TED-Net: Dispersal Attention for Perceiving Interaction Region in Indirectly-Contact HOI Detection

Yuxiao Wang, Qi Liu, *Senior Member, IEEE*, and Yu Lei

***Abstract*— Human-Object Interaction (HOI) detection is a fertile research ground that merits further investigation in computer vision, and plays an important role in image high-level semantic information understanding. To achieve superior object detection performance, existing HOI models predominantly concentrate on the corresponding bounding box information of humans and objects, respectively, and ignore their surrounding information, thus it results in imprecise inference of instance interaction, which is severe for indirectly-contact interaction images (Intersection-over-Union = 0). To address that, a novel Triple stream Enhanced encoder-decoder Dispersal Network (TED-Net), equipped with human, object, and instance interaction decoding streams, is proposed to decouple instances' relationships. Meanwhile, we design a dispersal attention mechanism to capture indirectly-contact interaction information and an auxiliary discrimination mechanism to improve the ability of instance interaction decoding stream for action category recognition. Experimental results show that the proposed TED-Net achieves the best performance among HOI models using the ResNet-50 backbone on the (big) HICO-Det dataset and comes third on the (small) V-COCO dataset in leaderboard. Additionally, two indirectly-contact interaction datasets, namely, HICO-Det-IC and V-COCO-IC, are constructed to demonstrate the usefulness and effectiveness of our TED-Net in interacting between indirectly-contact instances, with an average of +3.80 mAP on HICO-Det-IC and +5.46 mAP on V-COCO-IC. Code is available at https://drliuqi.github.io/.**

***Index Terms*— Human-object interaction, object detection, computer vision.**

I. Introduction

HUMAN-OBJECT interaction (HOI) detection is an activity understanding of estimating the interaction between a human and an object in an image. To enable computers to better understand human activities, Gupta and Malik [2] introduced the groundbreaking task of "Visual Semantic Role Labeling" for the first time. The emergence of this new task and its corresponding datasets not only facilitate improved scene understanding by computers but also contribute to a comprehensive visual understanding of human activities. Subsequently, the task of HOI detection [3] emerges. Given an image, the HOI task needs to find out the position of human-object in the image and to judge the action type of human-object. Specifically, the output of HOI detection system is a series of triples (⟨"human", "object", "action"⟩), which describes the person, the object, and the interaction type in an image, respectively. For example, ⟨"person", "bicycle", "ride"⟩ indicates that the person rides a bicycle. The HOI detection task not only locates the instances in an image, but also detects the relationship between instances. Research on HOI detection offers valuable insights for other advanced visual tasks, such as scene graph generation [4], action recognition [5], [6], video understanding [7], and so on.

Current neural network-based HOI detection approaches can be roughly grouped into two categories: two-stage pipeline [8], [9], [10], [11], [12], [13], [14] versus one-stage solution [1], [15], [16], [17], [18], [19]. Two-stage pipeline first utilizes pre-trained detection framework [3], [20], [21], [22] to detect instances (viz., humans and objects) in images and are then paired one-to-one in the interactive discrimination stage to recognize the action category. Two-stage pipeline model is relatively simple and easy to understand, and yet runs slowly due to the one-to-one pairing operation. Therefore, optimizing network models to reduce cost consumption is also one of the important research directions [23], [24].

One-stage architecture has shown its advantages over two-stage pipeline counterpart in HOI detection, which directly outputs the HOI triplet. Previous one-stage models usually require to define new interaction points or bounding boxes between humans and objects to match the real relationship pairs. Although one-to-one pairing operation is not required and the running time is significantly improved as compared to two-stage pipeline models, those pre-defined interaction points or bounding boxes are not always valid when encountering different practical situations.

To tackle this problem, a query-based one-stage HOI detection method [25] is proposed, using a transformer instead of interaction points or bounding boxes to extract interaction information. On the basis of that, one-stage architecture

Manuscript received 11 October 2023; revised 8 December 2023; accepted 22 January 2024. Date of publication 26 January 2024; date of current version 3 July 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in part by the Fundamental Research Funds for the Central Universities under Grant 2023ZYGXZR085, in part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1674, and in part by the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004. This article was recommended by Associate Editor R. Du. *(Corresponding author: Qi Liu.)*

Yuxiao Wang and Qi Liu are with the School of Future Technology, South China University of Technology, Guangzhou 511400, China (e-mail: ftwangyuxiao@mail.scut.edu.cn; drliuqi@scut.edu.cn).

Yu Lei is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611730, China (e-mail: leiyu1117@my.swjtu.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3358952

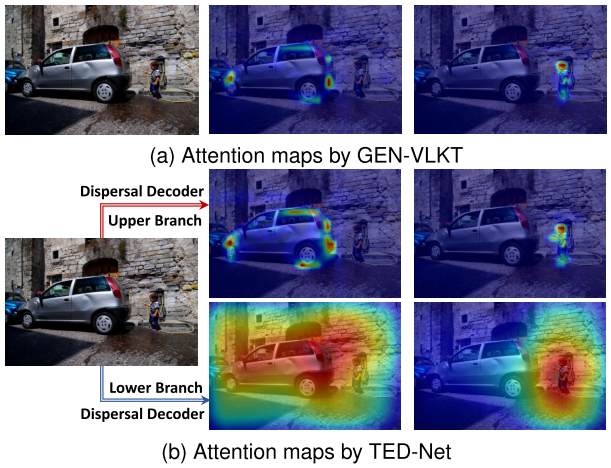

Fig. 1. TED-Net is compared with the GEN-VLKT via attention map when perceiving interaction regions. (a) The attention maps of existing models (like GEN-VLKT [1]) from human and object decoders, which include (partial) positional boundary information and ignore interaction region (viz., a water pipe) between human and object. (b) The attention maps of TED-Net are shown from upper and lower branches of the dispersal decoder. The same with the existing HOI detectors, the upper branch aims to capture the positional boundary information, while the proposed lower branch focuses on the surrounding information including human, object, and their interaction area.

has two solutions, namely, single stream networks [26], [27], [28] and dual stream networks [1], [29]. The former applies a decoder stream to simultaneously predict human, object bounding boxes, and object, action categories. Different from the single stream-based networks, the latter exploits the instance stream to detect the bounding boxes and instance categories, and the interaction stream to predict the action. However, it is challenging to double-decode humans and objects for each query in images containing rich HOI triples. Furthermore, the instance stream decoder typically learns only the boundary information of the instances. For indirectly-contact interaction images, i.e., Intersection over Union (IoU) $= 0$, it performs not well for the existing models to detect action categories in the interaction stream using the information provided by the instance stream since the instance stream only captures boundary information and ignores the surrounding information of humans and objects, as shown in Figure 1a.

In this work, we propose a Triple stream Enhanced encoder-decoder Dispersal Network, shorted as TED-Net, to decouple HOI pairs and capture surrounding information around humans and objects. Our motivation is shown as follows. The judgment of the action category between human and object depends on the human body, object state, and their surrounding environment. The interaction information of contact images (IoU > 0) mainly lies in the boundary box and the interior of human and object, while it mostly depends on the surrounding environment information for indirectly-contact images (IoU $= 0$) (e.g., the waterpipe in Figure 1b). Our TED-Net has significantly improved the existing methods on HICO-Det and V-COCO datasets. Specifically, our TED-Net has achieved a 0.73 mAP gain on the HICO-Det dataset, a 3.80 mAP gain on the HICO-Det-IC dataset, and a 5.46 mAP promotion on the V-COCO-IC dataset on average compared with the previous state-of-the-art (SOTA) method GEN-VLKT [1].

II. Related Works

Deep learning achieves advanced success in various fields, such as image re-ranking [30], image recognition [31], [32], [33], object detection [20], and image segmentation [34]. Current neural network-based HOI detection approaches can be roughly grouped into two categories: two-stage pipeline versus one-stage solution.

Two-stage pipeline. Gupta and Malik [2] first proposed the visual semantic role labeling task, which associates objects in a scene with different semantic roles for each action. Traditional HOI detection models first applied the pre-trained object detectors [20], [21] to detect humans and objects, and then associated with the human-object pairs to estimate their HOI types, e.g., iCAN [35] and HO-RCNN [3]. The instance-centric attention network (iCAN), proposed by Gao et al. [35], used appearance information as a cue to highlight areas relevant to each detected human or object. However, relying solely on the appearance of humans and objects neglects the contextual information, which is not conducive to HOI detection. Therefore, Wang et al. proposed a contextual attention framework to learn contextually-aware appearance features [36]. Additionally, the reference [11] designed a human-centered interactive inference framework to utilize the instances' context information.

