HODN: Disentangling Human-Object Feature for HOI Detection

Shuman Fang, Zhiwen Lin, Ke Yan, Jie Li, Xianming Lin*,
Rongrong Ji, *Senior Member, IEEE*

Abstract—The task of Human-Object Interaction (HOI) detection is to detect humans and their interactions with surrounding objects, where transformer-based methods show dominant advances currently. However, these methods ignore the relationship among humans, objects, and interactions: 1) human features are more contributive than object ones to interaction prediction; 2) interactive information disturbs the detection of objects but helps human detection. In this paper, we propose a *Human and Object Disentangling Network* (HODN) to model the HOI relationships explicitly, where humans and objects are first detected by two disentangling decoders independently and then processed by an interaction decoder. Considering that human features are more contributive to interaction, we propose a *Human-Guide Linking* method to make sure the interaction decoder focuses on the human-centric regions with human features as the positional embeddings. To handle the opposite influences of interactions on humans and objects, we propose a *Stop-Gradient Mechanism* to stop interaction gradients from optimizing the object detection but to allow them to optimize the human detection. Our proposed method achieves competitive performance on both the V-COCO and the HICO-Det datasets. It can be combined with existing methods easily for state-of-the-art results.

Index Terms—Human-Object Interaction Detection, Transformer, Visual Attention, Disentangling Features.

I. INTRODUCTION

INSTANCE-LEVEL detection no longer satisfies the requirement of understanding the visual world, but the relationships inference among instances has attracted considerable research interest recently, where Human-Object Interaction (HOI) detection plays a major role. In addition, other high-level semantic understanding tasks, such as activity recognition [39], [42] and visual question answering [26], [35], can benefit from HOI. The goal of HOI detection aims to detect humans and surrounding objects and infer the interactive relations between them, which can be typically represented as triplets of ⟨*human, object, interaction*⟩. Hence, HOI detection consists of three parts: human detection, object detection, and interaction classification.

This work is done when S. Fang works as an intern at Youtu Laboratory, Tencent, Shanghai 200233, China.

S. Fang, J. Li, X. Lin (Corresponding Author), and R. Ji are with the Media Analytics and Computing Laboratory, Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China. (fangshuman@stu.xmu.edu.cn, lijie.32@outlook.com, linxm@xmu.edu.cn, rrji@xmu.edu.cn)

Z. Lin and K. Yan are with Youtu Laboratory, Tencent, Shanghai 200233, China. (xavier.lin@foxmail.com, kerwinyan@tencent.com)

R. Ji is also with the Institute of Artificial Intelligence, Xiamen University, and Fujian Engineering Research Center of Trusted Artificial Intelligence Analysis and Application, Xiamen University, 361005, China. (rrji@xmu.edu.cn)

*Corresponding Author: Xianming Lin (linxm@xmu.edu.cn)

Based on variants of RCNN [10], [31], conventional methods usually detect instances firstly and enumerate all human-object pairs to predict secondly [3], [6]–[8], [12], [13], [18]–[20], [27], [36], [37], [48], or directly identify the pairs that are likely to interact [15], [22], [40], [49]. These methods suffer from the lack of contextual information due to the locality of convolutional layers and pooling layers. Nowadays, transformer-based HOI detectors [4], [16], [33], [45], [51] are proposed to handle this problem in an end-to-end manner. Benefiting from the attention mechanism, these networks can extract global features rather than local ones. However, in the very beginning, the transformer-based methods [33], [51] utilize a simple pipeline with a single encoder-decoder pair to detect the triplets of HOI, which struggles with handling localization and classification simultaneously. To deal with this problem, recent methods [4], [16], [45] disentangle the HOI task as tasks of instance detection and interaction classification. These methods usually utilize two decoders, whether parallel [45] or cascaded [4], [16], to handle the corresponding tasks. Among them, DisTR [50] takes a further step to disentangle both encoders and decoders for the two sub-tasks. The disentanglement of sub-tasks lets different modules focus on their corresponding tasks and improve the whole performance. However, for the instance detection task, previous works regard a pair of human and object as one instance, and process the instance features as a whole, which ignores the distinct effects between humans and objects. We argue that humans and objects play different roles in HOI detection, and the mutual effect among humans, objects, and interactions should also be analyzed.

To dig out the comprehensive relationships among humans, objects, and interactions, we conduct experiments and analyze from two respects, *i.e.*, 1) how humans and objects impact interactions and 2) how interactions impact humans and objects. In Section IV-B, we analyze these by masking the regions of different parts in the images and removing the module for interaction prediction. The results show that: 1) both humans and objects make contributions to interaction prediction, but humans contribute much more; 2) human detection needs the help of interactions while object detection will be disturbed.

Motivated by the different effects between humans and objects, we emphasize the necessity of utilizing disentangled human-object features. Hence, we propose an end-to-end transformer-based framework, termed *Human and Object Disentangling Network* (HODN), to explicitly model the relationships among humans, objects, and interactions. A simple comparison with previous works is visualized in Figure 1.

Fig. 1. Comparison of recent disentangling transformer-based methods. Previous works usually disentangle the HOI task into instance detection and interaction prediction by introducing two decoders. (a) Parallel methods [4], [16] adopt two relatively independent decoders. (b) Cascaded methods [45] handle these parts in series. (c) DisTR [50] takes a further step based on parallel methods, where both encoder and decoder are disentangled. (d) Our HODN contains two parallel detection decoders for human and object features and one interaction prediction decoder. Enc: Encoder; Dec: Decoder; Inst: Instance; Inter: Interaction.

As depicted, we use two separate detection decoders, *i.e.*, *human decoder* and *object decoder*, to extract human and object features independently, which will be processed by an *interaction decoder*. To make sure the human features contribute more there, a *Human-Guide Linking* (HG-Linking) method is utilized for the interaction decoder to focus on the human-centric regions. In particular, the interaction decoder receives the human features as the positional embeddings for all attention layers. These position embeddings, *a.k.a.*, interaction queries, are used to guide the decoder where to focus on. To link humans with the surrounding objects, object features are fed into the first layer of the interaction decoder to provide prior knowledge. With this, the interaction decoder can not only make use of the information of humans and objects but also make the human features dominant and the object ones auxiliary. Furthermore, due to another HOI relationship that interactive information obstructs object detection but helps human detection, we propose a training strategy, termed *Stop-Gradient Mechanism* (SG-Mechanism) to process interaction gradients separately. During back-propagation, the SG-Mechanism stops interaction gradients from passing through the object decoder but maintains them into the human decoder the same as the common practice. This not only inhibits the negative impact of interactions on object detection but keeps the slightly positive on human detection, which brings the best detection performance for both humans and objects.

