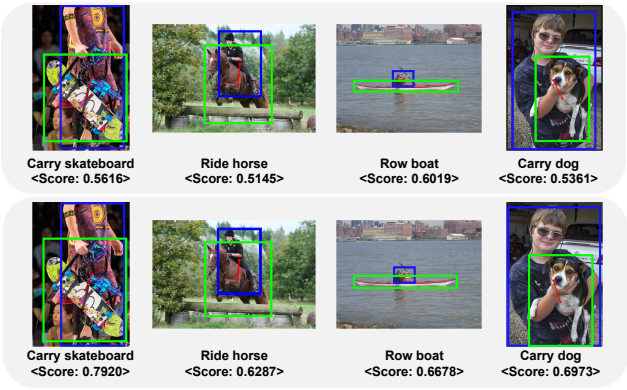


Figure 4: Qualitative examples of the predicted HOI and corresponding action score, where the top row refers to CDN, and the bottom row is the result of DCT.

4.7 Qualitative Analysis

To thoroughly understand the superior performance of our model, we showed some qualitative results of predicted HOI and corresponding action score in Figure 4, where the top row refers to CDN and the bottom row is about DCT. We found that DCT can always produce higher action scores than CDN, reflecting that DCT is more confident in determining action class, which can be attributed to that the initialization of interaction decoder comes from the output of designed spatial-semantic fusion module, rather than simply initialized using zeros. In this way, more HOI detection related information are integrated to enhance the action classification performance. Besides, knowledge transferred from the object branch can also help to improve the score for target category in multi-label classification task.

5 CONCLUSIONS

To solve the shortcomings of existing query-driven HOI detection methods, we propose a novel divide-and-conquer transformer network with knowledge transfer named DCT. In particular, DCT consists of a two-level decoder. For the first-level decoder, three independent task-aware decoders are designed to generate query sets for human and object box prediction branches and object instance category prediction branch, respectively. Then, the spatial location features of human-object pair and instance semantic information of object are fused by the designed spatial-semantic fusion module. Thereafter, the fused spatial-semantic query set is adopted to initialize the learnable query set of the second-level interaction decoder. Moreover, a multi-task knowledge transfer module is also introduced to mutually integrate the prior knowledge between interaction prediction and object category classification branches. Extensive experiments results demonstrate the effectiveness of the proposed model.

Acknowledgements: This research is partially supported by NSF IIS-2309073, ECCS-2123521 and Cisco unrestricted gift. This article solely reflects the opinions and conclusions of its authors and not the funding agencies.

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-End Object Detection with Transformers. In *ECCV*. 213–229.

[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to Detect Human-Object Interactions. In *WACV*. 381–389.

[3] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A Benchmark for Recognizing Human-Object Interactions in Images. In *CVPR*. 1017–1025.

[4] Mingfei Chen, Yue Liao, Si Liu, Zhiyuan Chen, Fei Wang, and Chen Qian. 2021. Reformulating HOI Detection As Adaptive Set Prediction. In *CVPR*. 9004–9013.

[5] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. 2022. Category-Aware Transformer Network for Better Human-Object Interaction Detection. In *CVPR*. 19516–19525.

[6] Haoshu Fang, Yichen Xie, Dian Shao, and Cewu Lu. 2021. DIRV: Dense Interaction Region Voting for End-to-End Human-Object Interaction Detection. In *AAAI*. 1291–1299.

[7] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. 2020. DRG: Dual Relation Graph for Human-Object Interaction Detection. In *ECCV*. 696–712.

[8] Chen Gao, Yuliang Zou, and Jia-Bin Huang. 2018. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. In *BMVC*. 41.

[9] Georgia Gkioxari, Ross B. Girshick, Piotr Dollár, and Kaiming He. 2018. Detecting and Recognizing Human-Object Interactions. In *CVPR*. 8359–8367.

[10] Saurabh Gupta and Jitendra Malik. 2015. Visual Semantic Role Labeling. *arXiv preprint arXiv:1505.04474* (2015).

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *ICCV*. 2980–2988.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

[13] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. 2020. Visual Compositional Learning for Human-Object Interaction Detection. In *ECCV*. 584–600.

[14] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. 2020. Visual Compositional Learning for Human-Object Interaction Detection. In *ECCV*. 584–600.

[15] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. 2021. Detecting Human-Object Interaction via Fabricated Compositional Learning. In *CVPR*. 14646–14655.