Unlike previous works, Xu et al. explored how human behavior in various scenarios could be detected and recognized by inferring human intent [37]. Then FCMNet [38], FG [39], PDNet [40], GTNet [41], and ConsNet [42] made use of text information to extract semantic information, leading to further improvements. Moreover, some works, including [43], [44], [45], and [46], were proposed to focus on the region of interest (RoI) with the assistance of spatial and body posture information.

While the HOI methods mentioned above are advantageous for feature expression, they impose a substantial workload. Afterward, two-stage graph-based models were proposed [44], [47], [48] and achieved satisfactory performance. Taking the GPNN [49] as an example, nodes and edges were exploited to represent instances and interactions, respectively, and the corresponding adjacency matrix was developed to build the relationship structure. SG2HOI [50] embedded scene graph information as the key contextual cues and developed a relation-aware message-passing module to get relationship information for interaction. Nevertheless, two-stage methods suffer from the cost of computation due to the one-to-one pairing operation.

One-stage solution. Different from two-stage methods, one-stage models realized the direct conversion from the input image to the HOI triplet [1], [16], [17], [51], [52], [53], [54]. PPDM [15] proposed the midpoint of human and object being as the interaction center and redefined the HOI triplet, which has the advantages of computation acceleration and accuracy improvement. Similar to PPDM using the same midpoint definition, IP-Net [55], however, cannot achieve

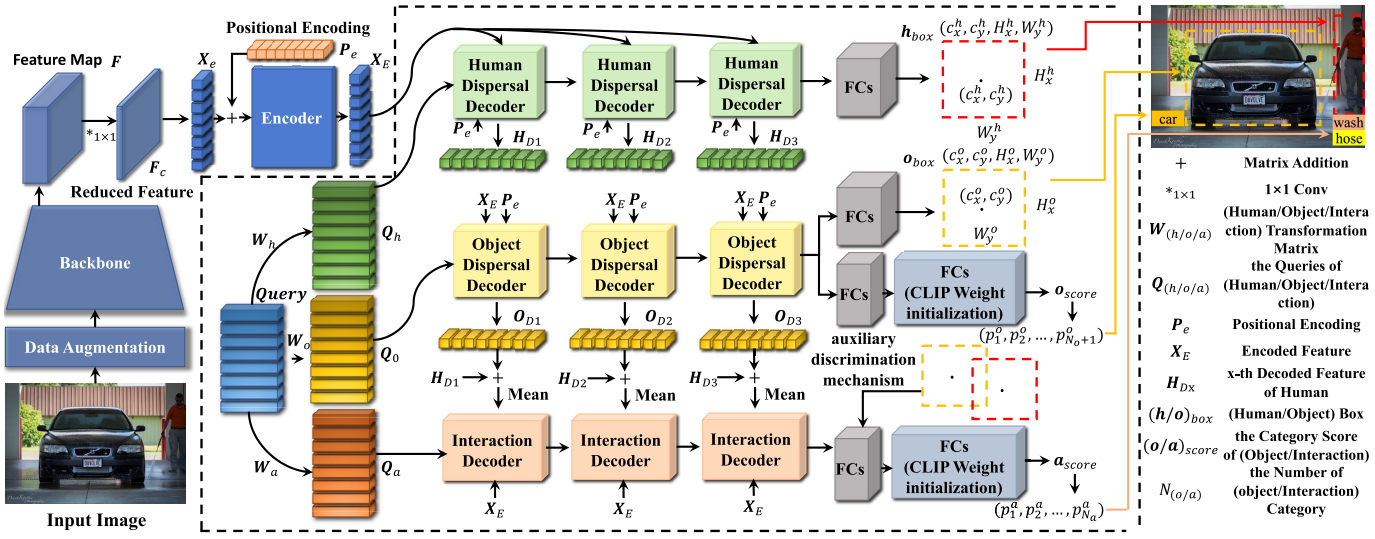

Fig. 2. Overall architecture of the proposed TED-Net. After performing data augmentation, the input image is fed to the ResNet-50 backbone for feature extraction. The resulting feature map is compressed and transmitted to the encoder for global feature encoding. Three query embeddings Q_h,Q_o,Q_a are obtained from the decoding process. To enhance interaction understanding, CLIP is finally used to initialize the classifier.

convinced results under different circumstances [26]. To that end, the query-based transformer was applied for HOI detection, such as HOITrans [26] and QPIC [27], where they queried the RoI through the query embedding to obtain the decoder's output. To address the drawbacks of query-driven HOI detectors with costly post-matching, GEN-VLKT [1] proposed a guided-embedding network to build a two-branch pipeline without post-matching. Besides, it integrated Contrastive Language-Image Pre-Training (CLIP) [56] to enhance interaction understanding.

III. Methods

In this section, the proposed TED-Net is developed and we present a detailed introduction of our architecture, as shown in Figure 2.

Backbone. We follow the ResNet-based detectors to adopt a ResNet-50 [57] as the feature extractor, and the resulting feature map is $\boldsymbol{F} \in \mathbb{R}^{H_r \times W_r \times C_r}$. Given an input image $\boldsymbol{x} \in \mathbb{R}^{H_i \times W_i \times C_i}$, H_i, W_i, C_i represent its' height, width and RGB channel, respectively. ResNet-50 performs 32 times down-sampling through multi-layer convolution, where $H_r = 1/32H_i$, $W_r = 1/32W_i$, and $C_r = 2048$. The 1×1 convolution operation is applied to compress $\boldsymbol{F}$ with the result of $\boldsymbol{F_C} \in \mathbb{R}^{H_r \times W_r \times C_c}$ at $C_c = 256$.

Encoder. As depicted in Figure 3, the encoder is composed of N stacked sub-encoders (in this work, N=6). Each sub-encoder consists of a multi-head self-attention layer, a feed-forward layer, and two add & normalize layers. The encoder is fed with two inputs: $\boldsymbol{X}_e$ and positional encoding $\boldsymbol{P}_e$. The feature map $\boldsymbol{F_C}$ is expanded at the pixel level to obtain $\boldsymbol{X}_e \in \mathbb{R}^{(H_r \times W_r) \times C_c}$ that consists of $H_r \times W_r$ tokens of length C_c. $\boldsymbol{X}_e$ with positional embedding $\boldsymbol{P}_e \in \mathbb{R}^{(H_r \times W_r) \times C_c}$ is fed to the encoder to output $\boldsymbol{X}_E$ [25], that is:

$$\boldsymbol{X}_E = Encoder(\boldsymbol{X}_e, \boldsymbol{P}_e). \tag{1}$$

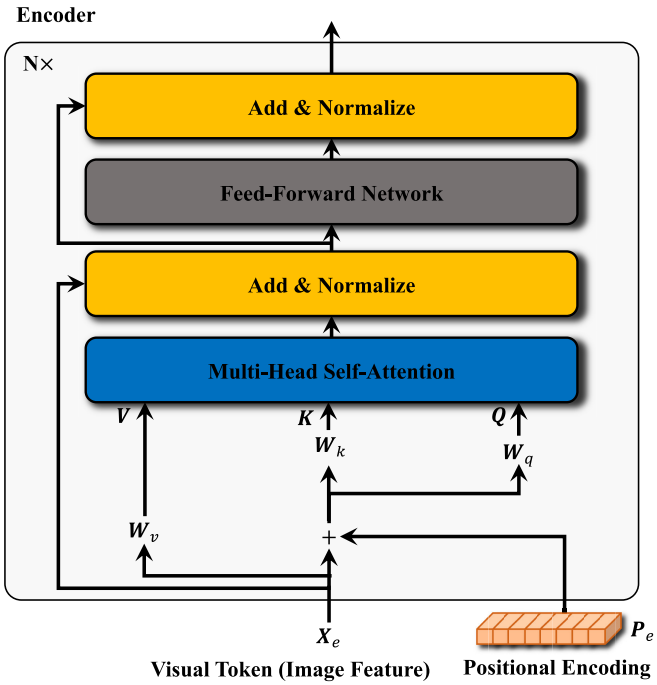

Fig. 3. Introduction of the encoder. The input features $\boldsymbol{X}_e$ are multiplied by $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, and $\boldsymbol{W}_v$ to obtain the matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. They are fed into a multi-head attention mechanism. Afterward, the obtained attention features are added to $\boldsymbol{X}_e$ and normalized.

