RR-Net: Relation Reasoning for End-to-End Human-Object Interaction Detection

Dongming Yang, Yuexian Zou*, *Senior Member, IEEE,* Can Zhang, *Member, IEEE* Meng Cao, and Jie Chen, *Member, IEEE*

***Abstract*—The task of Human-Object Interaction (HOI) detection devotes to learn how humans interact with surrounding objects via inferring fine-grained triplets of ⟨ human, verb, object ⟩. Most HOI feature learning techniques are dependent on pre-detected instance regions or human body-part regions, which are computationally expensive and hardly applicable to end-to-end detectors in real applications. In this paper, based on an end-to-end HOI detector, we make a first try to explore region-independent relation reasoning for HOI detection. We first present a Relation-aware Frame, which brings a progressive structure for interaction inference. Upon the Relation-aware Frame, an Interaction Intensifier Module and a Correlation Parsing Module are carefully designed, where: a) interactive semantics from humans can be exploited and passed to objects to intensify interactions, b) interactive correlations among humans, objects and interactions are integrated to promote predictions. Based on modules above, we construct a fully differentiable and end-to-end trainable network named Relation Reasoning Network (abbr. RR-Net). Extensive experiments show that our proposed RR-Net leads to competitive results compared with the state-of-the-art methods on both V-COCO and HICO-DET benchmarks and improves the baseline about 7.6% and 11.1% relatively, validating that this first effort in exploring region-independent relation reasoning has brought obvious improvement for end-to-end HOI detection.**

***Index Terms*—relation reasoning, interactive representation, human-object interaction, end-to-end.**

I. Introduction

FINE-GRAINED understanding of visual contents is one of the fundamental problems in computer vision. The task of Human-Object Interaction (HOI) detection aims to localize and infer triplets of ⟨ human, verb, object ⟩ from an image, which produces wealthy information for understanding complex human activities. This task can be applied in a wide range of fields such as intelligent surveillance, video analysis, human-robot interaction [1], etc. Beyond comprehending instances in isolation, e.g., object [2]–[7] and pedestrian [8], [9] detection, segmentation [10]–[12], human pose estimation [13], [14] and human parsing [15]–[17], detecting HOIs requires a deeper understanding of visual semantics to depict complex relationships between ⟨ human, object ⟩ pairs. HOI detection also presents more difficulties compared with traditional action recognition [18]–[20], e.g., an individual can simultaneously take multiple interactions with surrounding objects. Moreover, associating ever-changing roles with various objects leads to finer-grained and diverse samples of interactions. Therefore, HOI detection is challenging and usually needs to mine high-level and HOI-specific representations in solutions.

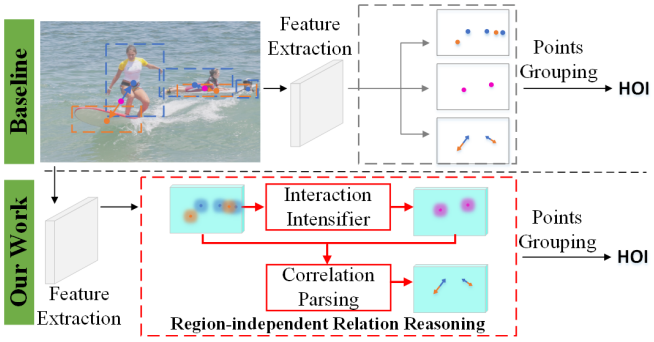

Fig. 1: A simplified visualization of how our method facilitates HOI detection. The baseline framework (above) lacks learning relation information, while we present a framework (bottom) that performs relation reasoning to learn HOI-specific representations without regions.

The conventional two-stage HOI detection methods [21]–[28] first obtain human and object region proposals from pre-trained object detectors [2], [3] to generate ⟨ human, object ⟩ pairs and then perform interaction classifications. These methods are with low efficiency because their two stages are separated and cannot be integrated. Recently, it has become a new trend to develop end-to-end HOI detectors. Inspired by the point-based object detectors [29], [30], as illustrated in Figure 1 (above), the latest end-to-end HOI detectors pose HOI detection as a point detection and grouping problem, which effectively improve the speed of HOI detection.

We revisit existing HOI detection methods and find that previous HOI feature learning techniques are all region-based where attention features [21]–[23], [31]–[33] from detected instances (e.g., humans and objects), spatial configuration [5], [22], [24], [25] and human pose [34]–[37] are exploited. However, these techniques are limited in facilitating end-to-end HOI detectors because: 1) Firstly, dominant models

Dongming Yang, Can Zhang, Meng Cao are with the School of ECE, Peking University, Shenzhen, China, 518055 (e-mail:yangdongming, zhangcan, mengcao@pku.edu.cn)

Yuexian Zou is with the School of ECE, Peking University, Shenzhen, China, also with the Peng Cheng Laboratory, Shenzhen, China, 518055 (e-mail:zouyx@pku.edu.cn)

Jie Chen is with the School of ECE, Peking University, Shenzhen, China, also with the Peng Cheng Laboratory, Shenzhen, China, 518055 (e-mail:chenj@pcl.ac.cn)

Copyright ©20xx IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

(* Corresponding author:Yuexian Zou)

focus on region-level relation modeling or explicit feature enhancements. However, region-independent relation reasoning is significant for learning HOI-specific representations among humans, objects and interactions. 2) Secondly, region-based feature learning is dependent on pre-detected instance regions or human body-part regions, which is computationally expensive and hardly applicable to end-to-end detectors in real applications. Besides, employing human pose or body-part features brings additional workload. Therefore, studying relation reasoning to integrate region-independent HOI representations has become the key to improve end-to-end HOI detection.

As illustrated in Figure 1 (bottom), based on an end-to-end HOI detector, we propose to capture and transmit HOI-specific representations among humans, objects and interactions where no human or object regions are needed. The contributions of our work are summarized as follows:

- A Relation-aware Frame is first presented to build a progressive structure for interaction inference, which imitates the human visual mechanism of recognizing HOI by comprehending visual instances and interactions coherently.
- Upon the Relation-aware Frame, an Interaction Intensifier Module is proposed, where interactive semantics from humans can be exploited and easily passed to objects. Therefore, responses between human and different objects can be relatively valued to intensify interactions.
- A Correlation Parsing Module is designed subsequently, in which interactive correlations among visual targets can be exploited and integrated. Such a module helps to better group HOI triplets without using instance regions.
- We offer a new end-to-end framework named Relation Reasoning Network (abbr. RR-Net), which is the first to carry out region-independent relation reasoning for HOI detection.

