Pattern Recognition 146 (2024) 110021

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Parallel disentangling network for human–object interaction detection

Yamin Cheng [a], Hancong Duan [a,*], Chen Wang [a], Zhijun Chen [b]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[b] SKLSDE Lab, School of Computer Science and Engineering, Beihang University, China

ARTICLE INFO

Keywords:
Human–object interaction detection
Transformer

ABSTRACT

Human–object interaction (HOI) detection aims to localize and classify triplets of human, object and interaction from a given image. Earlier two-stage methods suffer both from mutually independent training processes and the interference of redundant negative human–object pairs. Prevailing one-stage transformer-based methods are free from the above problems by tackling HOI in an end-to-end manner. However, one-stage transformer-based methods carry the unnecessary entanglements of the query for different tasks, i.e., human–object detection and interaction classification, and thus bring in poor performance. In this paper, we propose a new transformer-based approach that parallelly disentangles human–object detection and interaction classification in a triplet-wise manner. To make each query focus on one specific task clearly, we exhaustively disentangle HOI by parallelly expanding the naive query in vanilla transformer as triple explicit queries. Then, we introduce a semantic communication layer to preserve the consistent semantic association of each HOI through mixing the feature representations of each query triplet of the correspondence constraint. Extensive experiments demonstrate that our proposed framework outperforms the existing methods and achieves the state-of-the-art performance, with significant reduction in parameters and FLOPs.

1. Introduction

Human–Object Interaction (HOI) detection, formally defined as the task to predict a set of triplets ⟨Human, Interaction, Object⟩ within a static image, can be decoupled into two subtasks: detecting human–object pairs and classifying the interaction between them. Recognizing the interaction of human–object pairs can help a machine detailedly understand high-level human-centric scenes. For instance, the left input image in Fig. 1(a) describes "A person is kicking a football, and another persons have no interaction with the football", rather than "There are three men and a football". Besides being beneficial to understanding human behavior in complex real-world environments, it also contributes to many downstream tasks, such as activity analysis [1–3], image understanding [4,5], image retrieval [6], etc.

Most existing HOI detection methods fall into two paradigms: two-stage methods and one-stage methods. Two-stage methods [7–9] first localize humans and objects by off-the-shelf object detectors [10,11] and then feed the features of the human–object pairs, which are generated by matching humans and objects one by one, into an interaction classifier. Despite the encouraging results, two-stage methods often suffer from the interference of redundant negative human-object pairs and the sub-optimal solution resulting from the independent optimization on two subtasks. To alleviate these problems, the one-stage approaches directly treat HOI detection as multi-task learning and focus on localizing instances and inferring interaction in parallel [12,13], sequentially [14] or in a unified manner [15], allowing their learning processes to benefit from each other. However, one-stage transformer-based methods usually struggle to make a good trade-off on multi-task learning for two subtasks, since detection primarily tends to focus on the edge of the local region and interaction classification needs contextually essential cues. For instance, as depicted in Fig. 1(a), the unified method [15] that simultaneously detects HOI learns one type of naive and entangled query representations that directly but vaguely aggregate all the features of location and classification, weakening the efficiency of the method. To address this issue, as shown in Figs. 1(b) and (c), the parallel [12] and cascade [14] methods try to break HOI detection into two separated branches or two sequential steps, and efficiently strengthen the performance of the HOI detection. But these one-stage methods also make the pipeline more complex and costly in computation or take a non-exhaustive breaking strategy—that human–object (HO) pairs entangle as ever, or both. Nevertheless, this elegant property, disentangling HOI detection, still brings a lot of conveniences that each branch (or step) can focus on its task and produce good results.

In this paper, we move one step further, take the essence and discard the dregs from previous one-stage methods. To this end, we devise

* Corresponding author.
E-mail address: duanhancong@uestc.edu.cn (H. Duan).

https://doi.org/10.1016/j.patcog.2023.110021
Received 24 April 2022; Received in revised form 29 August 2023; Accepted 28 September 2023
Available online 4 October 2023

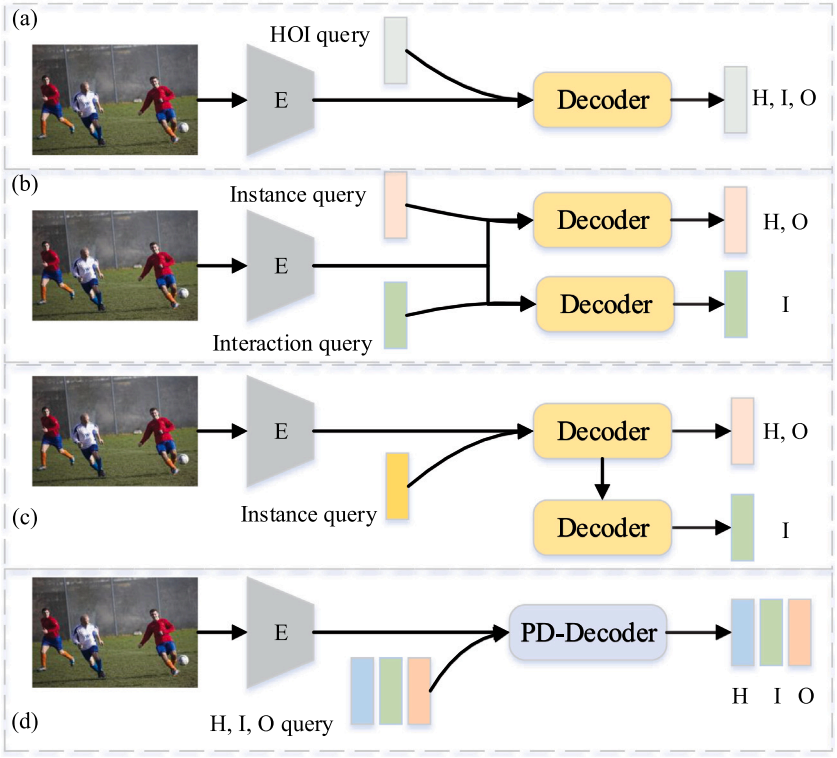


Fig. 1. Comparison of one-stage approaches on HOI detection. 'E' refers to the visual feature extractor. The resulting H, I and O are utilized to predict instances (human and object) or classify the interaction. (a) The unified method, typically uses one naive query to predict all HOI results. (b) The parallel method, generally decouples the decoder as two branches. (c) The cascade method, directly disentangles detection and interaction classification in a cascade manner. (d) Our proposed method, parallelly disentangles HOI in a triplet-wise manner.