The encoding process in Eq. 1 is as follows: $\boldsymbol{X}_e$ and $\boldsymbol{P}_e$ are first fed into the multi-head self-attention and add & normalize modules to obtain the feature $\boldsymbol{Z}$, that is:

$$\boldsymbol{Z} = LayerNorm(\boldsymbol{X}_e + MHSA(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})), \tag{2}$$

where $LayerNorm$ is used to normalize the activation values for each layer, and $MHSA$ is a multi-head self-attention module. Moreover, the process of calculating the attention feature

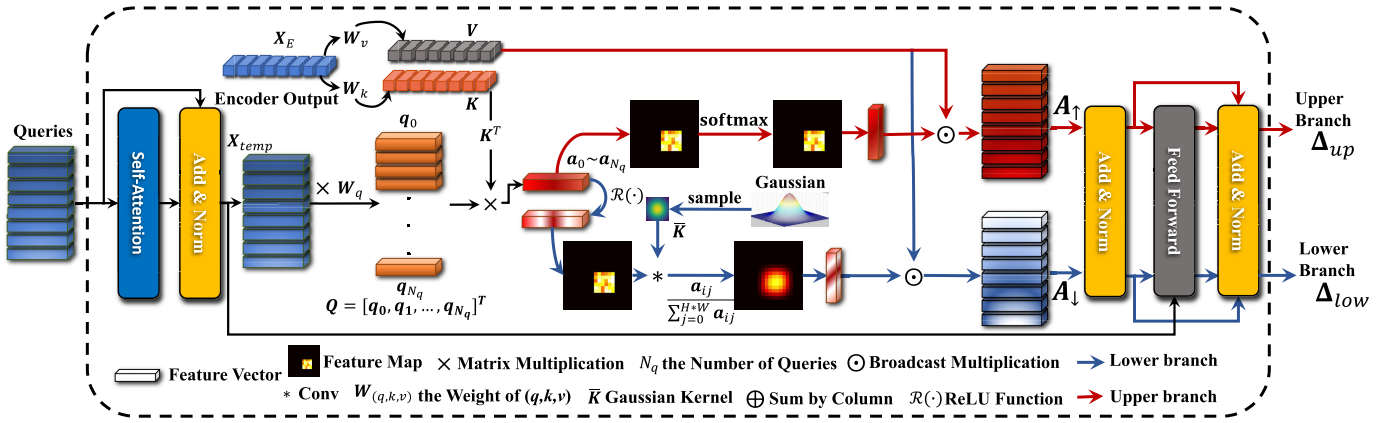

Fig. 4. Details of the dispersal decoder. The upper branch (red line) is the conventional transformer used to extract instance information, while the lower one is the combination of Gaussian convolution and conventional transformer for perceiving surrounding information. $\boldsymbol{W}_k$, $\boldsymbol{W}_v$, and $\boldsymbol{W}_q$ are the weight matrices of transformer $\boldsymbol{K}$, $\boldsymbol{V}$, and $\boldsymbol{Q}$, respectively. Each vector of $\boldsymbol{V}$ (or $\boldsymbol{K}$, or $\boldsymbol{Q}$) is computed by $\boldsymbol{X}_E \times \boldsymbol{W}_{v_i}$ (or $\boldsymbol{X}_E \times \boldsymbol{W}_{k_i}$, or $\boldsymbol{X}_{temp} \times \boldsymbol{W}_{q_i}$).

map $\boldsymbol{A}$ in the *MHSA* layer is formulated as:

$$\boldsymbol{A} = Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\boldsymbol{V}), \quad (3)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ denote the query, key, and value of the input tokens, respectively. $\boldsymbol{Q}\boldsymbol{K}^T$ represents the attention matrix, and d_k represents the dimension of $\boldsymbol{K}$.

Then we extract the global contextual features via:

$$\boldsymbol{X}_E = LayerNorm(\mathbf{Z} + Feed\text{-}Forward(\mathbf{Z})), \quad (4)$$

where $Feed\text{-}Forward$ consists of two fully connected layers.

Query generation. In our TED-Net, three query matrices are generated from a common query matrix to represent humans, objects, and interactions, respectively. To be specific, each input $\boldsymbol{query} \in \mathbb{R}^{N_q \times C_q}$ matrix is multiplied by the corresponding query transformation matrix $\boldsymbol{W}_s \in \mathbb{R}^{C_q \times C_q}$, namely:

$$\boldsymbol{Q}_s = \boldsymbol{query} \times \boldsymbol{W}_s, \quad s \in h, o, a. \quad (5)$$

It will simultaneously predict N_q humans, objects and action categories based on $\boldsymbol{Q}_h \in \mathbb{R}^{N_q \times C_q}$, $\boldsymbol{Q}_o \in \mathbb{R}^{N_q \times C_q}$ and $\boldsymbol{Q}_a \in \mathbb{R}^{N_q \times C_q}$, where N_q and C_q denote the number and dimension of the query, respectively. On the basis of that, $\boldsymbol{Q}_h$, $\boldsymbol{Q}_o$ and $\boldsymbol{Q}_a$ will be automatically linked since each of them comes from the same query.

Dispersal decoder. The two-branch dispersal decoder is designed to enhance the attention of instances' surrounding information, where the upper branch is the conventional transformer, and the lower one is the proposed combination of Gaussian convolution and conventional transformer. As shown in Figure 4, the upper branch (red line) is focused on the boundary information to detect the human-object bounding boxes, while the lower branch (blue line) is to capture the surrounding information for HOI. For ease of readability, $\boldsymbol{\Delta}_{up}$ and $\boldsymbol{\Delta}_{low}$ are used to denote the upper and the lower branches, respectively. Taking the human dispersal decoder as an example, the attention feature map $\boldsymbol{A}_\uparrow$ (red cube) is computed by Eq. 3 in $\boldsymbol{\Delta}_{up}$, and then the output of dispersal decoder is generated. In lower branch $\boldsymbol{\Delta}_{low}$, the attention features $\boldsymbol{a}_i (i = 0, 1, \ldots, N_q)$ from $\boldsymbol{\Delta}_{up}$ are activated by the ReLU function. A convolution kernel $\bar{\boldsymbol{K}} \in \mathbb{R}^{n \times n}$ is built by a series of sampling values from two-dimensional Gaussian distribution ($\mathcal{N}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, where $\mu_1 = 0$, $\mu_2 = 0$, and $\rho = 0$). That is:

$$\bar{\boldsymbol{K}}(x, y) = (2\pi\sigma_1\sigma_2)^{-1} \, exp^{-\frac{1}{2}\left(\frac{x}{\sigma_1^2} + \frac{y}{\sigma_2^2}\right)}, \quad (6)$$

which is acted on feature $\boldsymbol{a}_i$ to expand the region area. The next step is to prevent excessive value via performing normalization operation:

$$\boldsymbol{a}_i = \frac{\boldsymbol{a}_{ij}}{\sum_{j=0}^{H \times W} \boldsymbol{a}_{ij}}, j = 1, 2, \ldots, H \times W, \quad (7)$$

in which H and W represent the height and width of the input feature map, respectively. Therefore, we obtain the dispersal feature map $\boldsymbol{A}_\downarrow$ and the output $\boldsymbol{\Delta}_{low}$ of the decoder.

Interaction decoder. As shown in Figure 5, the query embedding $\boldsymbol{Q}_a$, the output $\boldsymbol{H\_O}$ of dispersal decoder, and the encoded feature tokens $\boldsymbol{X}_E$ are input to the interaction decoder, where $\boldsymbol{H\_O}$ is computed by

$$\boldsymbol{H\_O} = \frac{(\boldsymbol{H}_D + \boldsymbol{O}_D)}{2}, \quad (8)$$

which $\boldsymbol{H}_D$ and $\boldsymbol{O}_D$ are the outputs of human dispersal decoder and object dispersal decoder, respectively. The HOI will be classified by the interaction decoder and the auxiliary discrimination mechanism.