Extensive experiments are conducted on two commonly used benchmarks, including V-COCO [38] and HICO-DET [39] datasets. Our method provides obvious performance gain compared with the baseline and outperforms the state-of-the-art methods by a sizable margin. Detailed ablation study and analysis of our method are also provided to identify contributions from individual components and facilitate future research.

This paper is partly based on our preliminary work [40] and we extend it in a number of significant aspects. First, we carry out further exploration studies on Relation-aware Frame and introduce a relation function group $G()$, which can better exploit association among HOI cognitive processes. Second, we bring more detailed derivation of our method. Third, we add extensive experiments and more comprehensive analysis to validate our method. Finally, our newly proposed RR-Net achieves novel performance enhancements.

II. Related Work

A. Frameworks in HOI Detection

Beyond detecting instances from images [8], [41], visual semantic role labeling [38] is introduced to learn interactions [42] between humans and objects for the first time. Since human and object instances are basic elements in detecting HOIs, conventional HOI detection methods mostly extend from object detection frameworks to develop two-stage detectors [21], [22], [39], [43]. The first stage of these detectors is generating $\langle$ human, object $\rangle$ proposals from human and object region proposals obtained from a pre-trained object detector. Then, proposal classification in the second stage will predict interactions for each $\langle$ human, object $\rangle$ pair. Specifically, InteractNet [21] extends the object detector Faster R-CNN [2] with an additional branch for HOI detection. HO-RCNN [39] and iCAN [22] then introduce a three-branch architecture with one branch each for a human candidate, an object candidate, and an interaction pattern, which has become the template framework for most later HOI detectors. Cascaded human-object interaction recognition methods [26], [27] introduce cascade architectures for a multi-stage, coarse-to-fine HOI understanding, where an instance localization/detection network and an interaction recognition network are densely connected at each stage. TIN [43] further proposes a framework with multiple networks to perform feature extraction, interactiveness discrimination and HOI classification, respectively.

More recently, it has become popular to treat object detection as a point estimation task [29], [30] to improve detection efficiency. Following the same idea, IPNet [44] and PPDM [45] pose HOI detection as a point detection and grouping problem, which propose to build end-to-end HOI detectors. Specifically, PPDM [45] contains two parallel branches, where the first branch predicts three points (i.e., human point, object point and interaction point) and another branch predicts two displacements from the interaction point to its corresponding human and object points. Then, human with its corresponding object can be paired using the interaction point and displacements as the bridge. However, due to the lack of relation reasoning, these end-to-end HOI detectors learn limited HOI-specific representations in their methods, thus may be deficient in modeling HOIs (e.g., bring problem of miss-match between human and object).

B. Feature Learning in HOI Detection

Recent advances mainly focus on HOI feature learning to resolve ambiguity in interaction predictions, including incorporating contextual cues and developing region-based models.

1) Contextual Cues: Several contextual cues have been explored to improve HOI detection, where spatial configuration and human pose [13], [14] have become the most popular cues. Given a $\langle$ human, object $\rangle$ proposal with bounding boxes, the spatial configuration [22], [25], [39] is a coarse region layout encoded by a binary image with two channels. Tanmay [24] develops both a coarse region layout of human-object box-pair and a fine-grained layout of human pose, in which the absolute position and relative configuration of the human and object boxes are encoded. Additionally, human pose obtained from pre-trained pose estimators [3], [13] is unitized as a valid cue to support HOI detection. With the given human pose vector, pair-wise human body-parts [46] are explored as a set of local regions centered at human keypoints with a certain size. Furthermore, PMFNet [34] develops a multi-branch network to learn a pose-augmented representation to

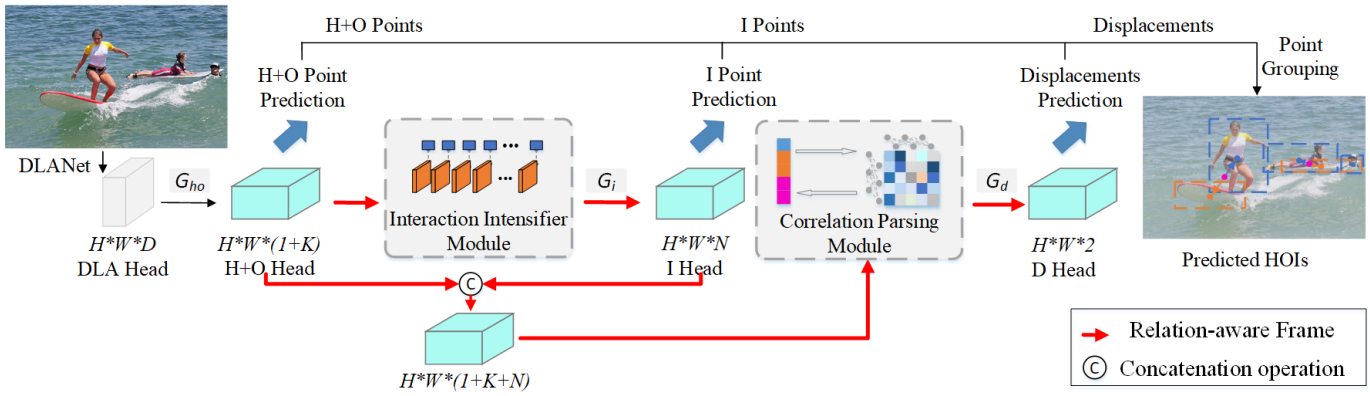

Fig. 2: An overview of RR-Net. In our framework, the instance points prediction, interaction points prediction and displacements prediction are treated as highly correlated processes. An Interaction Intensifier Module and a Correlation Parsing Module are further developed to achieve efficient relation reasoning to learn region-independent HOI representations.

incorporate interaction context, object features and detailed human body-part regions. Recently, FCMNet [47] combines the spatial encoding, context encoding and motion encoding [48] by a fusion module to improve HOI detection. Although extracting region-based spatial encodings or human body-part regions improve the detection results, these methods are not favored since additional computation and premade annotations are indispensable.

2) Appearance Attention and Graph Models: Inspired by attention mechanism [49]–[51] and relation networks [5], [6], [52]–[55] in action recognition [18], [56], recent researches focus on appearance attention [22], [23], [25], [31], [34], [57], [58] to improve HOI detection. The iCAN [22] exploits an instance-centric attention mechanism to enhance the semantic information from regions. In addition, Contextual Attention [57] proposes a deep contextual attention framework for HOI detection, in which context-aware appearance features for human and object are captured. The iHOI [23] and HOR [32] exploit human intention to guide the attended contextual regions for detecting human-object interactions. The PMFNet [34] focuses on pose-aware attention for HOI detection by employing attentional human parts. The VSGNet [25] proposes a spatial attention branch, which refines visual features by the spatial configuration of human-object pair. Moreover, several graph-based models [28], [36], [59]–[61] are introduced, where HOIs are represented with graphs and learned in an end-to-end manner. The GPNN [28] brings an idea of parsing and incorporating structural knowledge, which adopts an iterative graph with GRU [62] to transmit information among multiple instances. The in-GraphNet [60] and DRG [59] propose to describe human-object pairs and aggregate scene contexts via learnable graph-based structures. Lately, RPNN [35] introduces an object-body graph and a human-body graph to capture contexts between body-parts and surrounding instances (i.e., human and object), where fine-grained human body-part regions are required as prior information.