To evaluate the performance of our method, we conduct extensive experiments by following the previous works on the widely used datasets, *i.e.*, V-COCO [9] and HICO-Det [3], where our method achieves competitive results. Moreover, we combine two latest works [23], [47] with our method, bringing 2.40% and 0.91% relative gains respectively, to achieve the state-of-the-art performance. Visualization experiments also verified that our method can localize humans and objects more preciously and can focus on the interaction regions. We owe these to the disentangling of different parts and the established comprehensive relationships among them.

Concretely, we summarize our work as: 1) we found that humans contribute more to interaction prediction, and interactions have opposite influences on the detection of humans and objects. 2) We propose HODN to explicitly model the relationships, where an HG-Linking is utilized by the interaction decoder to make human features dominant and

object ones auxiliary, and an SG-Mechanism is proposed to handle interaction gradients differently. The HODN achieves competitive experimental results and can be easily combined with existing methods for state-of-the-art performance.

II. RELATED WORK

Based on the used detector, HOI methods can be categorized into two main streams: traditional methods and transformer-based methods, where the former are based on variants of R-CNN [10], [31] and the latter are on DETR [2].

Traditional Methods

Traditional methods can be further divided into two-stage and one-stage methods. Two-stage methods [3], [7], [8], [18] rely on an off-the-shelf object detector to localize all instances, including humans and objects. Then they enumerate all human-object combinations, crop the features inside the localized regions, and process cropped features by multi-stream networks. Typically, the multi-stream networks include a human stream, an object stream, and a pairwise stream. To improve the performance, some works introduce extra streams with extra knowledge, including spatial features [12], [27], [36], word embeddings [6], [13], human postures [19], [20], [37], [48], or above in combination. Another line of works also introduces graph neural networks to model the relationship of human-object pairs [25], [28], [36], [38], [43]. However, two-stage methods suffer from finding the interactive human-object pairs in an overwhelming number of permutations.

To alleviate the number of redundant non-interactive pairs, one-stage methods introduce the concept of interaction point as the anchor to directly identify the pairs that are likely to interact [15], [22], [40], [49]. By parallelly processing the instance detection branch and the interaction point prediction branch, they can match the most similar combination between the two independent branches to complete HOI detection. In particular, the introduced interaction point is the midpoint of the center points of human and object [22], [22], or the center point of the human-object pair union box [15]. Instead of using one interaction point, GGNet [49] infers a set of points to predict interactions more robustly. However, either one point or a set of points, such a concept may introduce a lot of useless regions and even some misleading information, *i.e.*, one-stage methods may not work in situations when the human and

the object in the interaction are far apart or when multiple instances are overlapping.

In summary, no matter the one-stage or the two-stage methods, traditional methods are highly dependent on the quality of object detection and can not capture the interactive relationships accurately due to the lack of contextual information in CNNs.

Transformer-based Methods

Considering that transformer architectures have succeeded in many computer vision tasks [5], [34], recent works introduce the transformer into the HOI detection task based on DETR [2]. Instead of relying on a pre-trained object detector, transformer-based methods generate a set of HOI triplets. The attention mechanism in the transformer encoder can extract the global feature, including humans, objects, and context. During the transformer decoder stage, these methods introduce a set of anchor-like learnable positional embeddings, *a.k.a* queries, to predict HOI triplets directly in an end-to-end manner. With the global feature, the network can easily mine the interactive information between humans and objects so that transformer-based methods can get quite advanced performance. Moreover, as the number of queries determines that of predicted results, transformer-based methods address the problem of a large number of redundant human-object pairs suffered by the two-stage ones. And due to the learnable nature of queries, the regions where the network is interested can be optimized to solve the problem faced by the one-stage methods.

In the beginning, transformer-based works examine how to design the transformer decoders, of which there exist two main streams, *i.e.*, single-decoder and two-decoder methods. The single-decoder methods, *e.g.*, QPIC [33] and HOI-Transformer [51], use only one decoder to handle HOI detection. The mixed features may make the decoder unable to focus on target regions. Some works have realized the differences between attention regions of instances (the pairs of humans and objects) and interactions, and then they disentangle the HOI task into two sub-tasks: *i.e.*, instance detection and interaction prediction. These two-decoder works can be further divided into cascaded methods [45] and parallel ones [4], [16]. By providing instance features to guarantee the performance of interaction prediction, the cascaded method CDN [45] detects instances firstly and predicts interactions secondly. However, object detection will be disturbed by interactions due to the cascaded linking way. To keep the instance decoder away from interaction impacts, HOTR [16] and AS-Net [4] use two parallel decoders to deal with instance detection and interaction prediction independently. However, these methods still suffer from mis-grouped instance-interaction pairs. Recent parallel work, DisTR [50], utilizes an attention module to fuse instance features into interaction representations to provide joint configurations of them. With the connection of instances and interactions, DisTR proposes to further disentangle features, where the transformer encoder and decoders are decoupled. Another work, HOD [46], disentangles the decoder into three independent parts. Then, HOD introduces random erasing for the object decoder to improve generalization and pose information for the human decoder to augment representations.

No matter how the decoders are designed, all existing methods ignore the differences between humans and objects, *i.e.*, they do not dig out the relationships among humans, objects, and interactions, indicating that there exists much space for them to improve.

Some other transformer-based methods also try to introduce additional information to boost performance. OCN [44], PhraseHOI [21] and GEN-VLKT [23] use word embeddings to assist in interaction prediction. Besides linguistic features, STIP [47] also adopts spatial features with graph methods. These methods also fail to explore the relationships among humans, objects, and interactions. We argue that our method is flexible so that can combine with them easily to achieve higher performance.

III. METHOD

The overall architecture of our proposed HODN is shown in Figure 2 and stated in Section III-A. Our framework is proposed to model the relationships among humans, objects, and interactions based on: 1) human features make more contributions to interaction prediction than object ones; 2) interactive information disturbs object detection but assists in human detection. We propose a *Human-Guide Linking method* (HG-Linking) in Section III-B to help the interaction decoder predict interactions by holding the former relationship. In Section III-C, we introduce the detail of a training strategy named *Stop-Gradient Mechanism* (SG-Mechanism), which is proposed to satisfy the latter one.

A. Overall Architecture

1) Global Feature Extractor: Given an image $x \in \mathbb{R}^{3 \times H \times W}$, the CNN backbone extracts the visual feature map $z_b \in \mathbb{R}^{C \times H' \times W'}$ from it, where H and W are the height and width of the input image, H' and W' are those of the feature map, and C is the number of channels. Then the visual feature map z_b is reduced in channel dimension from C to d by a projection convolution layer with a kernel size of 1×1. Next, the spatial dimensions of it are collapsed into one dimension by using a flatten operator. The processed feature map $z_{src} \in \mathbb{R}^{d \times (H' \times W')}$ combined with a positional encoding $p \in \mathbb{R}^{d \times (H' \times W')}$ is fed into the transformer encoder to get the global memory feature z_e. In this stage, the multi-head self-attention can focus on not only regions of humans and objects but also global contextual information.