Auxiliary discrimination mechanism. The inspiration behind ADM (Auxiliary Discrimination Mechanism) stems from intuitive observations: certain HOI actions exhibit conditional constraints. Take, for instance, the scenario of a human riding a bicycle; the positioning necessitates the human to be above the bicycle, rather than the bicycle being positioned above the human. Similarly, in an action like throwing a frisbee, the bounding boxes of the human and the frisbee should not intersect while the frisbee is in flight. Moreover, it is expected that the human bounding box would encompass a significantly larger area compared to that of the frisbee bounding box. These conceptual insights are visually represented in Figure 6.

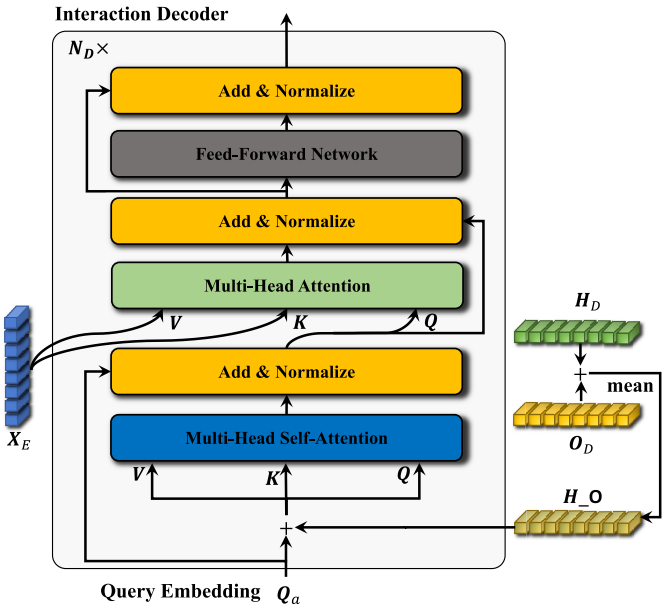

Fig. 5. Introduction of the interaction decoder. $\boldsymbol{H}_D$ and $\boldsymbol{O}_D$ represent the outputs of human dispersal decoder and object dispersal decoder, respectively. The sum of $\boldsymbol{H}_D$ and $\boldsymbol{O}_D$, denoted as $\boldsymbol{H\_O}$, is obtained. After adding $\boldsymbol{H\_O}$ to the query embedding $\boldsymbol{Q}_a$, they are transformed by $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$. Then they are input into a multi-head self-attention mechanism. The output of the multi-head self-attention mechanism is added to $\boldsymbol{Q}_a$ and normalized. $\boldsymbol{X}_E$ represents the output of the encoder. It combines with the normalized result to generate new $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ matrices.

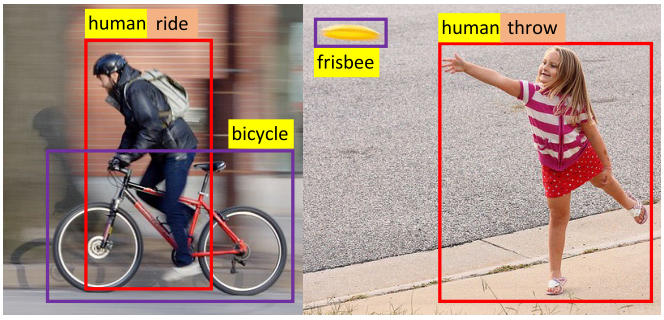

Fig. 6. Examples of the auxiliary discrimination mechanism in HOI scenarios. The auxiliary discrimination mechanism considers specific conditions for HOI actions. For example, in a scene where the human is riding a bicycle, the human should be positioned above the bicycle rather than the bicycle being above the person. Similarly, like throwing a frisbee, the bounding boxes of the human and the frisbee should not intersect.

Specifically, consider two interactive examples: ⟨"human", "bicycle", "ride"⟩ and ⟨"human", "frisbee", "throw"⟩, the former describes the interaction information as "human is riding above the bicycle", and the IoU must be greater than 0. For the latter without direct contact between the girl and the frisbee, its IoU equals 0. As we all know, the object category also affects the action category as well, such as "human-ride/hold-bicycle" instead of "human-throw/eat-bicycle".

Therefore, an auxiliary discrimination mechanism is designed. The output of the human decoding stream is $\boldsymbol{h}_{box} = (c_x^h, c_y^h, H_x^h, W_y^h)$, where c_x^h and c_y^h denote the coordinates of the midpoint of human box, H_x^h and W_y^h represent the human box's height and width, respectively. The output of object decoding stream is defined as $\boldsymbol{o}_{box} = (c_x^o, c_y^o, H_x^o, W_y^o)$. The $\boldsymbol{o}_{score} \in \mathbb{R}^{(N_o+1)}$ is the prediction score of each category, and $N_o + 1$ is the number of object categories. Then the relative position $\boldsymbol{r}_p$, the direction of object to human $\boldsymbol{d}_{o,h}$, the IoU of the human-object bounding boxes $iou_{h,o}$, the area ratio of the human-object bounding boxes $area_{ratio}$, and the object category information $\boldsymbol{o}_{class}$ are, respectively, computed by

$$\boldsymbol{r}_p = (c_x^h - c_x^o, c_y^h - c_y^o), \tag{9}$$

$$\boldsymbol{d}_{o,h} = \begin{cases} (-1, 1), & c_x^h - c_x^o < 0, c_y^h - c_y^o > 0, \\ (1, 1), & c_x^h - c_x^o > 0, c_y^h - c_y^o > 0, \\ (1, -1), & c_x^h - c_x^o > 0, c_y^h - c_y^o < 0, \\ (-1, -1,) & c_x^h - c_x^o < 0, c_y^h - c_y^o < 0, \end{cases} \tag{10}$$

$$iou_{h,o} = \frac{\boldsymbol{h}_{box} \cap \boldsymbol{o}_{box}}{\boldsymbol{h}_{box} \cup \boldsymbol{o}_{box}}, \tag{11}$$

$$area_{ratio} = \frac{H_x^h \times W_y^h}{H_x^o \times W_y^o + \varepsilon}, \tag{12}$$

$$\boldsymbol{o}_{class} = one\_hot(\boldsymbol{o}_{score}). \tag{13}$$

Then these metrics are concatenated with FCs (Fully Connected) layer to connect with CLIP weight.

FCs and CLIP. FCs from the first two rows in Figure 2 are utilized to predict bounding boxes of humans and objects, while the rest FCs are used for object category and action class classification. The CLIP model is not directly involved in the proposed pipeline. We feed pre-trained CLIP with 600 categories from HICO-DET dataset (or 274 categories from V-COCO dataset) and use its weights (denoted as CLIP weight Initialization in Fig. 2) to connect the former FCs.

Loss function. As the number of decoder's query embeddings is $N_q = 64$, our TED-Net can predict 64 interaction pairs simultaneously. We follow the query-based approaches using the Hungarian algorithm [58] to assign a unique prediction with each ground truth. The loss function is formulated as:

$$\mathcal{L}_{match\_loss} = \sum_{i}^{N_q} \mathcal{L}_{match}(g^i, \hat{y}^i), \tag{14}$$

$$\mathcal{L}_{match}(g^i, \hat{y}^i) = \beta_1 \sum_{p \in o,a} \alpha_p \mathcal{L}_{cls}^p + \beta_2 \sum_{q \in h,o} \alpha_q \mathcal{L}_{box}^q + \beta_3 \sum_{r \in h,o} \alpha_r \mathcal{L}_{iou}^r, \tag{15}$$

where g is the ground truth, and $\hat{y}$ denotes the predicted interaction pair. $\mathcal{L}_{cls}^p$, $\mathcal{L}_{box}^q$, $\mathcal{L}_{iou}^r$ represent the classification loss, the regression loss of box, the IoU loss of human-object bounding boxes, respectively. α and β are hyper-parameters of matching loss to achieve a trade-off between classification, regression, and IoU. The overall loss is given by

$$\mathcal{L}_{total} = \beta_1 \sum_{p \in o,a} \alpha_p \mathcal{L}_{cls}^p + \beta_2 \sum_{q \in h,o} \alpha_q \mathcal{L}_{box}^q + \beta_3 \sum_{r \in h,o} \alpha_r \mathcal{L}_{iou}^r + \beta_4 \mathcal{L}_{glo}, \tag{16}$$

in which $\mathcal{L}_{glo}$ is the distillation loss of the CLIP model.