Overall, appearance attention models and graph models are enlightening but complex in design. In addition, the computational cost of these models is large. For example, implementing attention mechanisms upon M human regions and N object regions needs $M + N$ standalone calculations. Analogously, constructing graphs with $M * N$ possible $\langle$ human, object $\rangle$ pairs needs to calculate $M * N$ candidate samples or spatial encodings. In this paper, unlike the previous region-based attention and relation modeling, we bring region-independent relation reasoning to integrate and transmit implicit HOI representations for HOI inference.

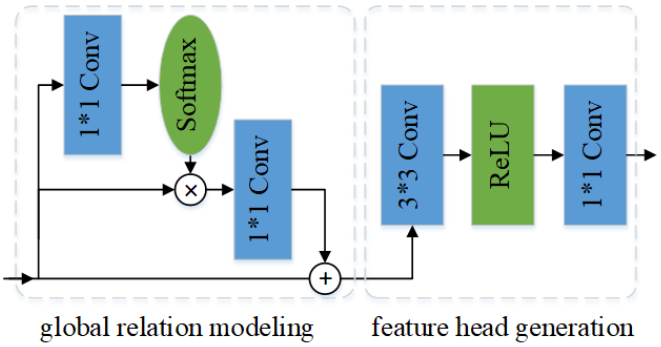

Fig. 4: Architecture of relation function group $G()$. The $G()$ first models global relation context from previous head features and then produces next head features.

III. Proposed Method

In this section, we introduce our work for HOI detection. We start with an overview of proposed RR-Net (Section III-A). Then, we present the Relation-aware Frame (Section III-B) and describe how we build the coherent structure. Afterwards, the Interaction Intensifier Module (Section III-C) and Correlation Parsing Module (Section III-D) are introduced. Finally, we outline our training and inference procedures (Section III-E).

A. Overview of RR-Net

An overview of our proposed RR-Net is shown in Figure 2. The DLA-34 [63] is employed as the backbone network in our implementation to extract the global DLA feature from an

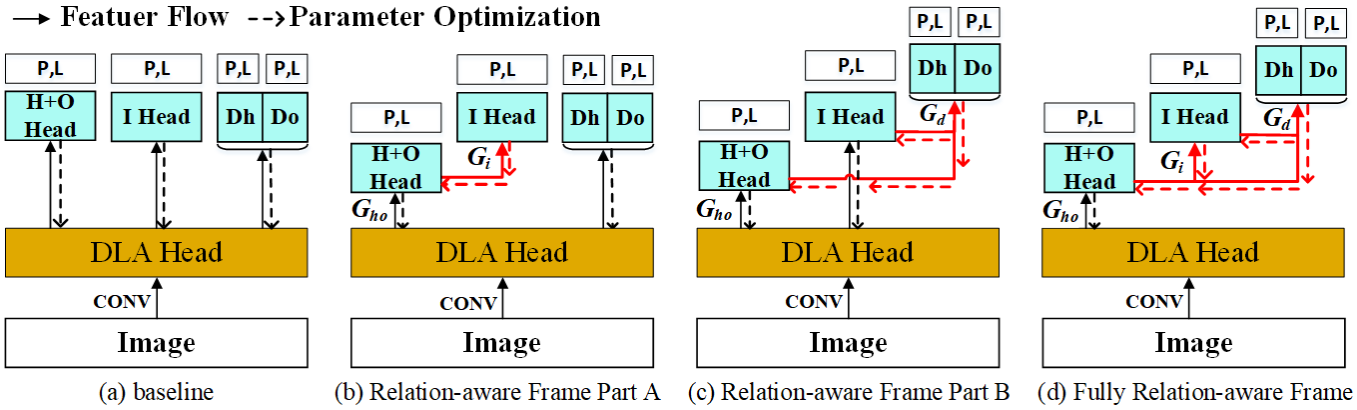

Fig. 3: Relation-aware Frame visualization. (a) In baseline, different heads are built independently, and head parameters are optimized severally in backward propagation. (b) Relation-aware Frame Part A: building I head upon H and O heads. (c) Relation-aware Frame Part B: building two D heads upon all points heads (i.e., H, O and I heads). (d) Fully Relation-aware Frame: building I head upon H and O heads. Meanwhile, building two D heads upon all points heads. Here, P means prediction operation and L means loss function.

image. We formulate HOI detection as a point detection and grouping problem [44], [45], where predicting HOI triplets consists of three steps: 1) predicting human points (H points), K-class object points (O points) and N-class interaction points (I points). Here, I point is defined as the midpoint of a corresponding ⟨ human, object ⟩ pair. 2) predicting two displacements from the I point to the H point and O point respectively, where each displacement contains an abscissa value and an ordinate value. 3) grouping each I point with the H point and O point respectively to generate a set of triplets based on the predicted points and displacements. The pseudo code description of RR-Net is presented in Algorithm 1.

We notice that existing works rarely study the association among HOI cognitive processes or explore interactive information in these processes. However, recognizing HOIs is a holistic task, in which each process is closely related to the others and provides significant interactive information for HOI detection. Based on the observation above, we first propose a Relation-aware Frame (Section III-B) to provide a coherent structure for interaction inference. Upon the Relation-aware Frame, an Interaction Intensifier Module (Section III-C) and a Correlation Parsing Module (Section III-D) are then developed to achieve region-independent relation reasoning so that interactive semantics and correlations among visual targets can be efficiently integrated. Next, we will introduce the proposed framework in detail.