a Parallel Disentangling Network, dubbed PDN, that exhaustively disassembles human–object detection and interaction classification in a triplet-wise manner. The proposed method keeps the advantages of unified one-stage methods, directly locating the interactive human–object pairs and interaction, and brings the advantages of parallel and cascade one-stage methods into our framework, disentangling HOI detection. As shown in Fig. 1(d), we design a parallel disentangling decoder, namely PD-Decoder, based on the one-stage transformer-based paradigm. In PD-Decoder, we instantiate the concept of 'query' as triple expanding queries: human queries, object queries and interaction queries, which stay the triplet-wise. These expanding queries can help eliminate unnecessary entanglements attributed to one naive query and focus on triple explicit semantic features, dedicated to detection or interaction classification, respectively. However, this also causes a core problem, i.e., how to preserve the consistent semantic association of each HOI. To address this problem, we design a semantic communication layer. The semantic communication layer couples the identical index query from three types of queries to formulate the query triplet set with the corresponding constraint (or called the intra-triplet set. The intra-triplet set is the combination of three queries for one HOI, where one element is a query), then interchanges the representation information within each intra-triplet set, and enables the inter-triplet sets mutually independent (the inter-triplet set is the combination of intra-triplet sets for different HOIs), since the feature representations aggregated by each intra-triplet set are dedicated to detecting one HOI.

In summary, we make the following key contributions:

- We present a new approach that parallelly disentangles human–object detection and interaction classification in a triplet-wise manner, which is beneficial for learning explicit semantic representations;
- We introduce triple expanding queries to learn respective semantic features directly, and design a semantic communication layer to preserve the consistent semantics association of each HOI;
- We conduct an extensive evaluation of our proposed approach on two datasets, V-COCO and HICO-DET, showing that PDN outperforms previous state-of-the-art methods and achieves relative performance gains of 3.46% and 1.25% respectively, with substantial reduction in parameters (↓ 13.6%) and FLOPs (↓ 6.7%).

2. Related work

Here we briefly review the prior literature in two related fields: transformer and human–object interaction detection,

2.1. Transformer

Recently, Transformer [16] has drawn much attention to AI community due to its strong capability of modeling long-range relation. In natural language processing (NLP), Transformers have so far been the de-facto architecture and driven the advancement of various sequence prediction tasks, e.g., language modeling and machine translation. Inspired by the tremendous success of transformer architectures in the field of NLP, many researchers attempt to further propel the transformer wave in computer vision (CV) tasks, e.g., object detection [17], action recognition [18], image captioning [19], visual ground [20]. In detail, the seminal work DETR [17] formulates object detection as a set prediction problem. It introduces a small set of learnable object queries, reasons global context and object relations with attention mechanism, and outputs the final set of predictions in parallel. Subsequently, several concurrent works [21,22] distinctly promote the model performance by mending positional ambiguity, slow training convergence and high computation complexity derived from DETR itself, respectively. Another line of transformer model, ViT families [23,24], uniformly partitions an input image into a sequence of non-overlapping image patches and applies a pure transformer directly to sequences of image patches to classify the input image. Later, many follow-up works

further upgrade the ViT architecture for not only better visual recognition but other high-level vision tasks, such as object detection [25] and semantic segmentation [26]. Vision Transformer has convincingly shown its strong potential as an alternative to the previously dominant convolutional neural networks.

Unlike most aforementioned works, we mainly focus on the general architecture variants and track HOI detection task.

2.2. Human–object interaction detection

With computer vision witnessing a steady momentum of breakthroughs in recent years, many researchers not only focus on detecting individual persons and objects in isolation but also identify instances and interactions between them in parallel, i.e., HOI, since it enhances a more comprehensive understanding of what is happening in the scene. In a nutshell, mainstream algorithms generally encode an image or a local image region to obtain their vision representations using Convolutional Neural Networks (CNNs) [27,28], Graph Neural Networks (GNNs) [29,30] and Transformer [16], and decode the representations into interaction prediction. As mentioned in the introduction, most of existing HOI methods can be mainly divided into two categories: two-stage and one-stage methods.

Two-stage HOI methods. Current two-stage detectors [7,31,32], detecting instances first and predicting interaction based on the detected instances, can be regarded as the instance-driven method. Particularly, ICAN [31] first proposes an instance-centric attention module that learns to dynamically highlight regions in an image conditioned on the appearance of each instance. Along this line, a bundle of representative works largely boosts HOI performance through some extra features, such as human pose [7], human intention [33], and language feature [29]. For instance, TSE [27] proposes a spatial enhancement approach to enforce fine-level spatial constraints in two directions between human body parts and object parts. AD-HOI [34] proposes a novel instance part-level attention deep framework for HOI detection, where the fine-grained part-level mutual context of a human–object pair is extracted to improve HOI detection. Also, auxiliary models, e.g., neural-symbolic methods [35] and graph models [29,36], can be easily introduced to help improve model performance. Additionally, recent advances in HOI detection focus on tackling the long-tailed distributions of HOI classes. Examples include transferring knowledge from seen categories to unseen ones by an analogy transformation [37], performing data augmentation of semantically similar objects [38].

One-stage HOI methods. Most one-stage detectors [12,14,15], directly locating the interaction point or interactive human–object pairs, are interaction-driven. In particular, based on a transformer encoder–decoder architecture, the pioneering work HOI-trans [15] directly introduces one type of naive query representations to predict HOI instances in a unified way. After that, a large bulk of work has been proposed investigating different detection strategies, e.g., detecting sequentially [14] or in parallel [12], etc. HOTR [12] explores a transformer encoder–decoder structure with a shared encoder and two parallel decoders, i.e., instance decoder and interaction decoder, to predict a set of object detection and interaction, then associates the human and object of the interaction. CDN [14], detecting HOI triplets in two cascade decoders, first resorts to the Human–Object Pair Decoder (HO-PD) to predict a set of human–object bounding-boxes pairs based on a set of learnable queries. Next, taking the output of the last layer of HO-PD as queries, an isolated interaction decoder is utilized to predict the action category for each query. Thereafter, the HOI triplets are formed by the output of the above two cascade decoders. From a methodological perspective, the inherent essence of these robust architectures lies in disentangling HOI detection. Besides, QPIC [39] proposes a query-based HOI detector, which incorporates contextually important information aggregated image-wide.

Different from them, we exhaustively disentangle HOI by introducing triple explicit queries to focus on the separate feature representations.

3. Method

In this work, we propose a new transformer-based approach that parallelly disentangles human–object detection and interaction classification in a triplet-wise manner. Looking at Fig. 2, the macro framework builds on a vanilla transformer as a cornerstone. The design principle is to exhaustively disentangle HOI in a triplet-wise manner (i.e., three queries of the intra-triplet set can correspond to the representations of human–object-interaction one by one) and preserve the consistent semantic association of HOI.