2) The HOI Decoders: The global memory feature z_e, along with the positional encoding p, is then utilized by two parallel decoders to provide contextual information. The two parallel decoders, the *human decoder* and the *object decoder*, are used to detect their corresponding targets independently. The human decoder transforms a set of randomly initialized learnable positional embeddings $Q_H = \{q_i^h \mid q_i^h \in \mathbb{R}^d\}_{i=1}^N$ (*human queries*) into Q_H^{out} (*human features*) layer by layer, where d is the channel dimension of the encoder and N is the number of positional embeddings. So does the object decoder, which transforms another set of learnable positional embeddings Q_O (*object queries*) with the same size as Q_H into Q_O^{out} (*object features*). In this stage, the human

Fig. 2. Overview framework of *Human and Object Disentangling Network* (HODN). Given an input image, the global memory feature z_e is extracted by the CNN backbone and the transformer encoder. Then two parallel decoders, *i.e.*, human decoder and object decoder, introduce two sets of learnable positional embeddings (Q_H and Q_O) to obtain human features and object features (Q_H^{out} and Q_O^{out}). The interaction decoder receives them to mine interactive information by the Human-Guide Linking method (described in Section III-B). Finally, the outputs of three decoders are sent into corresponding FFNs to get HOI predictions. During the back-propagation stage of training, the *Stop-Gradient Mechanism* (presented in Section III-C) is used to process the interaction gradients in a particular way.

features and the object features with the same subscript are considered as the human-object pair automatically, *i.e.*, the human-object pair features can be represented as $\left\{ (q_i^{h,out}, q_i^{o,out}) \mid q_i^{h,out} \in Q_H^{out}, q_i^{o,out} \in Q_O^{out} \right\}_{i=1}^{N}$. Note that to guarantee this, the two sets of queries, Q_H and Q_O, are initialized to be equal. Then, the *interaction decoder* receives the human-object pair features, *i.e.*, Q_H^{out} and Q_O^{out}, to query possible human-object pairs and dig out interaction knowledge between them. The output from the interaction decoder is denoted as Q_A^{out}.

3) Final Prediction Heads: The final part of our framework includes four feed-forward networks (FFNs), *i.e.*, a human-box FFN, an object-box FFN, an object-class FFN, and an interaction FFN. The outputs of three decoders, *i.e.*, Q_H^{out}, Q_O^{out}, and Q_A^{out}, are then fed into the corresponding FFNs to get human-box vectors, object-box vectors, object-class vectors as well as interaction-class vectors. These vectors share the same length in design, which is the same size of queries. Therefore, we can get a set of HOI predictions with the size of N, each of which is presented as ⟨*human box, object box, object class, interaction class*⟩. With these HOI predictions, we follow previous works [16], [33], [45], [51] for training and inference.

B. Human-Guide Linking

We follow DETR [2] to design the human decoder and the object decoder. However, unlike the human or object decoder which only processes one kind of information, the interaction decoder needs to fuse both human features and object features. Considering that human features contribute more than objects, we argue that naively taking the addition of two features as inputs will not bring satisfactory performance. Hence, a specific link method between the interaction decoder and the others two decoders is non-trivial.

We first review the architectural details of the vanilla transformer decoder, which takes a set of learnable positional embeddings, the global memory feature z_e, and the positional encoding p as inputs to localize the bounding boxes and predict the classes of targets, where the learnable positional embeddings, *a.k.a.*, queries, are designed to learn the potential target regions. The transformer decoder is a stack of decoder layers, each of which is composed of three main parts: a self-attention module, a cross-attention module, and a feed-forward network (FFN). The self-attention module in the i-th layer can be formulated by the following:

$$A_i^{self} = \text{softmax}\left(\frac{(Q_{i-1}^{out} + Q)(Q_{i-1}^{out} + Q)^\top}{\sqrt{d}} \right) Q_{i-1}^{out}, \quad (1)$$

where d denotes the channel dimension of the decoder, Q is the queries, and Q_{i-1}^{out} is the outputs of the previous decoder layer. Note that the Q_0^{out} is initialized with zeros, so the self-attention module in the first layer is meaningless and can be skipped as mentioned in DETR [2]. From Eq. 1, it can be seen that with the help of the queries, the self-attention module can capture the relationships among outputs of the previous layer and inhibit duplicate ones. The cross-attention module in the i-th layer can be formulated as:

$$A_i^{cross} = \text{softmax}\left(\frac{(A_i^{self} + Q)(z_e + p)^\top}{\sqrt{d}} \right) z_e, \quad (2)$$

where p is the positional encoding used by the encoder and z_e is the global memory feature output from the encoder. During the cross-attention module, A_i^{self} is aggregated with the highly responsive parts in z_e and is further refined to improve the detection of targets, where the queries provide information of the attention position. Note that both in self-attention and cross-attention, the queries are utilized to supply the spatial information of the distinct regions. Hence, we argue that the

Fig. 3. The details of the *Human-Guide Linking* depicted in Section III-B.

queries act as a guide to force the decoder where to focus on, which is also verified by DETR [2].

Different from the vanilla transformer decoder, including the human decoder and the object decoder, the interaction decoder needs to model the HOI relationships. As analyzed before, there is a strong correlation between interactions and humans. Hence, based on the functionality of two attention modules, we propose a *Human-Guide Linking* method to make human features more contributive. The detail of this specific linking method is given in Figure 3. In particular, we regard the human features $\boldsymbol{Q}_H^{out}$ from the human decoder as the positional embeddings, *a.k.a.*, interaction queries, which are utilized by all attention layers to guide the interaction decoder where to concentrate on. With this, the attention of the interaction decoder can be around humans. However, just making the interaction decoder pay attention to human-centric regions is not sufficient. Object features $\boldsymbol{Q}_O^{out}$ should also be considered albeit less contributive than human ones. Instead of enumerating permutations that may generate $N \times N$ possible pairs, we consider a one-to-one same-subscript matching strategy to assign the human-object pairs. So we regard the additions of $\boldsymbol{Q}_H^{out}$ and $\boldsymbol{Q}_O^{out}$ as assigned pairs and feed them into the self-attention module in the first layer to dig out the relation between them and remove duplication prediction. Particularly, we change the attention formula of the self-attention in the first layer as below:

$$A_1^{self} = \text{softmax}\left(\frac{(\boldsymbol{Q}_H^{out} + \boldsymbol{Q}_O^{out})(\boldsymbol{Q}_H^{out} + \boldsymbol{Q}_O^{out})^\mathsf{T}}{\sqrt{d}}\right)\boldsymbol{Q}_O^{out}. \tag{3}$$

Unlike the vanilla decoder where the self-attention in the first layer is meaningless, we make full use of it, which can construct the relationship between humans and objects quickly. The attention modules in the following decoder layers only take human features into account, not object features anymore.