Algorithm 1: RR-Net

Input: Image
Output: Predicted ⟨ human, verb, object ⟩ triplets

```
Predicting H points, K-class O points;
// IIM on Relation-aware Frame Part A
Predicting N-class I points;
// CPM on Relation-aware Frame Part B
Predicting displacements;
for h, o, i, dp^ih, dp^io do
    // Points & Displacements
    if S^h_(x^h,y^h), S^o_(x^o,y^o), S^i_(x^i,y^i) ∈ TOP T then
        Calculating all grouping distances by Eq. 12;
        if (i, dp^ih) → h, (i, dp^io) → o, then
            Grouping the ⟨ h, i, o ⟩ triplet;
            Calculating the final score S_hoi by Eq. 13;
        end
    end
end
```

B. Relation-aware Frame

As illustrated in Figure 3 (a), to predict H points, O points, I points and two displacements, the baseline builds different feature heads independently, and their parameters are optimized severally in backpropagation. We argue that such a structure neglects the implicit relation among these predictions. It is noted that the human visual system usually detects instances and groups related ones coherently from the scene to recognize a HOI. Considering the triplet ⟨ human, ride, bike ⟩ as an example, the person and bike may be detected first from the scene. Then, features obtained during the above process are not discarded but further used to determine the category and elements of the interaction. Based on the above inspiration, we propose a coherent structure for interaction inference, which is termed as a Relation-aware Frame.

Specifically, we consider points and displacements prediction as highly correlated processes, so that we link all prediction procedures together by the relation function group $G(\Delta) = [G_{ho}(\Delta), G_i(\Delta), G_{dh}(\Delta), G_{do}(\Delta)]$. It should be mentioned that we have merged $G_{dh}(\Delta)$ and $G_{do}(\Delta)$ as $G_d(\Delta)$ in Figure 3 for simplicity. For predicting humans and objects, we first generate H and O head feature $f_{ho} \in \Re^{H*W*(1+K)}$ from global DLA head feature $f_{dla} \in \Re^{H*W*D}$ by the function $G_{ho}(\Delta)$, where $H*W$ denotes range, D denotes feature dimension and $(1+K)$ is the category synthesis of human and K-class objects. The two stages of our Relation-aware Frame are then established to break the independent predictions of baseline, which are:

- Part A: building I head feature $f_i \in \Re^{H*W*N}$ upon feature f_{ho} by the function $G_i(\Delta)$, rather than parallelly upon DLA head.
- Part B: building D head features $f_{dh} \in \Re^{H*W*2}$ and $f_{do} \in \Re^{H*W*2}$ upon all point head features f_p by the function $G_d(\Delta)$, where $f_p = concat(f_{ho}, f_i)$.

For ease of understanding, we provide Figure 3 (b) and (c) to show these two parts respectively. As illustrated in Figure 3 (d), in the fully Relation-aware Frame, output head features of each previous stage are used as input in the next stage.

To enhance head features in Relation-aware Frame and make each of them from previous stage usable in the next stage, we propose relation function group $G(\Delta)$ shown in Figure 4. In the $G(\Delta)$, we first develop a compact block to model global relation context, which can be regarded as an instantiation of general Non-Local [50] idea. The difference is we do not compute the long-range dependencies of query position, but simply model the global context shared among the feature map. This modeling can be abstracted into three steps: 1) a $1 * 1$ convolution followed by a softmax function to obtain the global attention weights, and then an attention pooling with a matrix multiplication ($\otimes$) to obtain the global relation features; 2) a $1 * 1$ convolution to perform feature transformation; 3) an addition operation ($\oplus$) to merge global relation context to original features.

After global relation modeling, the feature head generation is then computed by:

$$F(X) = Conv2d(ReLU(Conv2d(X))), \tag{1}$$

where the first $3 * 3$ convolution transforms the features from input dimension to hidden dimension, and the second $1 * 1$ convolution acts as a fully connected layer to reduce the feature dimension from hidden dimension to specified dimension to match the predictions. The next procedures of predicting H points, O points and I points can be regarded as heatmap estimation processes by sigmoid functions.

C. Interaction Intensifier Module

Detecting a HOI not only needs to learn appearance features from the human and object, but also needs to reason high-level information that implied between the human and object, which we called interactive semantics. For example, when we see a person with the posture of sit, it reminds us of chairs, but not footballs or skateboards. It can even remind us that the interaction location of this person might be its lower body. It is because we have already learned interactive semantics between $human$ and $chair$. To exploit interactive semantics above to intensify interactions, we propose an Interaction Intensifier Module (abbr. IIM) upon the Relation-aware Frame Part A.

After running $G_{ho}(\Delta)$, each obtained feature map of H and O heads contains characteristic information for prediction of corresponding instances. We innovatively employ these features to perform relation reasoning, where no human or object regions are needed. As mentioned above, the total number $(1 + K)$ of categories contains human and K other object categories. As illustrated in Figure 5 (a), We can further split f_{ho} into $f_h \in \Re^{H*W*1}$ and $f_o \in \Re^{H*W*K}$. Then, a multilayer perceptron (MLP) with two fully connected layers followed by a sigmoid operation is adopted on feature f_h to generate the embedded interactive vector $\beta = [\beta_1, \beta_2, ..., \beta_K]$, denoting relevance between humans and different objects. Afterwards, O head features f_o are weighted with vector β by a multiplication operation, generating the valued O head features f'_o.

$$\beta = sigmoid(MLP(f_h)), \tag{2}$$

$$f'_o = \beta * f_o, \tag{3}$$

where $\beta_k \in [0, 1]$ is the k-th interactive value for f_o. Finally, IIM outputs $f'_{ho} = concat(f_h, f'_o)$, which is in same form as inputs. Different from treating human and object features separately, with the formulation above, interactive semantics from humans can be dynamically passed to objects to intensify interactions and promote I point predictions.

D. Correlation Parsing Module

The implied correlations among humans, objects, and interactions, which we called interactive correlations, provide crucial information for HOI inference. For example, the implied correlation between $kick$ and $football$ should be stronger than that between $kick$ and $sandwich$. We propose Correlation Parsing Module (abbr. CPM) upon the Relation-aware Frame Part B to break away from the previous ideas of region-based feature enhancements and exploit the interactive correlations among these targets. The detailed design of CPM is provided in Figure 5 (b). The CPM takes all point head features $f_p \in \Re^{H*W*(1+K+N)}$ as input, where the human class, K object classes and N interaction classes are considered as $(1 + K + N)$ visual targets.