TABLE I
PERFORMANCE COMPARISONS ON HICO-DET DATASET

Method	Detector	Backbone	Default (**mAP**↑)			Know Object (**mAP**↑)		
			Full	Rare	Non-Rare	Full	Rare	Non-Rare
Two-stage Methods:								
PMFNet	COCO	ResNet-50-FPN	17.46	15.65	18.00	20.34	17.47	21.20
DRG	COCO	ResNet-50-FPN	19.26	17.74	19.71	23.40	21.75	23.89
VCL	COCO	ResNet-50	19.43	16.55	20.29	22.00	19.09	22.87
VSGNe	COCO	Res-DCN-152	19.80	16.05	20.91	-	-	-
FCMNet	COCO	ResNet-50	20.41	17.34	21.56	22.04	18.97	23.12
ACP	COCO	Res-DCN-152	20.59	15.92	21.98	-	-	-
PD-Net	COCO	ResNet-152-FPN	20.81	15.90	22.28	24.78	18.88	26.54
SG2HOI	COCO	ResNet-50	20.93	18.24	21.78	24.83	20.52	25.32
DJ-RN	COCO	ResNet-50	21.34	18.53	22.18	23.69	20.64	24.60
SCG	COCO	ResNet-50-FPN	21.85	18.11	22.97	-	-	-
IDN	COCO	ResNet-50	23.36	22.47	23.63	26.43	25.01	26.85
ATL	HICO-Det	ResNet-50	23.81	17.43	25.72	27.38	22.09	28.96
One-stage Methods:								
PPDM-Hourglass	HICO-Det	Hourglass-104	21.94	13.97	24.32	24.81	17.09	27.12
HOI-Trans	HICO-Det	ResNet-50	23.46	16.91	25.41	26.15	19.24	28.22
HOTR	HICO-Det	ResNet-50	25.10	17.34	27.42	-	-	-
AS-Net	HICO-Det	ResNet-50	28.87	24.25	30.25	31.74	27.07	33.14
QPIC	HICO-Det	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
FGAHOI	-	Swin-Tiny	29.94	22.24	32.24	32.48	24.16	34.97
CDN-S	HICO-Det	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42
CDN-B	HICO-Det	ResNet-50	31.78	27.55	33.05	34.53	29.73	35.96
PR-Net	HICO-Det	ResNet-50	31.17	25.66	32.82	-	-	-
Iwin-B	HICO-Det	ResNet-50-FPN	32.03	27.62	34.14	35.17	28.79	35.91
ERNet-L	HICO-Det	EfficientNetV2-L	32.94	27.86	34.45	-	-	-
GEN-VLKT	HICO-Det	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
GEN-VLKT-OF	HICO-Det	ResNet-50	33.51	28.52	35.00	36.58	32.14	37.91
TED-Net	HICO-Det	ResNet-50	**34.00**	**29.88**	**35.24**	**37.13**	**33.63**	**38.18**

* -OF indicates the optimal result provided by themselves in Github.

IV. RESULTS

A. Experimental Settings

Datasets. Two commonly-used benchmark datasets, namely, HICO-Det [3] and V-COCO [2], are applied to evaluate the performance of our TED-Net. HICO-Det totally has 47,776 images (38,118 images for training and 9,658 images for testing) [3]. These images have 80 and 117 categories of objects and actions, respectively, and 600 types of different interactions, where interactions are divided into 462 no-rare and 138 rare categories. V-COCO [2] contains a total of 10326 images from coco data [59], including 5400 images for training sets and 4964 images for test sets. In addition, V-COCO has 80 object categories, 29 action categories, and 4 body motions without interaction with any objects.

Evaluation metric. The mean Average Precision (mAP) is used as an evaluation metric. A true positive result for HOI triplet prediction is defined when satisfying: 1) IoUs of the predicted bounding box of human and object and their corresponding ground-truth are greater than 0.5; and 2) the predicted categories are correct. For V-COCO dataset, there are two mAP metrics: S1 for the 29 action categories with 4 body motions and S2 for the 25 action categories without no-object HOI categories [1].

Parameter Details. The ResNet-50 serves as the foundational backbone for feature extraction, employing a channel compression of 256. The encoder consists of 6 layers, while each of the decoders dedicated to human, object, and instance interactions encompasses 3 layers. The intermediary size of the feed-forward layers within the transformer blocks is set at 2048. The dimension of the embeddings (transformer's input dimension) is specified as 256. The attention mechanism within the transformer employs 8 attention heads. The query embedding count, denoted as N_q, is established at 64.

The optimization of our network is achieved using the AdamW optimizer during the training phase, with a learning rate of 1.5×10^{-4} (the backbone's learning rate being 1.0×10^{-5}) and a weight decay of 1×10^{-4}. The training process extends for 120 epochs. All experimental trials are executed on 2 Tesla A40 (48G) GPUs, employing a batch size of 16. The computational environment runs Ubuntu 22.04, with Python version 3.8, PyTorch version 1.7.1, torchvision version 0.8.2, and CUDA version 11.0.

B. Effectiveness for Regular HOI Detection

To verify the effectiveness of our TED-Net, experiments have been conducted as compared to the existing methods on the HICO-Det and V-COCO datasets, and the results are tabulated in Tables I and II. From Table I, we observe that our TED-Net achieves the best performance among all HOI detectors, which especially outperforms the SOTA GEN-VLKT-OF method by 0.49 mAP, 1.36 mAP, and 0.24 mAP for Full, Rare and No-Rare default setting, and 0.55 mAP, 1.49 mAP, and 0.27 mAP for Know Object setting, respectively. As can be seen from Table II, TED-Net achieves superior results on the V-COCO dataset, by +1.00 mAP and +0.50 mAP on two settings, respectively, as compared to GEN-VLKT.

Fig. 7. HOI detection samples. The first two rows illustrate the attention maps of our TED-Net. It is evident that the upper branch of the dispersal decoder effectively focuses on the positions of instances, while the lower branch is utilized to perceive interactions, e.g., reins and car washing tools.

TABLE II
PERFORMANCE COMPARISONS ON V-COCO DATASET

Method	Anchor	AP^{S1}_{role} (**mAP**↑)	AP^{S2}_{role} (**mAP**↑)
Two-stage Methods:			
GPNN	×	44.0	-
iCAN	×	45.3	52.4
TIN	×	47.8	54.2
VCL	×	48.3	-
DRG	×	51.0	-
IP-Net	×	51.0	-
VSGNet	×	51.8	57.0
PMFNet	×	52.0	-
PD-Net	×	52.6	-
FCMNet	×	53.1	-
ACP	×	53.23	-
IDN	×	53.3	60.3
One-stage Methods:			
HOI-Trans	*Q*	52.9	-
GG-Net	*P*	54.7	-
HOTR	*Q*	55.2	64.4
QPIC	*Q*	58.8	61.0
Iwin-B	*Q*	60.47	-
FGAHOI	*Q*	60.5	61.2
Iwin-L	*Q*	60.85	-
PR-Net	*Q*	61.4	-
ERNet-L	*Q*	61.6	-
CND-B	*Q*	62.29	64.42
GEN-VLKT	*Q*	62.41	64.46
TED-Net	*Q*	**63.41**	**64.96**

TABLE III
PREFORMANCE COMPARISONS ON HICO-DET-IC AND V-COCO-IC DATASETS

HICO-Det-IC			
	Default (**mAP**↑)		
Method	Full	Rare	Non-Rare
GEN-VLKT	27.25	23.73	28.03
TED-Net	**30.09**	**30.09**	**30.24**
V-COCO-IC			
	AP^{S1}_{role} (**mAP**↑)		AP^{S2}_{role} (**mAP**↑)
GEN-VLKT	33.59		32.35
TED-Net	**38.71**		**38.15**

On the other hand, two indirectly-contact interaction datasets, namely, HICO-Det-IC and V-COCO-IC, are constructed from HICO-Det and V-COCO datasets, respectively, in order to test our model's usefulness and effectiveness in detecting indirectly-contact interaction instances. HICO-Det-IC and V-COCO-IC datasets are built by keeping those images with IoU =0 and deleting others with IoU>0 from HICO-Det and V-COCO datasets, respectively. In the HICO-Det-IC (or V-COCO-IC) dataset, there are a total of 3571 (or 291) images and 9720 (or 1300) pairs of interactions in the training set, and 829 (or 306) images and 2686 (or 1445) pairs of interactions in the test set. We have conducted experiments on the HICO-Det-IC and V-COCO-IC datasets, where the results have tabulated in Table III. For fair comparison, it is worth noting that we re-train our model on these two new datasets without using any prior parameters as that on HICO-Det or V-COCO datasets. Compared to GEN-VLKT, ours shows its superior performance in detecting indirectly-contact interactions by +2.84 mAP, +6.36 mAP, and +2.21 mAP in HICO-Det-IC from the viewpoints of Full, Rare, and No-Rare default settings, and by +5.12 mAP and +5.80 mAP in V-COCO-IC dataset from the viewpoints of AP1 and AP2.