With human features dominant and object ones auxiliary, the interaction decoder can effectively model the interaction relationships between humans and objects.

C. Stop Gradient Mechanism

Considering another HOI relationship that interactions have a negative influence on object detection but a positive influence on human detection, we propose a special training strategy, *Stop Gradient Mechanism* (SG-Mechanism), to handle the relationship. As shown in our framework, the outputs from the HOI Decoders are fed into the Final Prediction Heads, *i.e.*, the $\boldsymbol{Q}_A^{out}$ are sent into the interaction FFN, the $\boldsymbol{Q}_H^{out}$ are sent into the human-box FFN, and the $\boldsymbol{Q}_O^{out}$ are sent into the object-box and object-class FFN. The final loss to be minimized is calculated in four parts: location loss of human boxes L_{loc}^h, that of object boxes L_{loc}^o, object classification loss L_o, and interaction classification loss L_a, formulating as:

$$L_{total} = L_{loc}^h + L_{loc}^o + \lambda_o L_o + \lambda_a L_a, \tag{4}$$

where λ_o and λ_a are the weights of two classification losses, and L_{loc} is computed by box regression L_1 loss and the GIoU loss [32] with weighting coefficients λ_{reg} and λ_{giou}, which can be formulated as:

$$L_{loc} = \lambda_{reg} L_{reg} + \lambda_{giou} L_{giou}. \tag{5}$$

The gradients w.r.t. L_{loc}^o and L_o update the parameters of the object decoder towards the optimal direction of object detection. In general, the gradients w.r.t. L_a will pass through both the human decoder and the object decoder since the interaction decoder receives the outputs from the two decoders. However, considering the negative influence on object detection, we stop the gradients w.r.t. L_a propagating into the object decoder to keep optimal updates for object detection, which means the update of parameters of the object decoder is only computed by losses related to objects:

$$w_o \leftarrow w_o - \alpha\left(\nabla_{w_o}\left(L_{loc}^o + \lambda_o L_o\right)\right), \tag{6}$$

where w_o denotes the parameters of the object decoder and α is the learning rate. For the learnable nature of the positional embeddings, the update of the object queries $\boldsymbol{Q}_O$ can be represented as:

$$\boldsymbol{Q}_O \leftarrow \boldsymbol{Q}_O - \alpha\left(\nabla_{\boldsymbol{Q}_O}\left(L_{loc}^o + \lambda_o L_o\right)\right). \tag{7}$$

Since human detection is not disturbed by interaction gradients but benefits from them, we maintain the update for the human decoder with parameters w_h as:

$$w_h \leftarrow w_h - \alpha\left(\nabla_{w_h}\left(L_{loc}^h + \lambda_a L_a\right)\right). \tag{8}$$

So does the update for learnable human queries:

$$\boldsymbol{Q}_h \leftarrow \boldsymbol{Q}_h - \alpha\left(\nabla_{\boldsymbol{Q}_h}\left(L_{loc}^h + \lambda_a L_a\right)\right). \tag{9}$$

With SG-Mechanism, the detection of both humans and objects can achieve the best performance. And with better detection performance, the following interaction prediction can be further improved.

(a) Masking objects (b) Masking humans

(c) Performance vs. Mask-degree

Fig. 4. The interaction category of both (a) and (b) is "throw_obj". For human understanding, it is easy to determine the interaction class in (a) but hard in (b). And in (c), the mAP results based on the V-COCO test set demonstrate that the HOI detection with masked humans is much worse than that with masked objects and the gap becomes larger with higher masking probability.

	R_o	P_o	mAP_o	R_h	P_H	AP_h
QPIC	73.72	21.51	43.69	98.67	6.92	74.95
w/o action	76.10 (↑)	24.54 (↑)	45.34 (↑)	98.46 (↓)	1.53 (↓)	71.02 (↓)
gap	2.38	3.03	1.65	-0.21	-5.39	-3.93
DisTR	77.11	23.51	44.84	98.64	4.99	73.80
w/o action	78.03 (↑)	23.72 (↑)	45.20 (↑)	98.44 (↓)	4.00 (↓)	72.67 (↓)
gap	0.92	0.21	0.36	-0.20	-0.99	-1.13
HODN	85.20	25.10	53.20	98.93	5.77	81.68
w/o action	86.01 (↑)	25.28 (↑)	53.31 (↑)	98.78 (↓)	5.04 (↓)	80.59 (↓)
gap	0.81	0.18	0.11	-0.15	-0.73	-1.09

setting, based on the number of HOI categories in the training set, we design three evaluation types: full, rare, and non-rare.

3) Implementation Details: We adopt ResNet-50 [11] followed by a six-layer transformer encoder as our global feature extractor. For HOI decoders, the layer number of the human, object, and interaction decoder are all set to 6. Layers in encoder and decoders have 8 heads, and the dimension inside the transformer architecture is 256. The query size N is set to 100 for V-COCO and 64 for HICO-Det since the average number of variant human-object pairs per image in V-COCO is larger than that in HICO-Det, which is as mentioned in CDN [45]. The parameters of our proposed network are initialized with MS-COCO pre-trained DETR [2]. For the missing parameters, we initialize them randomly. During training, the AdamW [29] optimizer with the batch size of 16, weight decay of 10^{-4}, the initial learning rate of 10^{-5} for the backbone, and 10^{-4} for other parts is used. For V-COCO, to eliminate overfitting, we train our HODN for 90 epochs, with learning rates decaying by 10 times every 10 epochs after the 60th epoch and freeze the parameters of the backbone. And for HICO-Det, We train the whole HODN for 90 epochs where learning rates are decayed by 10 times at the 60th epoch. The hyper-parameters weight coefficients in training loss $\lambda_{reg}, \lambda_{giou}, \lambda_o$ and λ_a are set to 1, 2.5, 1, 1, respectively. And the threshold of pair-wise NMS is set as 0.7 in inference, the same as CDN. As for application experiments, all experimental settings are the same as those used in the original paper [23], [47] for a fair comparison.