In particular, a project function first provides a pattern to convert all point head feature f_p into a $(1 + K + N)$-wide adjacency matrix A. The calculation can be divided into two parts: feature embedings and a linear combination. Specifically, given the input $X = f_p$, we embed two feature tensors (X^θ, X^ϕ) by $\theta(X)$ and $\phi(X)$ modeled by two $1 * 1$ convolutions. The obtained tensors are then reshaped from $H * W * (1 + K + N)$ to $L * (1 + K + N)$ by the $Planar$ operation, where their two-dimensional location pixels of $H * W$ are converted to one-dimensional vector L. Finally, the following linear combination transforms visual targets into the unified adjacency matrix $A \in \Re^{(1+K+N)*(1+K+N)}$, where $(1 + K + N)$ denotes the length and width of the adjacency matrix. Each element $a_{ij} \in A$ denotes the correlation between two visual targets.

$$A = Planar[\theta(X)]^T * Planar[\phi(X)]. \tag{4}$$

Based on the adjacency matrix, a message spreading process is further adopted to broadcast and integrate interactive correlations from all elements over the matrix, which can be computed by:

$$A' = Conv1D((Conv1D(A)^T)^T \oplus A), \tag{5}$$

where two operations of $Conv1D$ are single-layer 1D convolutions to build communication among all elements. Between the convolutions, the $\oplus$ implements addition point to point which

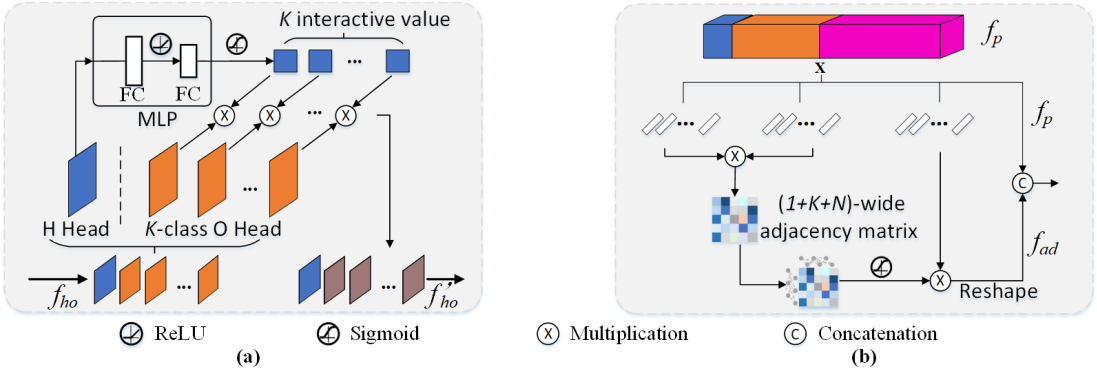

Fig. 5: The detailed design of our proposed (a) Interaction Intensifier Module (abbr. IIM) and (b) Correlation Parsing Module (abbr. CPM). The IIM takes H and O head features as inputs and passes interactive semantics from humans to objects to intensify interactions. The CPM exploits and integrates interactive correlations among visual targets and helps to group HOI triplets without region proposals.

updates the hidden element states according to the incoming information. The A' is the newly updated adjacency matrix with interactive correlations.

The following step provides a reverse projection of A' to convolution space, which outputs a feature tensor f_{ad}. Firstly, a $g(X)$ function and a reshape operation act as linear embedding to compute the representation of input signal as visual targets. Then a weighted linear combination is employed as the reverse projection:

$$Y = sigmoid(A') * Planar[g(X)]. \tag{6}$$

In this way, A' plays a role of dynamical weights so that interactive correlations among visual targets can be weighted aggregated. Finally, we reshape the obtained planar feature tensor $Y \in \Re^{L*(1+K+N)}$ to three-dimensional $f_{ad} \in \Re^{H*W*(1+K+N)}$. For the next displacements prediction, f_{ad} as the correlation encoding is concatenated with appearance feature f_p to provide correlative information among all visual targets.

E. Training and Inference

1) Training: HOI detection needs to generate not only the category of instances and interactions, but also the bounding boxes of instances. Therefore, we also need to regress the size and offset of instance boxes by $G_{wh}(\Delta)$ and $G_{off}(\Delta)$, which have the same computation with $G(\Delta)$ in Section III-B. In summary, the final loss can be calculated as the weighted sum of point prediction loss L_p, displacement prediction loss L_d and instance box regression loss L_r:

$$L = L_p + \lambda L_d + L_r, \tag{7}$$

where λ is set as 0.1 [30], [45]. The reason is that we do not normalize the displacements and directly use the raw pixel coordinates during training. Therefore, the displacement prediction loss L_d needs to be scaled by the constant λ.

As mentioned in Section III-B, three heatmaps can be produced by employing sigmoid upon the H, O and I head features respectively. We also map ground-truth H, O, and I points into three heatmaps with a Gaussian kernel respectively. Therefore, point detection can be transformed into a heatmap estimation problem, where $L_p = L_{ho} + L_i$ are optimized by element-wise focal loss [64]. For example, with predicted heatmap $M_i \in [0,1]^{H*W*N}$ and ground-truth heatmap $\tilde{M}_i \in [0,1]^{H*W*N}$ of interaction points, L_i is computed by:

$$L_i = -\frac{1}{U}\sum_{\forall n(x,y)} \begin{cases} (1-S^i_{nxy})^{\alpha} log(S^i_{nxy}), & if \tilde{S}^i_{nxy} = 1 \\ (1-\tilde{S}^i_{nxy})^{\gamma}(S^i_{nxy})^{\alpha} & otherwise \\ log(1-S^i_{nxy}), & \end{cases} \tag{8}$$

where U denotes the number of point location (x, y) on I head feature map, and $n \in N$ is the category of interaction points. S^i_{nxy} and $\tilde{S}^i_{nxy}$ are the predicted score and ground-truth label for the n-th interaction category at location (x, y), respectively. The α and γ are set as 2 and 4 following default setting [30].

The $L_d = L_{dh} + L_{do}$ is trained by $L1$ loss, where L_{dh} and L_{do} are losses for predicted displacement $dp^{ih} = (dp^{ih}_x, dp^{ih}_y)$ to H point and displacement $dp^{io} = (dp^{io}_x, dp^{io}_y)$ to O point respectively for an I point. Taking L_{dh} as an example,

$$L_{dh} = -\frac{1}{Q}\sum_{\forall(\tilde{x}^i,\tilde{y}^i)} |dp^{ih}_x - \tilde{dp}^{ih}_x| + |dp^{ih}_y - \tilde{dp}^{ih}_y|, \tag{9}$$

where $\tilde{dp}^{ih} = (\tilde{dp}^{ih}_x, \tilde{dp}^{ih}_y)$ is the ground-truth displacement from the I point $(\tilde{x}^i, \tilde{y}^i)$ to the corresponding H point $(\tilde{x}^h, \tilde{y}^h)$. Q denotes the number of ground-truth I points.