C. Ablation Study

The ablation study is conducted on the V-COCO and V-COCO-IC datasets to investigate the effect of each module in detail. The results on V-COCO are presented in Table IV. +T, +D, and +A denote the triple stream frame, dispersal attention mechanism, and auxiliary discrimination mechanism, respectively. $+D_{x\times y}$ means the use of a $x \times y$ size Gaussian convolution kernel in dispersal decoder. We take the SOTA GEN-VLKT as the baseline model (denoted as B). From Table IV, we see that both B+T and B+A achieve higher AP1 than the baseline. To verify the effectiveness of the dispersal attention mechanism, it is first equipped with the triple stream frame on the baseline model, that is B+T+D, where the sizes of D are 5×5, 7×7, and 9×9, respectively. The experimental results show that all performances have boosted to some extent, and the best result is at $B+T+D_{7\times7}$, by +0.88 mAP in AP1 and +0.33 mAP in AP2. In addition, the auxiliary discrimination mechanism integrates into the baseline model under the framework of triple stream frame, i.e.,

Fig. 8. HOI detection samples in HICO-Det. The red rectangle represents a human, the blue rectangle represents an object, and the green line indicates the line connecting their centers. The first row shows different interactions with the same object, such as ⟨"motorcycle", "sit on"⟩, ⟨"motorcycle", "ride"⟩, ⟨"motorcycle", "jump"⟩. The second row demonstrates different objects involved in the same interaction, such as ⟨"bottle", "hold"⟩, ⟨"baseball_bat", "hold"⟩, ⟨"laptop", "hold"⟩. The third row illustrates the recognition capability of TED-Net for no interactions between a human and an object.

TABLE IV
ABLATION STUDY ON V-COCO DATASET

Method	AP^{S1}_{role} (**mAP**↑)	AP^{S2}_{role} (**mAP**↑)
Baseline (GEN-VLKT)	62.41	64.46
B+T	62.73	64.11
B+D	/	/
B+A	62.79	64.17
B+T+$D_{5\times5}$	63.02	64.41
B+T+$D_{7\times7}$	63.29	64.79
B+T+$D_{9\times9}$	62.93	64.32
B+T+A	62.91	64.43
B+T+D+A(TED-Net)	**63.41**	**64.96**

* +T, +D, and +A denote the triple stream frame, dispersal attention mechanism, and auxiliary discrimination mechanism, respectively. +$D_{x\times y}$ means the use of a $x \times y$ size Gaussian convolution kernel in dispersal decoder..

TABLE V
ABLATION STUDY ON V-COCO-IC DATASET

Method	AP^{S1}_{role} (**mAP**↑)	AP^{S2}_{role} (**mAP**↑)
Baseline(GEN-VLKT)	33.59	32.35
B+T	34.65	33.44
B+D	/	/
B+A	35.28	33.89
B+T+$D_{7\times7}$	37.15	36.85
B+T+A	36.12	34.66
B+T+D+A(TED-Net)	**38.71**	**38.15**

TABLE VI
PERFORMANCE COMPARISONS OF IMPROBABLE INTERACTION ON HICO-DET DATASET

Method	**mAP**↑
Baseline (GEN-VLKT)	25.35
B+T	26.02
B+D*	**27.50**
B+A	25.58

* +D* indicates the use of the dispersal decoder. (The dispersal decoder cannot be used independently, so we conduct experiments based on the three-stream frame.)

B+T+A, causing 0.5 mAP improvement in AP1 and enjoying comparable performance in AP2. Our TED-Net reported as "B+T+D+A" in Table IV has achieved the best result, and a 1.00 mAP gain and 0.50 mAP gain over the baseline in AP1 and AP2, respectively. Similar results on V-COCO-IC dataset have been received in Table V. Among them, the B+T+D outperforms B+T by +2.5 mAP in S1 mode and +3.41 mAP in S2 mode. Additionally, B+T+D performs better than B+A by +1.87 mAP in S1 mode and +2.96 mAP in S2 mode.

We conduct four experiments to verify which component of TED-Net contributes the most to detect non-interactive human-object pairs (in the last row of Figure 8), and the results are shown in Table VI. The baseline is the GEN-VLKT model, +T indicates the use of three-stream frame, +D* indicates the use of dispersal decoder, and +A indicates the use of auxiliary discriminative mechanism. It can be observed that +D* (dispersal decoder) of TED-Net contributes the most to the improbable interaction detection.

Whether to use dispersal decoder and the choice of different backbones are two factors to affect performance. Therefore, we further conduct additional ablation studies on HICO-Det

Fig. 9. HOI detection samples in V-COCO. The first two rows describe the same action with different objects, such as ⟨"baseball bat", "hold"⟩, ⟨"toothbrush", "hold"⟩, ⟨"skis", "hold"⟩. The last row describes different real-life scenarios where the object and action categories are the same.

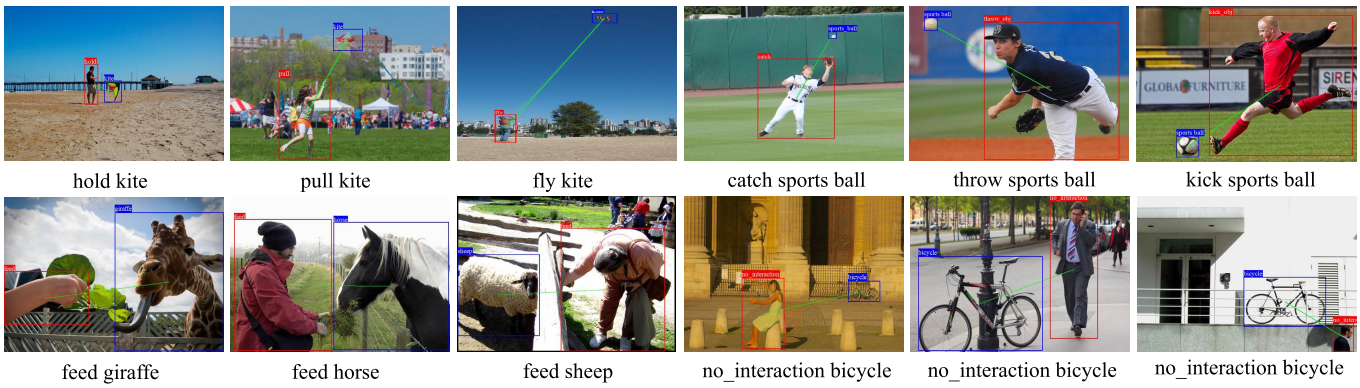

Fig. 10. Indirectly-contact HOI detection samples. The first row demonstrates TED-Net can detect the same object with different interactive actions, such as the "kite". For the same interactive action, TED-Net can cleverly distinguish different interactive objects (the first three images of the second row). The last three images in the second row showcase TED-Net's ability to accurately identify unlikely interactive behaviors.

TABLE VII
PERFORMANCE OF DISPERSAL DECODER ON V-COCO DATASET

Human dispersal decoder	Object dispersal decoder	AP^{S1}_{role} (**mAP**↑)	AP^{S2}_{role} (**mAP**↑)
✓		63.13	64.61
	✓	63.02	64.55
✓	✓	**63.41**	**64.96**

TABLE VIII
PERFORMANCE OF VARIOUS BACKBONES IN HICO-DET DATASET

Backbone	TED-Net parameters	Default (**mAP**↑)		
		Full	Rare	Non-Rare
Swin-Tiny	51M	31.18	28.35	32.01
Swin-Small	72M	**34.65**	**31.07**	**35.72**
ResNet-50	47M	34.00	29.88	35.24

* M represents million.

and V-COCO datasets, as shown in Table VII and Table VIII. It can be observed from Table VII that the optimal performance is achieved when the dispersal red decoder is applied to both human and object decoders. This is mainly because regions of HOI happen around both the human and the object, rather than being limited to either the human's or the object's surroundings.