IV. EXPERIMENT

A. Experiment Setup

1) Dataset: We follow the evaluation of most previous works and report the mean average precision (mAP) on two public benchmarks, *i.e.*, V-COCO [9] and HICO-Det [3]. V-COCO dataset, a subset of MS-COCO [24], contains 10,346 images (5,400 in the trainval set and 4,946 in the test set). The images in it are annotated with 80 object classes and 29 action classes. Among these action classes, four of them are not associated with semantic roles, so the role mAP of the V-COCO test set is only computed with the other 25 action classes. In HICO-Det, there are 38,118 images for training and 9,658 for testing annotated with 80 object classes and 117 action classes.

2) Metric: We use the mean average precision (mAP) to report performance. An HOI category is defined as an action class for V-COCO while a pair of an object class and an action class for HICO-Det. Same as the standard evaluation scheme, a detection is judged as a true positive if the predicted human boxes and object boxes both have IoUs larger than 0.5 with the corresponding ground-truth boxes and if the predicted HOI category is also correct. And the AP is calculated per HOI category. For the V-COCO dataset, we report the role mAP in two scenarios, where scenario 1 needs to predict the cases in which humans interact with no objects while scenario 2 ignores these cases. For the HICO-Det dataset, we report performance in two settings: default setting and known object setting. In the former setting, the performance is evaluated on all test images while in the latter one, each AP is calculated on images that contain the target object class. And in each

B. Relationships among humans, objects, and interactions

In this sub-section, we first verify whether humans and objects have different mutual effects on the interactions. In terms of *how humans and objects impact interactions*, we argue that humans make more contributions to interaction prediction than objects since human feature contains more information (such as human postures and facial expression) that is strongly related to interactions [1], [8], [25]. It can be easily verified by masking the regions of humans or objects in the images and inferring what the interaction is. As in Figure 4a, although we can not see the object, it is still easy to infer that a man is throwing something. On the contrary, in Figure 4b, it is quite hard to determine whether the interaction is "hit_obj" or "catch_obj" or any else with only the "sports

ball" visible. To further verify the hypothesis, we use a pretrained HOI detector, QPIC [33], to observe differences in results by masking humans and objects on the test set of V-COCO [9] respectively. Note that the average of human areas and that of object areas are similar, which means that masking humans or objects may cause the same degree of information missing, so the performance comparison between the two is principally fair. Results in Figure 4c show that, without either humans or objects, the performance drops sharply. Regardless of probability, the performance with masked humans is always lower than that with masked objects. Furthermore, the performance gap becomes larger as the degree of masking increases. That is to say, both of them make contributions to interaction prediction, but humans contribute much more than objects.

To further explore whether interactions impact humans and objects differently, we modify QPIC by removing the action-class feed-forward network which is used to classify the interactions, and we train it with the same setting as QPIC. The V-COCO benchmark serves as the training set and test set. Without the interaction classification, all features are directly optimized for human detection and object detection so that the detection performance of both should be improved intuitively. However, things do not turn out that way. We report the results with and without interaction classification in Table I, along with the gaps as the subtraction of results without interaction and results with interaction. As shown in Table I, without actions, object detection can be improved a lot (recall rate, precision rate, and mAP are increased by 2.37, 3.03, and 1.65, respectively), implying that object detection is disturbed by actions. On the contrary, the performance of human detection declines, with metrics decreasing by 0.21, 5.39, and 3.93, demonstrating the necessity for interactions. We also conduct similar experiments on DisTR [50] and our HODN since both methods claim the disentanglement for instances and interactions. From Table I, the slight performance discrepancies of object detection in DisTR indicate that separating detection from interactions is an efficient way to improve object detection. The better results compared with QPIC verify this. However, human detection in DisTR shows inadequate performance since the relatively independent instance detection stream stops interactions assisting human detection. From the results of QPIC and DisTR, we can conclude that interactions influence human detection and object detection quite differently and instance-level disentanglement only helps object detection. In terms of HODN, it shows significant superiority in both human and object detection. The gap between HODN and HODN without action is also less than other methods, we owe this to our disentangled human and object decoders and our SG-Mechanism strategy.

As analysis and experimental results above, we conclude the HOI relationships are: 1) for interaction prediction, human features make more contributions than object ones; 2) for human and object detection, interactive information assists in the former but obstructs the latter.

C. Quantitative Analysis

1) Performance Comparisons: We first evaluate the performance of our method on the HICO-Det test set with ResNet-

50 [11] as the backbone, and report result in Table II. Our method achieves a competitive result, *e.g.*, 33.14 mAP on the full evaluation for the default setting. Compared with transformer-based single-decoder works HoiTransformer [51] and QPIC [33], our HODN has achieved 41.26% and 14.00% relative mAP gain. Even when comparing PhraseHOI [21], OCN [44], and SSRT [14] which introduce prior knowledge of language, our method still attains relative improvements of 13.14%, 7.21%, and 9.16%. We argue that this gap comes from the limitation of the single-decoder that can not explicitly model the HOI relationships regardless of adopting auxiliary knowledge. Among multiple-decoders methods, our method outperforms 32.03% by HOTR [16], 14.79% by AS-Net [4], 4.28% by CDN [45], and 4.38% by DisTR [50]. Even with multiple decoders, these methods ignore the different potentiality between humans and objects, showing unsatisfactory performance. We can conclude that except for the GEN-VLKT [23] that introduces extra semantic information, our method achieves the best performance. We then conduct performance comparison experiments on the V-COCO test set. From Table III, our method achieves 67.0 role mAP in scenario 1 and 69.1 in scenario 2, outperforming all existing methods without extra knowledge. In particular, compared to methods only appearance feature referred to, we promote the state-of-the-art work, DisTR [45], with about 1.21% and 0.88% performance improvement under the two scenarios. Moreover, compared to methods with extra knowledge, our method is still competitive with similar performance with state-of-the-art method STIP [47] (67.0 of ours compared with 66.0 of STIP in scenario 1 and 69.1 compared with 70.7 in scenario 2). Note that both GEN-VLKT and STIP introduce extra knowledge and complicated structures. However, their performance varies on the two datasets. GEN-VLKT introduces the text encoder in CLIP [30] to initialize the weights of the interaction classifier FFN, which provides abundant prior knowledge on HICO-Det. However, for V-COCO with a large number of categories, the training samples are insufficient to train which leads to inefficiency. STIP uses graphs to construct the relationships among humans and objects. It is easy to form interactive relationships when dealing with a few HOI triplets. Nevertheless, it becomes quite hard when the number of triplets increases. Therefore, it performs excellent results on V-COCO, while showing insufficiency on dense triplets' datasets like HICO-Det. We argue that our method considers less extra knowledge which makes our method more generalized. Our method performs consistently on both benchmarks with somehow similar performance to the best one on each dataset, which implies the practicality of our method.