The $L_r = L_{wh} + L_{off}$ is optimized by $L1$ loss, which is in charge of regressing box size (h, w) and local offset $d = (d_x, d_y)$ for location of ground truth instance points.

$$L_{wh} = -\frac{1}{U+V}\sum_{\forall(\tilde{x}^h,\tilde{y}^h),(\tilde{x}^o,\tilde{y}^o)} |(h-\tilde{h})| + |(w-\tilde{w})|, \tag{10}$$

$$L_{off} = -\frac{1}{U+V}\sum_{\forall(\tilde{x}^h,\tilde{y}^h),(\tilde{x}^o,\tilde{y}^o)} |(d_x-\tilde{d}_x)| + |(d_y-\tilde{d}_y)|, \tag{11}$$

where $(\tilde{h}, \tilde{w})$ and $(\tilde{d}_x, \tilde{d}_y)$ are the width and height, as well as local offset for a ground-truth instance point. It is worth mentioning that the ground-truth instance point offset $(\tilde{d}_x, \tilde{d}_y)$ is caused by the process of projecting point location from original image to instance head feature map. $U + V$ denotes the number of ground-truth instance points.

In this way, the entire network is implemented to be fully differentiable and end-to-end trainable using gradient-based optimization.

2) Inference: After points detection, we select top T human, object and interaction points according to the corresponding confidence scores respectively. Then, point grouping process will determine the optimal triplets. Specifically, for each detected I point (x^i, y^i) and predicted displacements (dp_x^{ih}, dp_y^{ih}) and (dp_x^{io}, dp_y^{io}), several candidate instance points will be ranked according to two criteria, which are 1) have high confidence score, 2) be close to the human/object location that generated by interaction point location plus the human/object displacement. In words, the smaller its grouping distance, the more competitive it is. Taking the best candidate H point (x^h, y^h) for I point (x^i, y^i) as an example, its grouping distance D_{group}^{ih} can be defined as follow:

$$D_{group}^{ih} = \frac{1}{S_{(x^h,y^h)}^h} |(x^i, y^i) + (dp_x^{ih}, dp_y^{ih}) - (x^h, y^h)|, \quad (12)$$

where $S_{(x^h,y^h)}^h$ is the confidence score of candidate H point (x^h, y^h). The best candidate O point (x^o, y^o) will be obtained in same way.

For a grouped regressive H point (x^h, y^h) and an O point (x^o, y^o), their bounding boxes are determined as $B_h = (x^h, y^h, h^h, w^h)$ and $B_o = (x^o, y^o, h^o, w^o)$, where (h^h, w^h) and (h^o, w^o) denote the width and height of the human box and object box, respectively. With a set of detected HOI triplets, we use a fusion of scores to predict the final score S_{hoi} for each of them:

$$S_{hoi} = S_{(x^h,y^h)}^h * S_{(x^i,y^i)}^i * S_{(x^o,y^o)}^o, \quad (13)$$

where $S_{(x^h,y^h)}^h$, $S_{(x^i,y^i)}^i$ and $S_{(x^o,y^o)}^o$ are confidence scores of candidate H point, I point and O point.

IV. Experiments and Evaluations

In this section, we first outline our experimental setup (Section IV-A), including datasets, evaluation metrics and implementation details. We then report our overall quantitative results (Section IV-B) on two commonly used large-scale HOI detection benchmarks and compare the performance with the state-of-the-art methods. Next, we conduct ablation studies and detailed analysis (Section IV-C) to quantify the contributions from individual components. Finally, qualitative detection examples are provided to verify our method (Section IV-D) and limitations (Section IV-E) are discussed to facilitate future research.

A. Experimental Setup

1) Datasets and Evaluation Metrics: We evaluate our method and compare it with the state-of-the-arts on two commonly used large-scale benchmarks, including V-COCO [38] and HICO-DET [39] datasets. V-COCO [38] is a subset of MS COCO dataset [65], which includes 10,346 images. It contains 16,199 human instances in total and provides 26 common verb categories. HICO-DET [39] contains 47,776 images, where 80 object categories and 117 verb categories compose of 600 HOI categories. It provides more than 150K annotated ⟨ human, object ⟩ pairs. There are three different HOI category sets in HICO-DET, which are: (a) all 600 HOI categories (Full), (b) 138 HOI categories with less than 10 training instances (Rare), and (c) 462 HOI categories with 10 or more training instances (Non-Rare). A detected ⟨ human, verb, object ⟩ triplet is considered as a true positive if: 1) it has the correct interaction label; 2) both the predicted human and object bounding boxes have IoU of 0.5 or higher with the ground-truth boxes. Following the standard protocols, we use role mean average precision (mAP_{role}) [38] to report evaluation results.

2) Implementation details: During training, input images have the resolution of $512 * 512$, yielding a resolution of $128 *$

TABLE I: Performance comparison with state-of-the-arts on V-COCO dataset.

Method	Backbone	mAP$_{role}$(%)
Gupta et al. [38]	ResNet-50-FPN	31.8
InteractNet [21]	ResNet-50-FPN	40.0
iHOI [23]	ResNet-50	40.4
BAR-CNN [31]	Inception-ResNet	41.1
GPNN [28]	DeformableCNN	44.0
iCAN [22]	ResNet-50	45.3
Contextual Att [57]	ResNet-50	47.3
RPNN [35]	ResNet-50	47.5
TIN(RP_dC_d) [43]	ResNet-50	47.8
C-HOI [26]	ResNet-50	48.9
CP-HOI [27]	ResNet-50	50.4
DRG [59]	ResNet-50-FPN	51.0
IPNet [44]	Hourglass-104	51.0
VSGNet [25]	ResNet-152	51.8
PMFNet [34]	ResNet-50	52.0
CHGN [61]	ResNet-50	52.7
FCMNet [47]	ResNet-50	<u>53.1</u>
our baseline	DLA-34	51.4
RR-Net	DLA-34	**55.3**

TABLE II: Performance comparison on HICO-DET test set. Default: all images. Full: all 600 HOI categories. Rare: 138 HOI categories with less than 10 training instances. Non-Rare: 462 HOI categories with 10 or more training instances.

Method	Backbone	Default		
		full	rare	non-rare
Shen et al. [66]	VGG-19	6.46	4.24	7.12
HO-RCNN [39]	CaffeNet	7.81	5.37	8.54
InteractNet [21]	ResNet-50-FPN	9.94	7.16	10.77
iHOI [23]	ResNet-50	9.97	7.11	10.83
GPNN [28]	DeformableCNN	13.11	9.34	14.23
iCAN [22]	ResNet-50	14.84	10.45	16.15
Contextual Att [57]	ResNet-50	16.24	11.16	17.75
RPNN [35]	ResNet-50	17.35	12.78	18.71
TIN(RP_dC_d) [43]	ResNet-50	17.03	13.42	18.11
No-Frills [24]	ResNet-50	17.18	12.17	18.68
PMFNet [34]	ResNet-50	17.46	15.65	18.00
CHGN [61]	ResNet-50-FPN	17.57	16.85	17.78
C-HOI [26]	ResNet-50	18.89	13.51	20.19
DRG [59]	ResNet-50-FPN	19.26	**17.74**	19.71
CP-HOI [27]	ResNet-50	19.42	13.98	20.91
IPNet [44]	Hourglass-104	19.56	12.79	21.58
VSGNet [25]	ResNet-152	19.80	16.05	20.91
PPDM [45]	DLA-34	20.29	13.06	<u>22.45</u>
FCMNet [47]	ResNet-50	<u>20.41</u>	<u>17.34</u>	21.56
our baseline	DLA-34	18.87	11.32	21.13
RR-Net	DLA-34	**20.96**	13.43	**23.21**