Given one raw image, PDN closely follows the one-stage transformer-based general recipe and starts from a convolution stem, followed by a transformer encoder to extract visual features into a sequence. Then we leverage triple explicit queries with consistent semantic association to aggregate HOI semantic features in the parallel disentangling decoder. Afterward, the multi-layer perception yields the resulting predictions for each special task, respectively.

3.1. Preliminaries

Before elaborating the architecture of PDN, we provide a brief description of the basic components in the conventional transformer proposed in the previous literature [16].

The one core component in a transformer is the Multi-head Self-Attention (MSA) mechanism. In MSA, the inputs $X \in \mathbb{R}^{n\times d}$ are linearly transformed and further packed into three parts, namely queries $Q \in \mathbb{R}^{n\times d_k}$, keys $K \in \mathbb{R}^{n\times d_k}$ and values $V \in \mathbb{R}^{n\times d_v}$ where n is the sequence length, d, d_k, d_v are the dimensions of inputs, queries (keys) and values, respectively. The attention operation is conducted on Q, K, V as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V. \tag{1}$$

Finally, a linear layer is used to produce the output. Multi-head self-attention splits the queries, keys and values to h (usually 8) parts and perform the attention function in parallel, and then the output values of each head are concatenated and linearly projected to form the final output.

In addition to MSA module, another basic component in the transformer is the feed-forward networks (FFN), which is applied between self-attention layers for feature transformation and non-linearity. The FFN consists of two fully connected layers with a ReLU activation:

$$\text{FFN}(X) = FC(\sigma(FC(X))), FC(X) = WX + b, \tag{2}$$

where W and b are the learnable weight and bias term of fully connected layer (FC) respectively, and $\sigma(\cdot)$ is the activation function such as ReLU.

At last, in theory, a complete transformer layer contains the above self-attention module and a feed-forward layer followed by a layer normalization (LN) with residual connection. Please refer to the work [16] for a more comprehensive review of transformer.

3.2. Visual feature extractor

We seamlessly combine a Convolutional Neural Network (CNN) and a transformer encoder as the visual feature extractor. An input image I is fed into a CNN to generate a feature map of shape (H, W, C), carrying high-level semantic concepts. Then, one 1×1 convolution layer decreases its channel dimension from C to d. Next, a flatten operator is leveraged to collapse its spatial dimension into one dimension, generating the flatten feature of shape $(H \times W, d)$. Finally, building upon standard transformer architecture with a multi-head self-attention module and a feed-forward network (FFN), an encoder receives the flatten feature and the position encoding E_{pos} as input and yields the resulting global features $X_s \in R^{(H\times W,d)}$. Due to the powerful capability of the transformer-encoder to capture the long-range dependencies, the produced global features possess richer contextual information.

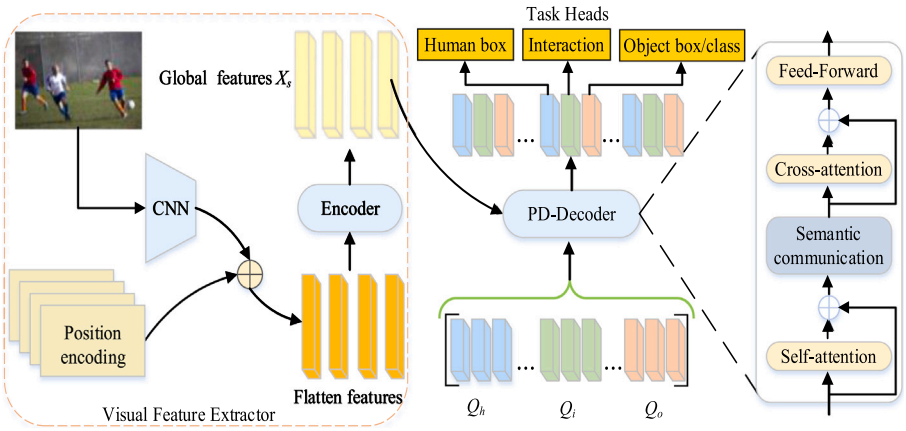


Fig. 2. Schematic illustration of our PDN. We first apply a CNN-transformer combined architecture to extract sequenced visual features X_s. Then, in the PD-decoder we exhaustively disentangle HOI based on X_s and triple random-initialized queries, i.e., human queries Q_h, object queries Q_o and interaction queries Q_i. Additionally, we design a semantic communication layer to preserve the consistent semantics association of HOI. The semantic communication layer formulates the query triplet set with the correspondence constraint and interchanges the information within each query triplet. Subsequently, the multi-layer perception (MLP) yields the resulting predictions for each specific task.

3.3. Parallel disentangling decoder

As illustrated in Fig. 2, other than the standard FFN, the PD-Decoder itself primarily contains three basic components: multi-head self-attention layer, semantics communication layer and multi-head cross attention layer. For convenience, we omit the multi-head attribute of attention layer and the layer normalization (LN), which stay the same as the previous literature [16], in PD-Decoder.

Self-attention layer. Beforehand, we randomly initialize three groups of learnable query vectors $Q_h \in R^{N_h \times C_q}, Q_i \in R^{N_i \times C_q}, Q_o \in R^{N_o \times C_q}$ as human queries, interaction queries, object queries to improve the query's capabilities of capturing multiple features, where triple query numbers N_h, N_i, N_o maintain one equal tunable size N. The queries are designed in such a way that one query concisely captures one semantic role of HOI triplet. N is therefore set to be large enough so that it is always larger than the number of actual human–object pairs in an image. After that, we arrange a self-attention layer, which takes the concatenation of triple queries Q_h, Q_i, Q_o as the query Q, key K and value V, to yield intermediate HOI query $Q_{hio} \in R^{(b,3\times N,d)}$, where b and d are the batch and dimension size. The formulation of the self-attention is given as:

$$Q_{hio} = \mathrm{softmax}(\frac{[Q_h, Q_i, Q_o]K^T}{\sqrt{d_k}})V, \tag{3}$$

where $[\cdot]$ is the concatenation operation; $\sqrt{d_k}$ maintains the scaling factor of the original transformer.