2) Application to Existing Works: As we analyzed before, many existing works only disentangle HOI detection into instance detection and interaction prediction, *i.e.*, they only construct the relationships between instances and interactions, not distinguish humans and objects. Note that our motivation and method are orthogonal to them. We can combine HODN with them by further disentangling instances into humans and objects and adopting our proposed HG-Linking and SG-Mechanism. Almost all two-decoder transformer-based methods can combine with ours. Here we take STIP [47] and GEN-

TABLE II
PERFORMANCE COMPARISON ON HICO-DET DATASET. EACH LETTER IN THE FEATURE COLUMN STANDS FOR **A**: APPEARANCE/VISUAL FEATURE, **S**: SPATIAL FEATURES [7], **L**: LINGUISTIC FEATURE OF LABEL SEMANTIC EMBEDDINGS, **P**: HUMAN POSE FEATURE.

Method	Backbone	Feature	Default			Known Object		
			Full	Rare	NonRare	Full	Rare	NonRare
Traditional Methods:								
iCAN [7]	ResNet-50	A+S	14.84	10.45	16.15	16.26	11.33	17.73
iHOI [41]	ResNet-50-FPN	A+S	13.39	9.51	14.55	-	-	-
TIN [20]	ResNet-50	A+S+P	17.22	13.51	18.32	19.38	15.38	20.57
DRG [6]	ResNet-50-FPN	A+S+P+L	19.26	17.74	19.71	23.40	21.75	23.89
VCL [12]	ResNet-50	A+S	19.43	16.55	20.29	22.00	19.09	22.87
VSGNet [12]	ResNet-152	A+S	19.80	16.05	20.91	-	-	-
FCMNet [27]	ResNet-50	A+S+P	20.41	17.34	21.56	22.04	18.97	23.12
ACP [17]	ResNet-152	A+S+P	20.59	15.92	21.98	-	-	-
PastaNet [19]	ResNet-50	A+P	22.65	21.17	23.09	24.53	23.00	24.99
ConsNet [28]	ResNet-50-FPN	A+S+L	22.15	17.12	23.65	-	-	-
IDN [18]	ResNet-50	A+S	23.36	22.47	23.63	26.43	25.01	26.85
UnionDet [15]	ResNet-50-FPN	A	17.58	11.72	19.33	19.76	14.68	21.27
IP-Net [40]	Hourglass-104	A	19.56	12.79	21.58	22.05	15.77	23.92
PPDM [22]	Hourglass-104	A	21.73	13.78	24.10	24.58	16.65	26.84
GG-Net [49]	Hourglass-104	A	23.47	16.48	25.60	27.36	20.23	29.48
Transformer-based Methods:								
HoiTransformer [51]	ResNet-50	A	23.46	16.91	25.41	26.15	19.24	28.22
HOTR [16]	ResNet-50	A	25.10	17.34	27.42	-	-	-
AS-Net [4]	ResNet-50	A	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [33]	ResNet-50	A	29.07	21.85	31.23	31.68	24.14	33.93
CDN [45]	ResNet-50	A	31.78	27.55	33.05	34.53	29.73	35.96
PhraseHOI [21]	ResNet-50	A+L	29.29	22.03	31.46	31.97	23.99	34.36
OCN [44]	ResNet-50	A+L	30.91	25.56	32.51	-	-	-
SSRT [14]	ResNet-50	A+L	30.36	25.42	31.83	-	-	-
DisTR [50]	ResNet-50	A	31.75	27.45	33.03	34.50	30.13	35.81
STIP [47]	ResNet-50	A+S+L	32.22	28.15	33.43	35.29	31.43	36.45
GEN-VLKT [23]	ResNet-50	A+L	33.75	29.25	35.10	36.78	32.75	37.99
HODN	ResNet-50	A	33.14	28.54	34.52	35.86	31.18	37.26
HODN + GEN-VLKT	ResNet-50	A+L	**34.56**	**30.26**	**35.84**	**37.86**	**33.93**	**39.03**

KLVT [23] as examples, which are state-of-the-art methods on V-COCO and HICO-Det, respectively. As shown in the last row of Table II, combined with our work, the performance of GEN-VLKT increases by 2.40% and 2.94%, attaining 34.56 mAP and 37.86 mAP under default and known objects settings on HICO-Det. As for STIP in the last row of Table III, STIP can achieve 66.5 mAP and 71.5 role mAP V-COCO, gaining improvements of 0.91% and 1.13%. Both verify that existing methods can benefit easily from our method and achieve new state-of-the-art results.

D. Ablation Study

1) HG-Linking: The specialness of HG-Linking (Section III-B) is to hold one of the HOI relationships, *i.e.*, making Q_H^{out} (human features) as the principal and serving Q_O^{out} (object features) as the auxiliary. To prove its effectiveness, we first verify the difference between the contributions of humans and objects. Particularly, we treat them as the same by passing the addition of Q_H^{out} and Q_O^{out} into the interaction decoder. Here, we adopt two strategies to link: 1) the interaction decoder takes the addition as the positional embeddings (interaction queries) like a vanilla decoder; 2) the interaction decoder introduces a set of randomly initialized learnable positional embeddings and receives the addition as an extra input of the 1st self-attention module. For convenience, we name the first strategy as "addition-guide" and the second as

"random-guide" and report the results in the first row and the second row of Table IV, respectively. We observe a performance decline of more than 1.3% and 3.7%, respectively, indicating the necessity of disentangling the features of humans and objects. The sharper performance drop of the "random-guide" compared to the "addition-guide" also indicates the non-trivial effect of the positional embeddings. Then we verify the superior contribution of humans compared with objects by oppositely treating human and object features, *i.e.*, "object-guide". In the 3rd row of Table IV, with "object-guide", the performance decreases with a margin of 3.8 and 3.9 role mAP, implying the more contributive potentiality of human features. The degradation of performance with replaced link methods proves the effectiveness and verifies that humans and objects have different degrees of contributions to interactions, and the former are more helpful ones.

2) SG-Mechanism: We further conduct the ablation study for SG-Mechanism (Section III-C). We design SG-Mechanism by holding another HOI relationship: interactions will aid human detection but interfere with object detection. To check its effectiveness, we remove SG-Mechanism to allow interaction gradients to optimize the object decoder. The performance gap is shown in the 4th row of Table IV indicates that this mechanism does protect object detection from being disturbed by interactions. As we stated, SG-Mechanism can eliminate the negative impact of interaction on objects, and better object detection can bring more accurate interaction prediction.