128 for all output head features. We employ standard data augmentation techniques (e.g., random scale, random shift) [30]. Our model is optimized with Adam [67]. The batch-size is set as 15 (on a single GPU) for V-COCO benchmark and 20 (on a single GPU) for HICO-DET benchmark. Following general experience [30], [45] and grid experiments, we train the model for 140 epochs, with the initial learning rate of 5e-4 which drops 10x at 90 and 120 epochs respectively. The number of top predictions T is set as 100. All our experiments are conducted by Pytorch on GPU of NVIDIA Tesla P100.

B. Overall Performance

We compare our method with several state-of-the-arts in this subsection. Meanwhile, we strip all components related to the Relation-aware Frame structure, Interaction Intensifier Module and Correlation Parsing Module as the baseline.

1) Performance on V-COCO: Comparison results on V-COCO in terms of mAP_{role} are shown in Table I. It can be seen that our proposed RR-Net has a mAP(%) of 55.3, obtaining the best performance among all methods. Although we do not adopt previous region-based feature learning (e.g., Contextual Att [57], VSGNet [25], CHGN [61], FCMNet [47]), or employ additional human pose (e.g., TIN [43], PMFNet [34]), our method outperforms these methods with sizable gains. Besides, our method achieves an absolute gain of 3.9 points, which is a relative improvement of 7.6% compared with the baseline, validating its efficacy in HOI detection task.

2) Performance on HICO-DET: Table II shows the comparisons of RR-Net with state-of-the-arts on HICO-DET. We report the quantitative evaluation settings of Full, Rare, and Non-Rare. Firstly, the detection results of our RR-Net are the best among all methods under the evaluation settings of Full and Non-Rare, demonstrating that our method is more competitive than the others in detecting most common HOIs. It is noted that our method is not preeminent in detecting rare HOIs (HOI categories with less than 10 training instances), because our relation reasoning modules need enough training samples to exert their effects. Besides, our RR-Net obtains 20.96 mAP on HICO-DET (Defualt Full), which achieves absolute gain of 2.1 point and relative gain of 11.1% compared with the baseline. These results quantitatively show the efficacy of our method.

3) Computational Complexity: When it comes to efficiency, our method provides its advantages over the methods based on region-based attention features [21], [22], [28] or additional inputs [34], [43] (e.g., estimated human pose). It is because region-based models are computationally expensive, and estimating human pose or body-parts takes extra time. However, our method captures HOI-specific representations where no human or object proposals are needed. We compare the inference time with the methods which have reported their speed on V-COCO. As shown in the table III, RR-Net is faster than most methods by a large margin, with the speed of 34.36 fps. Therefore, our method can meet the real-time requirements of applications. Meanwhile, the detection speed of our RR-Net is basically identical with the baseline. However, our RR-Net brings a relative improvement of 7.6% in detection performance.

TABLE III: Comparison of inference speed with state-of-the-arts.

Method	Inference Time (ms)	FPS	mAP$_{role}$(%)
InteractNet [21]	145	6.90	40.0
GPNN [28]	245	4.08	44.0
iCAN [22]	204	4.09	45.3
TIN(RP_dC_d) [43]	513	1.95	47.8
PMFNet [34]	253	3.95	52.0
our baseline	27	36.74	51.4
RR-Net(ours)	29	34.36	**55.3**

TABLE IV: Impact of adopting progressive Relation-aware Frame. R-Frame in the table is short for the Relation-aware Frame.

Method	mAP$_{role}$(%)
baseline	51.4
baseline+R-Frame Part A	52.7
baseline+R-Frame Part B	52.5
baseline+Fully R-Frame	53.1

C. Ablation Studies

1) Performance Impact of Proposed Components: We conduct several ablation studies in this subsection. V-COCO [38] serves as the testbed on which we further analyze the individual effect of components in our method.

We first study the performance impact of our proposed progressive Relation-aware Frame, Interaction Intensifier Module (IIM) and Correlation Parsing Module (CPM). As shown in Table IV, we perform an ablation study on Relation-aware Frame by dividing it into two parts. When we adopt the Relation-aware Frame Part A or Part B severally (See Figure 3), the mAP are 52.7 and 52.5, indicating the respective effects of the two parts. Adopting the fully Relation-aware Frame obtains the mAP of 53.1, which boosts the performance by 1.7 points compared with the baseline. Experimental analysis on proposed relation function group $G()$ is also given in Table V. We strip the global relation modeling out of the relation function group $G()$ to investigate its impact. As shown in the Table V, both Relation-aware Frame Part A and Part B benefit from the global relation modeling. Employing the global relation modeling in the fully Relation-aware Frame boosts the performance by 0.4 point. These results show that the proposed structure is effective.

Meanwhile, we investigate the impact of employing IIM and CPM in Table VI and Table VII . Compared with the baseline, employing IIM (upon Relation-aware Frame Part A) and CPM (upon Relation-aware Frame Part B) brings gains of 2.0 mAP and 1.6 mAP, respectively. Furthermore, improvements of 3.0 mAP and 2.8 mAP can be obtained while employing IIM and CPM respectively upon our fully Relation-aware Frame. Finally, RR-Net achieves an absolute gain of 3.9 mAP compared with the baseline. These evidences show that each proposed component of our region-independent relation reasoning indeed contributes to the final performance.