[16] A. S. M. Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. 2022. What to look at and where: Semantic and Spatial Refined Transformer for Detecting Human-object Interactions. In *CVPR*. 5343–5353.

[17] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J. Kim. 2020. UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection. In *ECCV*. 498–514.

[18] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. 2021. HOTR: End-to-End Human-Object Interaction Detection With Transformers. In *CVPR*. 74–83.

[19] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. 2022. MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection. In *CVPR*.

[20] Yonglu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Haoshu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. 2020. PaStaNet: Toward Human Activity Knowledge Engine. In *CVPR*. 379–388.

[21] Zhimin Li, Cheng Zou, Yu Zhao, Boxun Li, and Sheng Zhong. 2022. Improving Human-Object Interaction Detection via Phrase Learning and Label Composition. In *AAAI*. 1509–1517.

[22] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. 2020. PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection. In *CVPR*. 479–487.

[23] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. 2022. GEN-VLKT: Simplify Association and Enhance Interaction Understanding for HOI Detection. In *CVPR*. 20091–20100.

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*. 740–755.

[25] Yang Liu, Qingchao Chen, and Andrew Zisserman. 2020. Amplifying Key Cues for Human-Object-Interaction Detection. In *ECCV*. 248–265.

[26] Ye Liu, Junsong Yuan, and Chang Wen Chen. 2020. ConsNet: Learning Consistency Graph for Zero-Shot Human-Object Interaction Detection. In *ACMMM*. 4235–4243.

[27] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *ICLR*.

[28] Jihwan Park, Seungjun Lee, Hwan Heo, Hyeong Kyu Choi, and Hyunwoo J. Kim. 2022. Consistency Learning via Decoding Path Augmentation for Transformers in Human Object Interaction Detection. In *CVPR*. 1009–1018.

[29] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning Human-Object Interactions by Graph Parsing Neural Networks. In *ECCV*. 407–423.

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*. 91–99.

[31] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR*. 658–666.

[32] Yuzhang Shang, Zhihang Yuan, and Yan Yan. 2023. MIM4DD: Mutual Information Maximization for Dataset Distillation. In *NeurIPS 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.).

[33] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. 2018. Scaling Human-Object Interaction Recognition Through Zero-Shot Learning. In *WACV*. 1568–1576.

[34] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. 2021. QPIC: Query-Based Pairwise Human-Object Interaction Detection With Image-Wide Contextual Information. In *CVPR*. 10410–10419.

[35] Oytun Ulutan, A. S. M. Iftekhar, and B. S. Manjunath. 2020. VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In *CVPR*. 13614–13623.

[36] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. 2019. Pose-Aware Multi-Level Feature Network for Human Object Interaction Detection. In *ICCV*. 9468–9477.

[37] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. 2020. Learning Human-Object Interaction Detection Using Interaction Points. In *CVPR*. 4115–4124.

[38] Zhiliang Wu, Changchang Sun, Hanyu Xuan, Kang Zhang, and Yan Yan. 2023. Divide-and-Conquer Completion Network for Video Inpainting. *IEEE TCSVT* 33, 6 (2023), 2753–2766.

[39] Dan Xu, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. 2018. PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing. In *CVPR*. 675–684.

[40] Hangjie Yuan, Mang Wang, Dong Ni, and Liangpeng Xu. 2022. Detecting Human-Object Interactions with Object-Guided Cross-Modal Calibrated Semantics. In *AAAI*. 3206–3214.

[41] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. 2021. Mining the Benefits of Two-stage and One-stage HOI Detection. In *NeurIPS*. 17209–17220.

[42] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. 2022. Efficient Two-Stage Detection of Human-Object Interactions with a Novel Unary-Pairwise Transformer. In *CVPR*. IEEE, 20072–20080.

[43] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang Wen Chen. 2022. Exploring Structure-aware Transformer over Interaction Proposals for Human-Object Interaction Detection. In *CVPR*. 19526–19535.

[44] Sipeng Zheng, Boshen Xu, and Qin Jin. 2023. Open-Category Human-Object Interaction Pre-training via Language Modeling Framework. In *CVPR*. 19392–19402.

[45] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. 2020. Polysemy Deciphering Network for Human-Object Interaction Detection. In *ECCV*. 69–85.

[46] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. 2022. Human-Object Interaction Detection via Disentangled Transformer. In *CVPR*. 19546–19555.

[47] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. 2021. End-to-End Human Object Interaction Detection With HOI Transformer. In *CVPR*. 11825–11834.