Method	Backbone	Feature	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
Traditional Methods:				
iCAN [7]	R50	A+S	45.3	52.4
iHOI [41]	R50-FPN	A+P	45.8	-
TIN [20]	R50	A+S+P	48.7	-
DRG [6]	R50	A+S+P+L	51.0	-
VCL [12]	R50	A+S	48.3	-
VSGNet [12]	R152	A+S	51.8	-
FCMNet [27]	R50	A+S+P	53.1	-
ACP [17]	R152	A+S+P	53.2	-
PastaNet [19]	R50	A+P	51.0	57.5
ConsNet [28]	R50-FPN	A+S+L	53.2	-
IDN [18]	R50	A+S	53.3	60.3
UnionDet [15]	R50-FPN	A	47.5	56.2
IP-Net [40]	HOG104	A	51.0	-
PPDM [22]	HOG104	A	-	-
GG-Net [49]	HOG104	A	54.7	-
Transformer-based Methods:				
HoiTransformer [51]	R50	A	52.9	-
HOTR [16]	R50	A	55.2	64.4
AS-Net [4]	R50	A	53.9	-
QPIC [33]	R50	A	58.8	61.0
CDN [45]	R50	A	62.3	64.4
PhraseHOI [21]	R50	A+L	57.4	-
OCN [44]	R50	A+L	64.2	66.3
SSRT [14]	R50	A+L	63.7	65.9
DisTR [50]	R50	A	66.2	68.5
STIP [47]	R50	A+S+L	66.0	70.7
GEN-VLKT [23]	R50	A+L	62.4	64.5
HODN	R50	A	67.0	69.1
HODN+STIP	R50	A+S+L	**67.5**	**71.9**

#Row	Ablation Item	$AP^{\#1}_{role}$	$AP^{\#2}_{role}$
	HODN	**67.0**	**69.1**
1	human-guide → addition-guide	65.1 (↓ 1.9)	67.2 (↓ 1.9)
2	human-guide → random-guide	62.7 (↓ 4.4)	64.8 (↓ 4.3)
3	human-guide → object-guide	63.3 (↓ 3.8)	65.2 (↓ 3.9)
4	w/o SG-Mechanism	64.7 (↓ 2.3)	66.9 (↓ 2.2)
5	SG-Mechanism to Human	64.3 (↓ 2.7)	66.6 (↓ 2.5)

Moreover, we also apply SG-Mechanism to the detection of humans and show the result in row 5 of Table IV. Note that by adopting SG-Mechanism only for humans, the interaction gradients will backpropagate to the object decoder while not to the human decoder, which means not only the object decoder will be disturbed but also the human decoder can not benefit from interaction prediction. The larger degradation of the role

Fig. 5. Performance comparison based on the different number of parameters. All results are conducted under scenario 1 on the V-COCO test set. The light blue dashed line, representing the performance of the state-of-the-art, OCN [44], with the role mAP of 64.2, is viewed as the baseline to measure against. Our HODN with tiny, small, and base parameter settings is signified by three red stars. And the three red stars locate higher than the light blue dashed line, demonstrating all HODNs outperform previous methods.

mAP (from 64.7 to 64.3 and from 66.9 to 66.6) implies that interactions assist in human detection though slightly. We can conclude that only by adopting SG-Mechanism to objects, as our HODN does, the detection of both humans and objects can achieve the best performance.

E. Parameters vs. Performance

Considering that the disentangling decoders introduce more parameters, we conduct experiments about the performance versus the number of parameters on the V-COCO test set. Here we adopt two other smaller networks, named HODN-tiny (HODN-T) and HODN-small (HODN-S). In particular, two parameter-shared 3-layer decoders for humans and objects, as well as a 3-layer interaction decoder, are included in the HODN-T. And two shared 6-layer decoders and one 3-layer interaction decoder are adopted in the HODN-S. Note that we still utilize independent queries to disentangle the features of humans and objects even when parameters are shared. The comparison result is shown in Figure 5 by reporting the role mAP under scenario 1 since some of the existing works do not support evaluation for scenario 2. As shown in Figure 5, HODN outperforms previous works under various settings of the number of parameters. Even though the improvement compared with OCN [44] is slight, HODN-T introduces much fewer parameters (HODN-T with 39.8M and 41.8M with OCN). The role mAP rises dramatically as the number of parameters increases and reaches an optimal performance at HODN setting. Note that even compared to CDN-L (63.9 mAP in scenario 1), which introduces much more parameters (about 67.0M), our HODN (with about 57.9M parameters) still maintains a significant advantage. We conclude that there is no direct correlation between performance and the number of parameters. Our efficiency comes more from our well-designed framework than from the larger number of parameters.

F. Qualitative Results

The performance of HOI detection relies on the accuracy of instance location and interaction prediction. We argue that

Fig. 6. Visualization samples with ground-truth (left), as well as the detection results of QPIC (middle) and HODN (right). Bounding boxes of humans and objects are drawn with blue and yellow boxes. Interaction categories with confidence are depicted with blue characters. For clearer visualization, we zoom in interaction categories to the left-top corner of images with black characters for correct predictions and red for incorrect ones.

Fig. 7. Visualization of attention maps of decoders in QPIC and HODN with the corresponding ground-truth bounding boxes. The interaction categories are depicted on the left side of the images. As can be seen from the figure, decoders in HODN can capture finer areas according to targets, while QPIC mixes all of them and is biased toward object regions. H-dec: the human decoder, O-dec: the object decoder, I-dec: the interaction decoder.

HODN can detect much more objects, especially occluded ones or overlapping ones. We visualize some examples and use a classical HOI detector QPIC [33] for comparison. On one hand, the location of objects will be disturbed if introducing interaction gradients and that of humans will be improved, as we previously examined. As shown in the left-top images of Figure 6, QPIC locates two "surfboards" in one box, while our HODN does not conflate a bunch of objects. In the left-middle images, QPIC only locates a little part of the "bed", while our HODN can predict the entire "bed" even though most parts of it are occluded by a human. Furthermore, in the left-bottom images, even with only the legs visible, HODN can still locate the man entirely with the help of interactions, whereas QPIC fails. We owe this to SG-Mechanism, which assists in human detection and protects object detection from negative influence by interactions. On the other hand, with information like posture, human features are more vital to interaction prediction. However, object features, albeit helpful for interaction prediction, may introduce bias due to the imbalanced data distribution. Therefore, the prediction of interactions is likely to overfit objects if the network pays too much attention to object features. For example, in the right-top images of Figure 7, QPIC mispredicts the interaction "hold" as "ride" when recognizing the "motorcycle", since "A person is riding a motorcycle" happens with a high probability. A similar situation happens in the right-middle and the right-bottom images of Figure 7, where interaction "lay" has a high correlation with "bed" and "using snowboard" has a strong association with "snowboard", misleading the predictions of QPIC. On the contrary, HODN predicts interactions accurately. We owe this to HG-Linking, which makes the interaction decoder pay more attention to humans and less attention to objects.