Semantic communication layer. Then, we introduce one semantic communication layer to preserve the consistent semantic association of HOI. As seen in Fig. 3, the semantic communication layer starts with a gating mechanism—a group convolution and a gated activation unit [40], followed by a group convolution and Swish activation [41]. Furthermore, batchnorm, residual connection and dropout are also deployed to aid training deep models. By design, we couple the identical index query from three types of queries to formulate the query triplet set with correspondence constraint. In practice, we first reshape Q_{hio} as $Q_s = [Q_s^h, Q_s^i, Q_s^o] \in R^{(b,N,3,d)}$ to form the dependent query triplet set (i.e., the intra-triplet set) and N independence query triplet sets (i.e., the inter-triplet set). The semantic communication layer leverages the group convolution, which aims to enhance not only the intra-triplet correspondence constraint but also the inter-triplet mutual independence. It is worth noting that the communication of query set in the above self-attention layer not only reinforces the inherent correspondence constraint of dependent queries (i.e., queries of the intra-triplet set), but also leads to the mutual contamination of the independent queries (i.e., queries of the inter-triplet set). To alleviate the query feature contamination resulting from communication of independent queries and preserve the consistent semantic association of dependent HOI, the group number of group convolution remains equal to the query number N, which empowers each dependent query triplet to exchange the information internally. Then the semantic communication layer receives $[Q_s^h, Q_s^i, Q_s^o]$ as inputs and yields updated queries Q_c as follows:

$$Q_c = [Q_c^h, Q_c^i, Q_c^o] = f_s(Q_s^h, Q_s^i, Q_s^o), \tag{4}$$

where $f_s(\cdot)$ is the semantic communication layer. In this way, Q_s can provide the compact knowledge to allow the learning processes of the updated queries $[Q_c^h, Q_c^i, Q_c^o]$ to benefit from each other.

Cross-attention layer. Sequentially, we feed the updated queries Q_c, global information X_s and position encoding E_{pos} into the cross-attention layer, followed by the FFN. This process can be formulated as:

$$P_c = [P_c^h, P_c^i, P_c^o] = f_c(Q_c, X_s, E_{pos}), \tag{5}$$

where $f_c(\cdot)$ is the process of the cross-attention layer and FFN like the transformer-based model HOI-Trans [15]. Besides, like the visual feature extractor, E_{pos} and other default setting reduplicate the operation of HOI-Trans [15]. Note, here, that the output Q_c is reshaped back into the size $(b, 3\times N, d)$. P_c^h, P_c^i and P_c^o denote the intermediate features and are ultimately sent into the following three task heads, respectively.

3.4. Task heads

Looking at Fig. 2, we deploy three task heads to localize the human–object boxes and classify the interaction in parallel. Here each head consists of one or several MLP branches, and each branch is for a specific task, e.g., detection or classification. The final outputs of each MLP branch are a set of action category a_i, object category a_i^o, human bounding box b_i^h or object bounding box b_i^o, $i \in \{1, 2, \ldots, N\}$.

3.5. Learning and inference

In this subsection, we introduce the learning and inference processes concretely.

Learning. Following the set-based training process of HOI-Trans [15], we first match each ground-truth with its best-matching prediction by

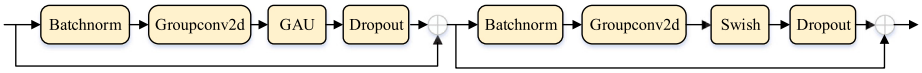


Fig. 3. An overview of our proposed semantic communication layer. The semantic communication layer contains a group convolution with an expansion factor of 2 projecting the number of channels and a gated activation unit (GAU), followed by a group convolution and then a swish activation layer.

the bipartite matching with the Hungarian algorithm. Then the loss is produced between the matched predictions and the corresponding ground truths for the final back-propagation. Besides, the HOI datasets usually have long-tail class distribution for both object class and action class. To alleviate the long-tail problem, we follow CDN [14] to arrange a dynamic re-weighting mechanism for further improvements with a decoupling training strategy. By removing the interactive score loss, PDN follows the leftover loss of CDN [14], which is composed of four parts: the box regression loss L_b, the intersection-over-union loss L_{GIoU}, the object class loss L_c^o, and the action category loss L_c^a. The target loss is the weighted sum of these parts as:

$$L = \sum_{k\in(h,o)} (\lambda_b L_b^k + \lambda_{GIoU} L_{GIoU}^k) + \lambda_o L_c^o + \lambda_a L_c^a, \tag{6}$$

where λ_b, λ_{GIoU}, λ_o and λ_a are the hyper-parameters for adjusting the weights of each loss.

Inference. As mentioned above, the detection results are represented by the following four components: human bounding box, object bounding box, interaction category and object category. Formally, we set the ith prediction results as $\langle b_i^h, b_i^o, c_i^o, c_i^a \rangle$, the final HOI triplet score s_i^{hoi} is given by $s_i^{hoi} = s_i^a s_i^o$, where s_i^a and s_i^o refer to the scores of interaction and object classification respectively (generated by sending the output of the corresponding branch of task heads into the softmax activation function). We further filter out these instance pairs based on the score s_i^{hoi}.

4. Experiments

In this section, we first introduce the adopted datasets, metrics and implementation. Next, we compare our proposed method with other state-of-the-art methods on two datasets, i.e., V-COCO and HICO-DET. Afterward, we deeply dig into the effectiveness of our proposed method through ablation studies and visualization of results. In the end, we provide the qualitative analysis of HOI samples to examine information encoded by our model, and comparisons of interaction classification performances on V-COCO dataset.

4.1. Datasets and evaluation metrics

Datasets. HICO-DET [42] is a common large-scale dataset for HOI detection. It includes 47,776 images (38,118 for training and 9658 for testing), annotated with 117 verbs and 80 object categories. Following the previous works [43], we report our performance in Full, Rare and Non-Rare Categories on HICO-DET.

V-COCO [44] consists of 5400 images in the train-val dataset and 4946 images in the test set. The training and validation set are collected from COCO [45] training set and the test images are from the COCO validation set. Each person is annotated with a binary label vector for 29 different actions (where each entry indicates whether the person is performing a certain action).

These datasets provide us with a large number of samples to investigate our model in human–object interaction for daily human activities.

Evaluation metrics. Following the standard evaluation, we leverage the commonly used mean Average Precision (mAP) to examine the model performance for both datasets. An HOI detection is considered as true positive only if it localizes the humans and objects accurately (i.e., the Interaction-over-Union (IOU) ratio between the predicted box and ground truth is greater than 0.5) and predicts the interaction correctly.

4.2. Implementation details

We instantiate a family of models with different parameters by both varying the convolution backbone and removing the semantic communication layer, as shown in Tables 1 and 2. Here S, M, L refer to small, middle and large, respectively. For PDN-S and PDN-M, we adopt ResNet-50 with a 6-layer transformer encoder as the visual feature extractor. For PD-Decoder, PDN-M is equipped with a stack of PD-Decoder blocks (6 layers), while PDN-S removes its semantic communication layer. PDN-L only replaces the ResNet-50 with ResNet-101 in PDN-M. The number of queries N are set to 30 for HICO-DET and 40 for V-COCO since the average number of positives for variant human–object pairs per image of HICO-DET is smaller than V-COCO. The reduced dimension size d is set to 256. In the task heads, the human and object box branches have 3 linear layers with ReLU while the object and action category branches have one linear layer.