To evaluate the effects of different backbones, a new vision transformer, called Swin Transformer [60], has been applied, and results are tabulated in Table VIII. Due to limitations on computing resources, viz., GPUs, we run the Swin-Tiny and Swin-Small versions. It can be observed that the proposed method with ResNet-50 as the backbone is superior to the Swin-Tiny based one in terms of parameters and mAP, and achieves comparable performance with fewer parameters compared to Swin-Small based one. As well, we observe that the proposed method performs better when the backbone changes to larger Swin Transformer. It is envisioned that interest of reasearchers can further improve our work when owning sufficient GPUs.

Fig. 11. Multiple interaction detection samples on HICO-Det. In examples 1, 2, 3, and 8, TED-Net successfully detects interactions between different individuals under the same object. Other examples showcase TED-Net's outstanding performance in detecting interactive actions between multiple humans and different objects.

TABLE IX
PERFORMANCE COMPARISONS ON HICO-DET-IC AND V-COCO DATASETS

Method	HICO-Det-IC (**mAP**↑)			V-COCO-IC (**mAP**↑)		Efficiency	
	Full	Rare	Non-Rare	AP^{S1}_{role}	AP^{S2}_{role}	#Params(↓)	FPS(↑)
GEN-VLKT	27.25	23.73	28.03	33.59	32.35	**42M**	**31.01**
TED-Net	**30.09**	**30.09**	**30.24**	**38.71**	**38.15**	47M	29.23

D. Model Complexity

The comparative analysis of performance between TED-Net and the prevailing state-of-the-art models on the HICO-Det-IC and V-COCO-IC datasets is presented in Table IX. Our approach, TED-Net, introduces a minor increase in model parameters, which is relatively modest when juxtaposed with the size of the GEN-VLKT. The operational speed, quantified in terms of FPS (Frames Per Second), experiences only a marginal decrease. Nonetheless, TED-Net demonstrates remarkable advancement in results, yielding an average enhancement of +3.80 mAP on HICO-Det-IC and +5.46 mAP on V-COCO-IC.

E. Visualization

The attention visualization results of TED-Net are shown in Figure 7. The visualizations reveal that the upper branch decoder of TED-Net is primarily used for localizing humans and objects, and the lower branch focuses on perceiving the surrounding interaction information, such as the reins in the first row image and the washing tool in the second row image. Additionally, Figure 1b also showcases the attention visualization results of TED-Net.

Figure 8 depicts the visualized outcomes generated by the TED-Net model on the HICO-Det dataset. It is noteworthy that, for convenience, only the interaction pair with the highest score from each input image has been retained. For instances involving multiple interaction pairs, please consult Figures 11 and 12. Upon analyzing an image, TED-Net classifies and detects the entities of "human", "action", and "object". Human bounding boxes are highlighted in red, object bounding boxes in blue, and the connection between their centers is indicated by a green line in Figure 8.

In Figure 8, the images in the first row exemplifies different interactions between a human and the same object. This demonstrates TED-Net's ability to effectively detect distinct interactions involving the same object. The second row portrays interaction detection outcomes for diverse objects

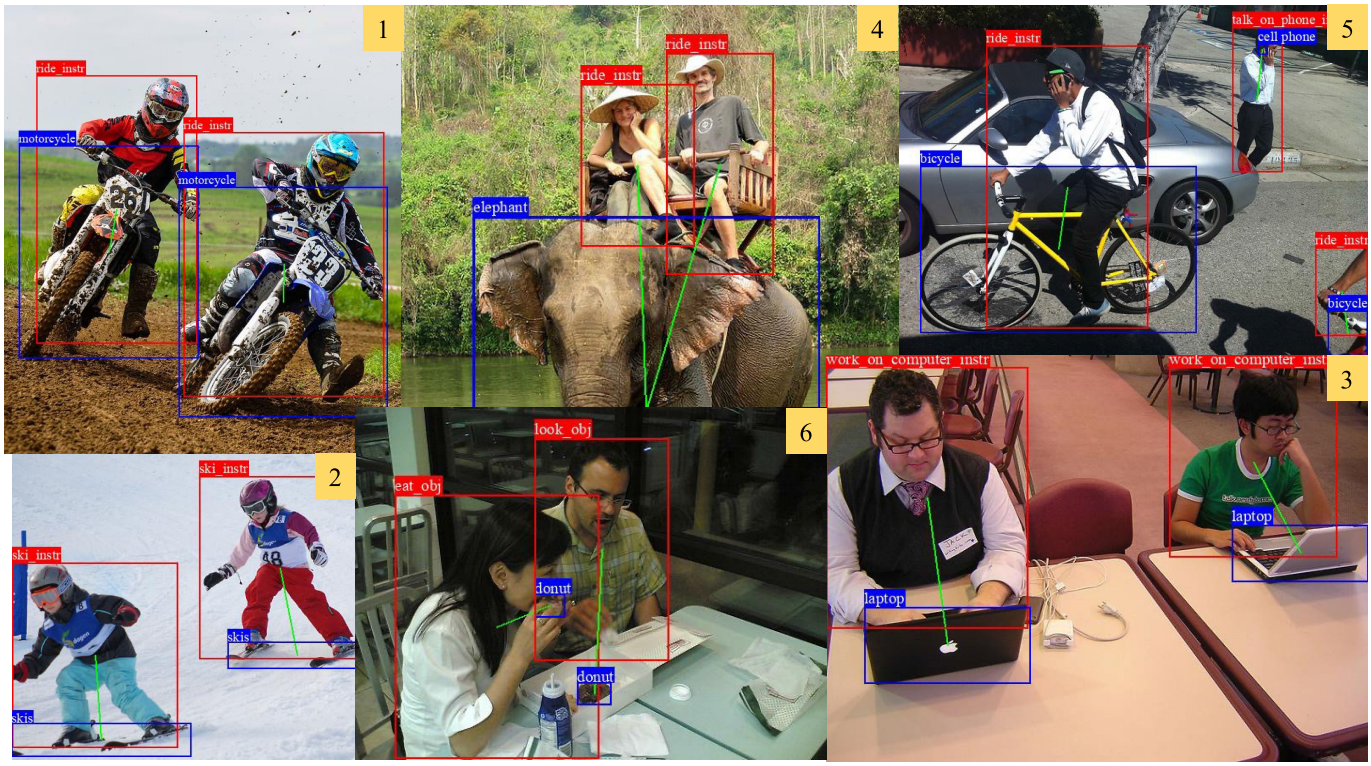

Fig. 12. Multiple interaction detection samples on V-COCO. Examples 1, 2, 3, and 4 illustrate scenarios where multiple humans interact with different objects (examples 1, 2, and 3) or a single object (example 4). On the other hand, Examples 5 and 6 vividly depict situations where multiple humans interact with different objects.

partaking in the same action ("hold"). Importantly, TED-Net excels in identifying human-object pairs that are improbable participants in an interaction, indicated by the "no interaction" action in various scenarios (last row).

Figure 9 showcases TED-Net's testing performance on the V-COCO dataset. Analogous to Figure 8, the visualizations present the detection outcomes for various actions, such as "hold", "lay", and others. Figure 10 showcases the visual outcomes produced by TED-Net when applied to indirectly-contacting images. To elaborate, the first three images of the first row in Figure 10 highlights TED-Net's success in not only identifying the "fly kite" interaction but also discerning rare interaction categories like "hold kite" and "pull kite" for the entity "kite". This accomplishment stems from TED-Net's holistic consideration of both self-related and surrounding environmental factors, which contributes to more precise predictions.

Furthermore, the last three images of the first row in Figure 10 exemplifies TED-Net's capacity to distinguish between the actions "catch sports ball", "throw sports ball", and "kick sports ball". This distinction is achieved by integrating diverse action-specific details and contextual cues present within the human form. In the second row of Figure 10, TED-Net adeptly determines the absence of interaction between the human and the bicycle. This showcases TED-Net's proficiency in recognizing scenarios where interactions are unlikely to occur.

Figures 11 and 12 provide schematic representations of multi-interaction detection within an image. In Figure 11, instances 1, 2, and 3 offer illustrations of interactions involving multiple humans with the same object. Furthermore, TED-Net demonstrates the capability to discern identical interactions occurring between multiple humans and distinct objects (instances 4, 5, and 6). The examples also encompass situations with multiple HOI within the same context, as depicted by actions like "ride a boat" and "sit on boat" in instance 8.