We also visualize the attention maps extracted from the last layer of the decoder of QPIC and decoders of HODN in Figure 7. To see more clearly, we also draw the ground truth bounding boxes of humans and objects. Our three decoders can concentrate on their own goal regions, i.e., the human decoder

pays attention to humans, the object decoder to objects, and the interaction decoder to the regions that contribute to understanding actions. For example, in the left-bottom images, the human decoder focuses on the man's head and limbs, while the attention of the interaction decoder becomes more fine-grained after combining object features: the highlight parts become the man's hands and the sports ball. In the right-bottom images, the interaction decoder shifts attention from the head and limbs of the man to the interaction regions where humans and horses come into contact. On the contrary, QPIC uses only one decoder to handle human detection, object detection, and interaction prediction. Consequently, it can not distinguish their difference. As in the visualization of attention maps, QPIC only focuses on object-around regions the most of time. Therefore, the missing objects or incorrect location may cause QPIC to fail to find the correct HOI triplets. This also explains why our method performs much better than it.

V. Conclusion

In this paper, we analyze the relationships among humans, objects, and interactions in two aspects: 1) for interaction, human features make more contributions; 2) for detection, interactive information helps human detection while disturbing object detection. Accordingly, we propose a *Human and Object Disentangling Network* (HODN), a transformer-based framework to explicitly model the relationships, which contains two parallel detection decoders for human and object features, and one interaction decoder for final interactions. A *Human-Guide Linking* method is used by the interaction decoder to make human features dominant and object ones auxiliary. Particularly, human features are sent into the interaction decoder as positional embeddings to make the decoder focus on human-centric regions. Finally, considering that interactive information has the opposite influence on human detection and object detection, we propose *Stop-Gradient Mechanism*, where interaction gradients are not utilized to optimize object detection but human detection. Since our method is orthogonal to the existing methods, they can be easily combined with our method and benefit from it. Extensive experiments conducted on V-COCO and HICO-DET demonstrate that our method brings a significant performance improvement over the state-of-the-art HOI detection methods.

Acknowledgements

References

[1] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," in *Association for the Advancement of Artificial Intelligence*, 2020.

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020.

[3] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proceedings of the IEEE/CVF Workshop on Applications of Computer Vision*, 2018.

[4] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating hoi detection as adaptive set prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[6] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," in *European Conference on Computer Vision*, 2020.

[7] C. Gao, Y. Zou, and J.-B. Huang, "ican: Instance-centric attention network for human-object interaction detection," in *British Machine Vision Conference*, 2018.

[8] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

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

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[12] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *European Conference on Computer Vision*, 2020.

[13] Z. Hou, B. Yu, Y. Qiao, X. Peng, and D. Tao, "Detecting human-object interaction via fabricated compositional learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[14] A. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[15] B. Kim, T. Choi, J. Kang, and H. J. Kim, "Uniondet: Union-level detector towards real-time human-object interaction detection," in *European Conference on Computer Vision*, 2020.

[16] B. Kim, J. Lee, J. Kang, E.-S. Kim, and H. J. Kim, "Hotr: End-to-end human-object interaction detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[17] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in *European Conference on Computer Vision*, 2020.

[18] Y.-L. Li, X. Liu, X. Wu, Y. Li, and C. Lu, "Hoi analysis: Integrating and decomposing human-object interaction," in *Advances in Neural Information Processing Systems*, 2020.

[19] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, H.-S. Fang, Z. Ma, M. Chen, and C. Lu, "Pastanet: Toward human activity knowledge engine," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[20] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[21] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," in *Association for the Advancement of Artificial Intelligence*, 2022.

[22] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "Ppdm: Parallel point detection and matching for real-time human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[23] Y. Liao, A. Zhang, M. Lu, Y. Wang, X. Li, and S. Liu, "Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[25] X. Lin, Q. Zou, and X. Xu, "Action-guided attention mining and relation reasoning network for human-object interaction detection," in *International Joint Conference on Artificial Intelligence*, 2021.

[26] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter-and intra-modality interactions," *IEEE Transactions on Multimedia*, 2020.

[27] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *European Conference on Computer Vision*, 2020.

[28] Y. Liu, J. Yuan, and C. W. Chen, "Consnet: Learning consistency graph for zero-shot human-object interaction detection," in *ACM International Conference on Multimedia*, 2020.

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021.

[31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015.

[32] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[33] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[34] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021.

[35] T. Tsai, A. Stolcke, and M. Slaney, "A study of multimodal addressee detection in human-human-computer interaction," *IEEE Transactions on Multimedia*, 2015.

[36] O. Ulutan, A. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[37] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[38] H. Wang, W.-s. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," in *European Conference on Computer Vision*, 2020.

[39] L. Wang, X. Zhao, Y. Si, L. Cao, and Y. Liu, "Context-associative hierarchical memory model for human activity recognition and prediction," *IEEE Transactions on Multimedia*, 2016.

[40] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[41] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Transactions on Multimedia*, 2019.

[42] W. Xu, Z. Miao, X.-P. Zhang, and Y. Tian, "A hierarchical spatio-temporal model for human activity recognition," *IEEE Transactions on Multimedia*, 2017.

[43] D. Yang and Y. Zou, "A graph-based interactive reasoning for human-object interaction detection," in *International Joint Conference on Artificial Intelligence*, 2020.

[44] H. Yuan, M. Wang, D. Ni, and L. Xu, "Detecting human-object interactions with object-guided cross-modal calibrated semantics," in *Association for the Advancement of Artificial Intelligence*, 2022.

[45] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, and X. Li, "Mining the benefits of two-stage and one-stage hoi detection," in *Advances in Neural Information Processing Systems*, 2021.

[46] H. Zhang, S. Wan, W. Guo, P. Jin, and M. Zheng, "Hod: Human-object decoupling network for hoi detection," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 2219–2224.

[47] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Exploring structure-aware transformer over interaction proposals for human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[48] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *European Conference on Computer Vision*, 2020.

[49] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[50] D. Zhou, Z. Liu, J. Wang, L. Wang, T. Hu, E. Ding, and J. Wang, "Human-object interaction detection via disentangled transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[51] C. Zou, B. Wang, Y. Hu, J. Liu, Q. Wu, Y. Zhao, B. Li, C. Zhang, C. Zhang, Y. Wei *et al.*, "End-to-end human object interaction detection with hoi transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.