During training, we implement the network with the parameters of DETR [17] trained with the COCO dataset. Following CDN [14], we set the weight coefficients λ_b, λ_{GIoU}, λ_o and λ_a to 2.5, 1, 1 and 1, respectively. We optimize the network by AdamW with the weight decay 10^{-4} and batch size 16. We first train the whole model for 90 epochs with a learning rate of 10^{-4} decreased by 10 times at the 60th epoch. Then, during the decoupling training process, we freeze the parameters of CNN and fine-tune the encoder and parallel disentangling decoder together with the human detection, object detection and interaction task heads for 10 epochs with a learning rate of 10^{-5}. Incidentally, all experiments are conducted on the 8 Tesla V100 GPUs.

4.3. Comparisons with state-of-the-art methods

We compare our PDN method with several existing approaches for evaluation, and present their performances on HICO-DET and V-COCO datasets in Tables 1 and 2.

Table 1 plainly summarizes the quantitative results on the HICO-DET dataset in the Default and Known Object modes. It is widely known the one-stage methods could provide much better performance than the two-stage ones. We speculate the reasons causing their bad performance lies in the drawbacks of two-stage methods: the interference of redundant negative human–object pairs and the sub-optimal solution resulting from the independent optimization on two subtasks. Here we only show the comparison between our method and the one-stage method. In contrast to PPDM [13] reformulating HOI detection as a point detection and matching problem, our PDN-S brings an especially significant mAP increase from 21.73 to 31.23, with a relative gain of 43.72%. When comparing to the prevailing one-stage transformer-based methods, our framework vastly outperforms HOI-Trans [15] which directly predicts HOI instance in a unified way and QPIC [39] which takes image-wide contexts into account by 6.57 and 4.11 mAP, respectively. The comparison with the work HOTR [12] which manipulates the instance and interaction decoder to predict the instance bounding box and relationship classification parallel is impressing. Here we gain 8.08 mAP improvement. Notably, with the same backbone network and learning strategy, the PDN-M outperforms CDN-S [14] 1.27% mAP and PDN-L surpasses the state-of-the-art CDN-L [14] by 3.46% mAP. Meanwhile, the experimental results also demonstrate our PDN outperforms the existing methods and benefits from disentangling HOI.

Table 2 clearly lists the performance comparisons on V-COCO dataset. The quantitative results read that PDN can beat many previous

Table 1
Comparisons with state-of-the-art methods on HICO-DET dataset.

Method	Backbone	Default			Known object		
		Full	Rare	NonRare	Full	Rare	NonRare
PPDM [13]	Hourglass-104	21.73	13.78	24.10	24.58	16.65	26.84
GG-net [46]	Hourglass-104	23.47	16.48	25.60	27.36	20.23	29.48
HOTR [12]	ResNet-50	25.10	17.34	27.42	–	–	–
HOI-Trans [15]	ResNet-101	26.61	19.15	28.84	29.13	20.98	31.57
QPIC [39]	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
CDN-S [14]	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42
CDN-L [14]	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38
PDN-S	ResNet-50	31.23	25.75	32.86	33.85	28.67	35.39
PDN-M	ResNet-50	31.84	26.38	33.47	34.45	29.34	35.97
PDN-L	ResNet-101	**33.18**	**27.95**	**34.75**	**35.86**	**30.57**	**37.43**

Table 2
Comparisons with state-of-the-art methods on V-COCO dataset.

Method	Backbone	AP^{S1}_{role}	AP^{S2}_{role}
HOI-Trans [15]	ResNet-101	52.9	–
AS-Net [47]	ResNet-50	53.9	–
GG-net [46]	Hourglass-104	54.7	–
HOTR [12]	ResNet-50	55.2	64.4
QPIC [39]	ResNet-50	58.8	61.0
CDN-S [14]	ResNet-50	61.7	63.8
CDN-L [14]	ResNet-101	63.9	65.9
PDN-S	ResNet-50	62.2	64.1
PDN-M	ResNet-50	62.9	64.8
PDN-L	ResNet-101	**64.7**	**66.7**

Table 3
Comparisons about FLOPs and parameters with state-of-the-art methods.

Method	Parameters (M)	FLOPs (G)
CDN-L	69.87	164
PDN-L	60.40	153

methods using lots of bells and whistles. In detail, PDN-L largely exceeds the adaptive set-based one-stage framework AS-Net [47] with 10.8 point mAP gains. Compared to GG-net [46] which mimics the two steps (Glance and Gaze) taken by humans to identify human–object interactions, PDN-L substantially raises the relative accuracy by 18.28%. Besides, PDN-L significantly outpaces the previous one-stage transformer-base method HOTR [12] and HOI-Trans [15] with 9.5 and 11.8 point mAP gains. Looking at the Table 2, we can find that PDN-L remarkably exceeds the state-of-the-art method CDN-L [14] with 1.25% point relation mAP gains. We also observe one gap between PDN-M and CDN-S [14]. The consistent performance improvement confirms the great potential of our method.

As for efficiency analysis, due to leaving out one transformer-style decoder (6 layers) with the quadratic computational complexity to the input image scale, PDN-L takes substantially fewer parameters and FLOPs than CDN-L [14]. More specifically, as can be seen, Table 3 shows PDN-L saves nearly 13.6% fewer parameters and 6.7% fewer FLOPs.

4.4. Ablation study

Here we thoroughly investigate how the major components and key parameters in our PDN influence the overall performance. We adopt PDN-L as the base model. Technically, based on PDN-L, PDN-B leaves out the semantic communication layer and PDN-W replaces the group convolution with the standard CNN. The experimental results are shown in Tables 4–9. Unless stated otherwise, the other setting follows the above implementation details and remains unchanged in the PDN series architecture. Incidentally, here HOI-Trans [15] takes ResNet-101 as the CNN backbone, uses the entangled (naive) queries to generate all results of human–object detection and interaction classification.

Table 4
Ablation study about the expanding queries on HICO-DET dataset.

Query type	Default			Known object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
Naive	30.25	24.58	31.94	32.06	28.19	33.21
Expanding	31.98	26.76	33.53	33.98	28.46	35.62

Table 5
Ablation study about the expanding queries on V-COCO dataset.