Turning to Figure 12, instances 1, 2, 3, and 4 exhibit scenarios where multiple humans engage with different objects (examples 1, 2, and 3) or a single object (example 4). Instances 5 and 6 portray divergent interactions involving various objects.

V. Conclusion

To enhance the capability of existing HOI detection methods for predicting indirectly-contact interactions, we propose a novel HOI detector, namely, TED-Net, consisting of human, object, and instance interaction decoding streams. For capturing the surrounding interaction information in indirectly-contact images, a dispersal attention mechanism is designed to act on the human and object decoding streams. For interaction understanding, an auxiliary discrimination mechanism is developed to assist in predicting HOI. Furthermore, two indirectly-contact HOI detection datasets, referred to as HICO-Det-IC and V-COCO-IC, are built for investigating the capability of TED-Net in detecting indirectly-contact interactions. Experimental results have shown the superior performance of TED-Net as compared to other competitors.

Specifically, in HICO-Det-IC and V-COCO-IC datasets, TED-Net achieves a remarkable improvement by +3.8 mAP and +5.46 mAP on average, respectively, as compared to the SOTA baseline using ResNet-50.

Conflict of Interest Statement

The authors declare no competing financial interests or personal relationships that could have influenced the work presented in this article.

References

[1] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20091–20100.

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

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

[4] Z. Fu et al., "DRAKE: Deep pair-wise relation alignment for knowledge-enhanced multimodal scene graph generation in social media posts," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3199–3213, Jul. 2023.

[5] H. Wang, B. Yu, J. Li, L. Zhang, and D. Chen, "Multi-stream interaction networks for human action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3050–3060, May 2022.

[6] H. Fan, T. Zhuo, X. Yu, Y. Yang, and M. Kankanhalli, "Understanding atomic hand-object interaction with human intention," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 275–285, Jan. 2022.

[7] N. Wang et al., "Exploring spatio-temporal graph convolution for video-based human-object interaction recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5814–5827, Oct. 2023.

[8] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "DRG: Dual relation graph for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 696–712.

[9] Z. Hou, X. Peng, Y. Qiao, and D. Tao, "Visual compositional learning for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 584–600.

[10] D.-J. Kim, X. Sun, J. Choi, S. Lin, and I. S. Kweon, "Detecting human-object interactions with action co-occurrence priors," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 718–736.

[11] K. Xu et al., "Effective actor-centric human-object interaction detection," *Image Vis. Comput.*, vol. 121, May 2022, Art. no. 104422.

[12] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13299–13307.

[13] Y. Ito, "HOKEM: Human and object keypoint-based extension module for human-object interaction detection," 2023, *arXiv:2306.14260*.

[14] B. Wan, Y. Liu, D. Zhou, T. Tuytelaars, and X. He, "Weakly-supervised hoi detection via prior-guided bi-level representation learning," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023, pp. 1–17.

[15] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "PPDM: Parallel point detection and matching for real-time human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 482–490.

[16] J. Lim, V. M. Baskaran, J. M. Lim, K. Wong, J. See, and M. Tistarelli, "ERNet: An efficient and reliable human-object interaction detection network," *IEEE Trans. Image Process.*, vol. 32, pp. 964–979, 2023.

[17] H. Peng et al., "Parallel reasoning network for human-object interaction detection," 2023, *arXiv:2301.03510*.

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

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

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[21] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[22] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.

[23] Q. Liu and X. Li, "Efficient low-rank matrix factorization based on ℓ_1, ε-norm for online background subtraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4900–4904, Jul. 2022.

[24] Q. Liu, X. Li, H. Cao, and Y. Wu, "From simulated to visual data: A robust low-rank tensor completion approach using lp-regression for outlier resistance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3462–3474, Jun. 2022.

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

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

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

[28] Z. Li, C. Zou, Y. Zhao, B. Li, and S. Zhong, "Improving human-object interaction detection via phrase learning and label composition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1509–1517.

[29] Z. Hou, B. Yu, and D. Tao, "Discovering human-object interaction concepts via self-compositional learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 461–478.

[30] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 2019–2032, May 2014.

[31] J. Zhang, J. Yang, J. Yu, and J. Fan, "Semisupervised image classification by mutual learning of multiple self-supervised models," *Int. J. Intell. Syst.*, vol. 37, no. 5, pp. 3117–3141, May 2022.

[32] J. Zhang, Y. Cao, and Q. Wu, "Vector of locally and adaptively aggregated descriptors for image feature representation," *Pattern Recognit.*, vol. 116, Aug. 2021, Art. no. 107952.

[33] J. Yu, M. Tan, H. Zhang, Y. Rui, and D. Tao, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 563–578, Feb. 2022.

[34] J. Chen, W. Lu, Y. Li, L. Shen, and J. Duan, "Adversarial learning of object-aware activation map for weakly-supervised semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3935–3946, Aug. 2023.

[35] C. Gao, Y. Zou, and J.-B. Huang, "ICAN: Instance-centric attention network for human-object interaction detection," 2018, *arXiv:1808.10437*.

[36] T. Wang et al., "Deep contextual attention for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5694–5702.

[37] B. Xu, J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Interact as you intend: Intention-driven human-object interaction detection," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1423–1432, Jun. 2020.

[38] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 248–265.

[39] A. Bansal, S. S. Rambhatla, A. Shrivastava, and R. Chellappa, "Detecting human-object interactions via functional generalization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10460–10469.

[40] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 69–85.

[41] A. S. M. Iftekhar, S. Kumar, R. A. McEver, S. You, and B. S. Manjunath, "GTNet: Guided transformer network for detecting human-object interactions," 2021, *arXiv:2108.00596*.

[42] Y. Liu, J. Yuan, and C. W. Chen, "ConsNet: Learning consistency graph for zero-shot human-object interaction detection," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4235–4243.

[43] Y.-L. Li et al., "Detailed 2D-3D joint representation for human-object interaction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10166–10175.

[44] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13617–13626.

[45] Y.-L. Li et al., "PaStaNet: Toward human activity knowledge engine," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 382–391.
[46] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9468–9477.
[47] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 843–851.
[48] D. Yang and Y. Zou, "A graph-based interactive reasoning for human-object interaction detection," 2020, *arXiv:2007.06925*.
[49] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 401–417.
[50] T. He, L. Gao, J. Song, and Y.-F. Li, "Exploiting scene graphs for human-object interaction detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15984–15993.
[51] X. Zhong, X. Qu, C. Ding, and D. Tao, "Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13229–13238.
[52] B. Kim, T. Choi, J. Kang, and H. J. Kim, "UnionDet: Union-level detector towards real-time human-object interaction detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 498–514.
[53] H.-S. Fang, Y. Xie, D. Shao, and C. Lu, "DIRV: Dense interaction region voting for end-to-end human-object interaction detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1291–1299.
[54] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, and C. Qian, "Reformulating HOI detection as adaptive set prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9004–9013.
[55] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4116–4125.
[56] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[58] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, nos. 1–2, pp. 83–97, Mar. 1955.
[59] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 740–755.
[60] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

Yuxiao Wang is currently pursuing the Ph.D. degree with the School of Future Technology, South China University of Technology (SCUT), China. His research interests include human-object interaction, including human-object interaction detection, human-object contact detection, semantic segmentation, and crowd counting. In addition, he has conducted research on weakly supervised learning methods.

Qi Liu (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the City University of Hong Kong, Hong Kong, China, in 2019. From 2018 to 2019, he was a Visiting Scholar with the University of California at Davis, CA, USA. From 2019 to 2022, he was a Research Fellow with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore. He is currently a Professor with the School of Future Technology, South China University of Technology. His research interests include human-object interaction, AIGC, 3D scene reconstruction, and affective computing. He was a recipient of the Best Paper Award of IEEE ICSIDP in 2019. Since 2022, he has been an Associate Editor of IEEE SYSTEMS JOURNAL and *Digital Signal Processing*. He was a Guest Editor of IEEE INTERNET OF THINGS JOURNAL and *IET Signal Processing*.

Yu Lei is currently pursuing the Ph.D. degree with the School of Information Science and Technology, Southwest Jiaotong University (SWJTU), China. Her research interests include crowd counting, semantic segmentation, intelligent transportation, and gesture recognition.