Query type	AP^{S1}_{role}	AP^{S2}_{role}
Naive	62.5	63.7
Expanding	63.3	64.8

Table 6
Ablation study about the semantic communication layer on HICO-DET dataset.

Method	Default			Known object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
PDN-B	31.98	26.76	33.53	33.98	28.46	35.62
PDN-W	32.31	27.04	33.88	34.77	29.39	36.38
PDN-L	**33.18**	**27.95**	**34.75**	**35.86**	**30.57**	**37.43**

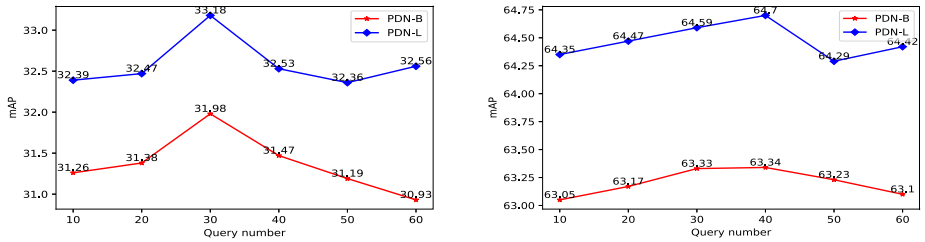
Table 7
Ablation study about the semantic communication layer on V-COCO dataset.

Method	AP^{S1}_{role}	AP^{S2}_{role}
PDN-B	63.3	64.8
PDN-W	63.8	65.3
PDN-L	**64.7**	**66.7**

Expanding queries. To analyze the contribution of disentangling HOI, we train ablated versions of our model separately, i.e., expanding queries in PD-Decoder. Here we adopt PDN-B which applies naive or expanding queries. As shown, Tables 4 and 5 indicate the second model using the expanding query respectively raises the performance of 5.72% and 1.28% mAP on HICO-DET and V-COCO when compared with the first model which adopts the naive queries to generate the instance bounding boxes and interaction classification. It indicates the superiority of disentangling HOI by expanding the naive queries as triple explicit queries. We believe that this is reasonable, because the expanding queries can make each query focus on one specific task clearly.

Semantic communication layer. To clarify the influence of semantic communication layer, we conduct experiment by removing semantic communication layer. Like the results reported in Tables 6 and 7, our PDN-L further yields 1.2 and 1.4 point mAP gain over PDN-B on HICO-DET and V-COCO, respectively. The mAP increases prove that the semantic communication layer can give the benefit. However, without the semantic communication of the correspondence constraint of the triple queries, i.e., replacing the group convolution with the standard CNN, PDB-W provides poor performance in comparison with PDN-L, suggesting the importance of the correspondence constraint of the triple queries.

Number of query. To verify the effectiveness of number of query, we conduct experiment by altering the query number. The shown result is tested on full class of default mode to HICO-DET dataset and Scenario 1 to V-COCO dataset. To some extent, increasing query number contributes to improving the results, but the tendency vanishes for larger number 40 for HICO-DET dataset. As seen in Fig. 4, our model get the best performance when we set the number of query to 40 for V-COCO and 30 for HICO-DET dataset. The reason lies in that the average number of positives for variant human–object pairs per image of HICO-DET is smaller than V-COCO.



(a) (b)

Fig. 4. Performance comparison about query number on HICO-DET (a) and V-COCO (b).

Table 8
Ablation study about the convolution kernel size on HICO-DET dataset.

Kernel size	Default			Known object		
	Full	Rare	Non-Rare	Full	Rare	Non-Rare
(2,1)	31.45	26.24	33.01	33.68	28.39	35.26
(2,3)	31.46	26.27	33.01	33.70	28.42	35.28
(2,5)	31.38	26.19	32.93	33.76	28.29	35.39
(3,1)	**33.18**	**27.95**	**34.75**	**35.86**	**30.57**	**37.43**
(3,3)	32.67	27.31	34.27	34.79	29.46	36.38
(3,5)	33.16	27.92	34.73	35.76	30.48	37.34

Table 9
Ablation study about the convolution kernel size on V-COCO dataset.

Kernel size	AP^{S1}_{role}	AP^{S2}_{role}
(2,1)	63.4	65.1
(2,3)	63.3	64.9
(2,5)	63.6	65.2
(3,1)	64.2	65.9
(3.3)	**64.7**	**66.7**
(3,5)	64.6	66.7

Convolution kernel size. To get further insights into the effect of the convolution kernel sizes of semantic communication layer, we successively sweep the kernel size in (2,1),(2,3), (2,5), (3,1),(3,3) and (3,5). Tables 8 and 9 clearly summarize the experimental result about convolution kernel size. Increasing the first dimension of key convolution kernel size from 2 to 3 can basically lead to the performance elevation on both datasets. However, the experiment results also show a slight fluctuation about detection performance with different second dimension of group convolution kernel sizes. We argue that the results are likely because keeping Human-Object-Interaction query information communication can better preserve the consistent semantic association of dependent HOI and the exchange of information on only two queries of the query triplet contributes less to preserving consistent semantic association.

5. Qualitative analysis

In this section, we give a depth analysis to interpret the visualization information of our PDN. We randomly sample three image from the HICO-DET validation set and evaluate them by the trained model. As shown in Fig. 5, we visualize the attention maps extracted from the last layer of the decoder, where the pixel's brightness indicates how much the feature was noticed. The first row shows HOI examples of original images. The rest rows visualize the attention maps of correspondence human, interaction and object, respectively. The samples showcase that the model works well at recognizing whether the subjects and objects are interactive in different scenes, which reveals the effect of our model PDN. The highlight regions that correspond to the different attentiveness fields give us some insight into what the model cares about and indicate that HOI semantic features are learned

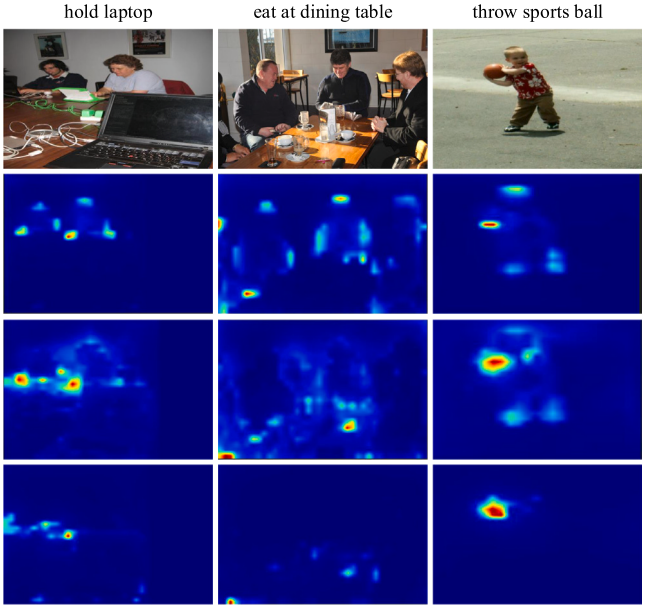


Fig. 5. Visualization of original images and attention maps in the last layer of PD-decoder. The first row shows HOI examples of original images. The rest rows visualize the attention maps of correspondence human, interaction and object, respectively.

effectively. We can see that three attention maps individually gaze at the discriminative region, i.e., the human and object attention maps emphasize the boundaries of persons and objects (see the images in the 2nd and 4th rows) while the interaction attention map mainly concentrates on the human–object contact areas (see the images in the 3rd row). As a special case, for 'hold laptop', the human attention map notably focuses on the human's profile; the object attention map highlights the boundaries of the laptop; while the interaction attention map contributes to interaction context, i.e., the human's hands holding the laptop. This means that our model makes them beneficial mutually, which leads to its superior performance over existing methods. Thus, this clearly demonstrates the PDN's capability of disentangling HOI.

Comparisons of per-class performances. We report the per-class performances in Table 10 on the V-COCO dataset. Compared with the existing methods, our proposed PDN achieves better performance in majority of the classes with the application of disentangling HOI. Additionally, the per-class performances manifest that some of the action classes perform incorrectly due to the failure of detecting some small objects in images, e.g. eat instruments which usually have small objects and commonly become occluded in the images. The shortcomings mainly stem from the deficit of components in processing image feature maps: the same instance feature extraction scale. As such, we believe a better module that learns to attend to multi-scale features may improve the performance.

Table 10
Per class AP comparisons to the existing methods in V-COCO. Our method demonstrates superior performance in majority of the classes. All the results are reported as percentage.

Method	InteractNet [48]	iCAN [31]	VSGnet [49]	PMFNet [7]	PD-Net [50]	IPGN [51]	PDN
hold-obj	26.38	29.06	48.27	44.01	45.07	51.12	59.93
sit-instr	19.88	26.04	29.9	29.51	31.86	31.82	58.46
ride-instr	55.23	61.9	70.84	70.33	71.80	69.72	71.59
look-obj	20.2	26.49	42.78	45.22	46.72	45.28	57.65
hit-instr	62.32	74.11	76.08	76.30	78.57	75.6	83.72
hit-obj	43.32	46.13	48.6	52.28	50.28	53.51	82.37
eat-obj	32.37	37.73	38.3	44.55	47.41	41.77	61.73
eat-instr	1.97	8.26	6.3	5.9	6.94	7.9	35.96
jump-instr	45.14	51.45	52.66	53.39	52.75	55.11	78.45
lay-instr	20.99	22.4	21.66	26.40	28.25	26.27	67.37
talk on pho	31.77	52.81	62.23	54.69	56.64	66.33	66.28
carry-obj	33.11	32.02	39.09	44.24	45.64	44.76	48.12
throw-obj	40.44	40.62	45.12	49.76	47.82	50.38	64.35
catch-obj	42.52	47.61	44.84	54.11	55.01	48.87	60.28
cut-instr	22.97	37.18	46.78	40.08	42.69	45.69	54.31
cut-obj	36.4	34.76	36.58	40.01	39.24	40.27	63.59
work on com	57.26	56.29	64.6	67.39	67.98	64.5	70.79
ski-instr	36.47	41.69	50.59	53.04	52.59	54.34	63.41
surf-instr	65.59	77.15	82.22	80.47	80.95	81.3	81.35
skateboard	75.51	79.35	87.8	86.81	88.00	86.6	88.53
drink-instr	33.81	32.19	54.41	46.76	53.84	56.27	59.94
kick-obj	69.44	66.89	69.85	72.70	74.50	68.96	84.53
read-obj	23.85	30.74	42.83	36.80	39.07	45.19	58.95
snowboard-ins	63.85	74.35	79.9	74.33	76.55	79.5	78.71
Average	40.0	45.3	51.76	52.05	53.34	53.79	66.7

6. Conclusion

In this paper, we present the Parallel Disentangling Network (PDN), a novel one-stage framework with disentangling HOI in a triplet-wise manner, for HOI detection. PDN keeps the advantages of unified one-stage methods, directly locating the human–object pairs and interaction, and brings the advantages of parallel and cascade one-stage methods into our framework, disentangling HOI detection. The key insight is to exhaustively disentangle HOI to learn respective semantic features through triple explicit queries. Besides, PDN also benefits both from formulating the query triplet set with correspondence constraint and introducing the semantic communication operation. These processes interchange the representation information within each intra-triplet set, and enable the inter-triplet sets mutually independent, since the feature representations aggregated by each intra-triple set are dedicated to detecting one HOI. We empirically verify the effectiveness of PDN and demonstrate that PDN outperforms existing methods and achieves a prominent mAP gain with significant reduction in parameters and FLOPs. In the future work, we will focus on designing a multi-scale detection framework that can be applied to handle the large variations in the scale of visual objects.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (U1813202).

References

[1] O.C. Kurban, N. Calik, T. Yildirim, Human and action recognition using adaptive energy images, Pattern Recognit. 127 (2022) 108621.
[2] J. Zhang, F. Shen, X. Xu, H.T. Shen, Temporal reasoning graph for activity recognition, IEEE Trans. Image Process. 29 (2020) 5491–5506.
[3] K. Gedamu, Y. Ji, Y. Yang, L. Gao, H.T. Shen, Arbitrary-view human action recognition via novel-view action generation, Pattern Recognit. 118 (2021) 108043.
[4] Z. Yang, P. Wang, T. Chu, J. Yang, Human-centric image captioning, Pattern Recognit. 126 (2022) 108545.
[5] J. Ji, Z. Du, X. Zhang, Divergent-convergent attention for image captioning, Pattern Recognit. 115 (2021) 107928.
[6] Q. Zhao, X. Wang, S. Lyu, B. Liu, Y. Yang, A feature consistency driven attention erasing network for fine-grained image retrieval, Pattern Recognit. 128 (2022) 108618.
[7] B. Wan, D. Zhou, Y. Liu, R. Li, X. He, Pose-aware multi-level feature network for human object interaction detection, in: IEEE International Conference on Computer Vision, ICCV, 2019.
[8] Y.-L. Li, X. Liu, H. Lu, S. Wang, J. Liu, J. Li, Detailed 2d-3d joint representation for human-object interaction, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020.
[9] Y.-L. Li, X. Liu, X. Wu, Y. Li, C. Lu, Hoi analysis: Integrating and decomposing human-object interaction, Adv. Neural Inf. Process. Syst. 33 (2020) 5011–5022.
[10] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015) 91–99.
[11] R.N. Strickland, H.I. Hahn, Wavelet transform methods for object detection and recovery, IEEE Trans. Image Process. 6 (5) (1997) 724–735.
[12] B. Kim, J. Lee, J. Kang, E.-S. Kim, H.J. Kim, HOTR: End-to-end human-object interaction detection with transformers, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021.
[13] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, J. Feng, Ppdm: Parallel point detection and matching for real-time human-object interaction detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020.
[14] A. Zhang, Y. Liao, S. Liu, M. Lu, Y. Wang, C. Gao, X. Li, Mining the benefits of two-stage and one-stage HOI detection, Adv. Neural Inf. Process. Syst. (2021).
[15] 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: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 5998–6008.
[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European Conference on Computer Vision, ECCV, Springer, 2020.
[18] W. Dong, Z. Zhang, C. Song, T. Tan, Identifying the key frames: An attention-aware sampling method for action recognition, Pattern Recognit. (2022) 108797.

[19] J.H. Tan, C.S. Chan, J.H. Chuah, End-to-end supermask pruning: Learning to prune image captioning models, Pattern Recognit. 122 (2022) 108366.
[20] J. Deng, Z. Yang, T. Chen, W. Zhou, H. Li, Transvg: End-to-end visual grounding with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1769–1779.
[21] Y. Wang, X. Zhang, T. Yang, J. Sun, Anchor detr: Query design for transformer-based detector, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 2567–2575.
[22] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2020.
[23] 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, 2020, arXiv preprint arXiv:2010.11929.
[24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[25] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.
[26] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021).
[27] L. Liu, R.T. Tan, Human object interaction detection using two-direction spatial enhancement and exclusive object prior, Pattern Recognit. 124 (2022) 108438.
[28] D.-G. Lee, S.-W. Lee, Human interaction recognition framework based on interacting body part attention, Pattern Recognit. (2022) 108645.
[29] C. Gao, J. Xu, Y. Zou, J.-B. Huang, Drg: Dual relation graph for human-object interaction detection, in: European Conference on Computer Vision, ECCV, Springer, 2020.
[30] J. Jiang, Z. He, S. Zhang, X. Zhao, J. Tan, Learning to transfer focus of graph neural network for scene graph parsing, Pattern Recognit. 112 (2021) 107707.
[31] C. Gao, Y. Zou, J.-B. Huang, ican: Instance-centric attention network for human-object interaction detection, 2018, arXiv preprint arXiv:1808.10437.
[32] D. Yang, Y. Zou, Z. Li, G. Li, Learning human-object interaction via interactive semantic reasoning, IEEE Trans. Image Process. 30 (2021) 9294–9305.
[33] B. Xu, J. Li, Y. Wong, Q. Zhao, M.S. Kankanhalli, Interact as you intend: Intention-driven human-object interaction detection, IEEE Trans. Multimed. 22 (6) (2019) 1423–1432.
[34] L. Bai, F. Chen, Y. Tian, Automatically detecting human-object interaction by an instance part-level attention deep framework, Pattern Recognit. 134 (2023) 109110.
[35] Y. Xie, Z. Xu, M.S. Kankanhalli, K.S. Meel, H. Soh, Embedding symbolic knowledge into deep networks, Adv. Neural Inf. Process. Syst. 32 (2019).
[36] H. Wang, W.-s. Zheng, L. Yingbiao, Contextual heterogeneous graph network for human-object interaction detection, in: European Conference on Computer Vision, ECCV, 2020.
[37] J. Peyre, J. Sivic, I. Laptev, C. Schmid, Weakly-supervised learning of visual relations, in: Proceedings of the Ieee International Conference on Computer Vision, 2017, pp. 5179–5188.
[38] A. Bansal, S.S. Rambhatla, A. Shrivastava, R. Chellappa, Detecting human-object interactions via functional generalization, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 10460–10469.
[39] M. Tamura, H. Ohashi, T. Yoshinaga, QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021.
[40] Y.N. Dauphin, A. Fan, M. Auli, D. Grangier, Language modeling with gated convolutional networks, in: International Conference on Machine Learning, PMLR, 2017.
[41] P. Ramachandran, B. Zoph, Q.V. Le, Searching for activation functions, 2017, arXiv preprint arXiv:1710.05941.
[42] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, J. Deng, Learning to detect human-object interactions, in: IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018.
[43] B. Kim, T. Choi, J. Kang, H.J. Kim, Uniondet: Union-level detector towards real-time human-object interaction detection, in: European Conference on Computer Vision, ECCV, 2020.
[44] S. Gupta, J. Malik, Visual semantic role labeling, 2015, arXiv preprint arXiv:1505.04474.
[45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, ECCV, 2014.
[46] X. Zhong, X. Qu, C. Ding, D. Tao, Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021.
[47] M. Chen, Y. Liao, S. Liu, Z. Chen, F. Wang, C. Qian, Reformulating hoi detection as adaptive set prediction, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2021.
[48] G. Gkioxari, R. Girshick, P. Dollár, K. He, Detecting and recognizing human-object interactions, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[49] O. Ulutan, A. Iftekhar, B.S. Manjunath, Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2020.
[50] X. Zhong, C. Ding, X. Qu, D. Tao, Polysemy deciphering network for human-object interaction detection, in: European Conference on Computer Vision, ECCV, 2020.
[51] H. Wang, L. Jiao, F. Liu, L. Li, X. Liu, D. Ji, W. Gan, IPGN: Interactiveness proposal graph network for human-object interaction detection, IEEE Trans. Image Process. 30 (2021) 6583–6593.

Yamin Cheng is currently pursuing the Ph.D. degree with School of Computer Science and Engineering in University of Electronic Science and Technology of China, China. His research interests include human–object interaction detection, computer vision and pattern recognition.

Hancong Duan received the B.S.degree in computer science from Southwest Jiaotong University in 1995, the M.E.degree in computer architecture in 2005, and the Ph.D. degree in computer system architecture from UESTC in 2007. Currently he is a professor of computer science at UESTC. His current research interests focus on deep learning, distributed storage and cloud computing.

Chen Wang received the B.S. and M.S. degrees from Ningbo University, China, in 2013 and 2016 respectively. He is currently a Ph.D. candidate with the School of Computer Science and Engineering in University of Electronic Science and Technology of China. His current research interests include computer vision and pattern recognition.

Zhijun Chen is currently a Ph.D. candidate with School of Computer Science and Engineering, Beihang University. His major research interests include machine learning and crowdsourcing.