2) Analysis of Relation-aware Frame: We perform visualization experiments in this subsection to give in-depth analysis of Relation-aware Frame. We show several examples in Figure 6 to illustrate how the Relation-aware Frame works, where different $\langle$ human, verb, object $\rangle$ triplets are displayed to

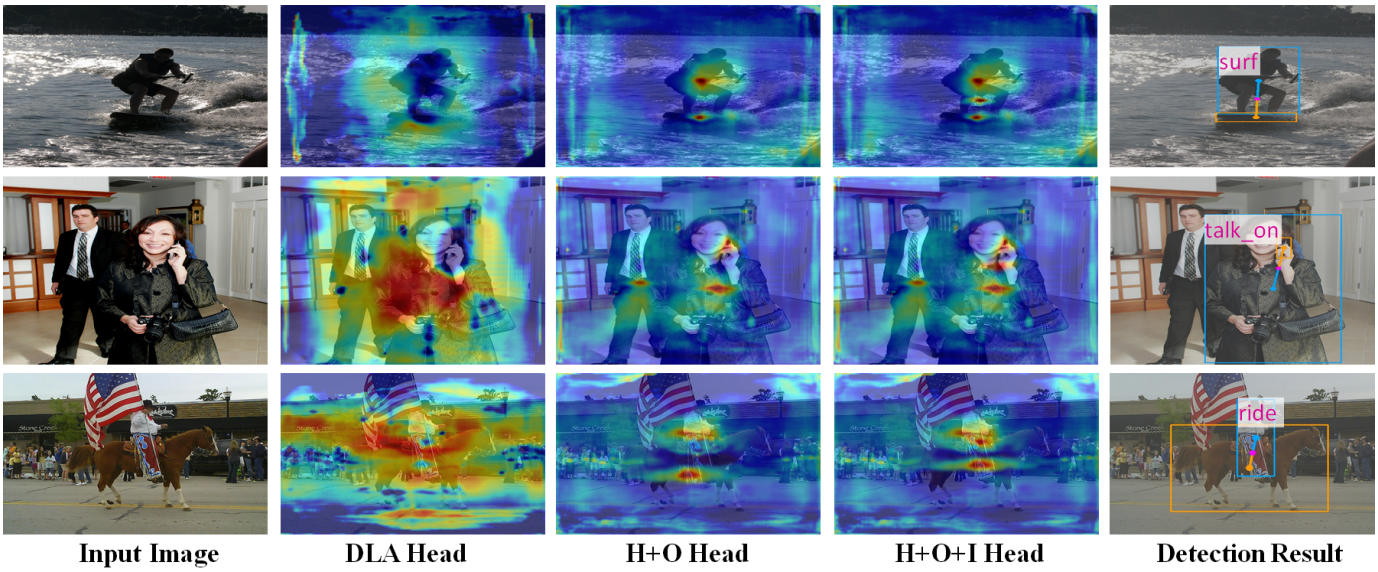

Fig. 6: Analysis of Relation-aware Frame. Each column from left to right is the input images, visualized focused area from DLA head features, H+O head features, concatenation of all point head features and final detection results. Compared with DLA head features, our developed H+O head features and H+O+I head features are more effective in predicting I points and displacements. Best viewed in color.

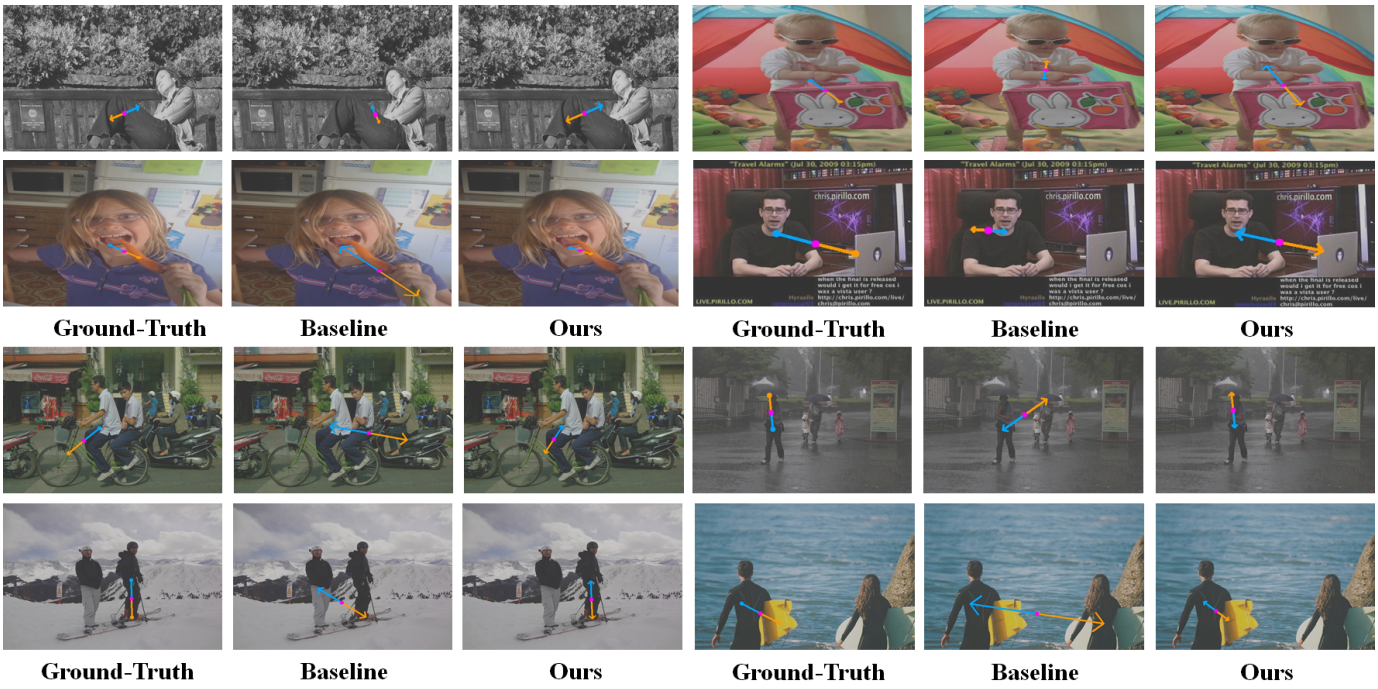

Fig. 7: Visualized effects of IIM and CPM. The first two rows show that our framework with IIM and CPM makes better predictions of interaction point and displacements compared with the baseline. The third and fourth rows illustrate that our method effectively prevents potential miss-match of ⟨ human, object ⟩ pairs from the baseline method. Best viewed in color.

Fig. 8: Visualization of HOI detection. The first two rows show our results compared with baseline, where texts below indicate the detected interaction (verb), and two numbers in turn represent scores predicted by baseline and our method. The rest of rows show multiple people take interactions with various objects concurrently detected by our method. Best viewed in color.

TABLE V: Analysis on relation function group $G()$. R-Frame in the table is short for the Relation-aware Frame.

Method	$mAP_{role}(\%)$		
	Part A	Part B	Fully R-Frame
w/o global relation modeling	52.4	52.2	52.7
w/ global relation modeling	52.7	52.5	53.1

TABLE VI: Impact of employing Interaction Intensifier Module (IIM).

Baseline	√		
Relation-aware Frame Part A		√	√
Relation-aware Frame Part B			√
IIM		√	√
$\text{mAP}_{role}(\%)$	51.4	53.4	54.4

TABLE VII: Impact of employing Correlation Parsing Module (CPM).

Baseline	√		
Relation-aware Frame Part A			√
Relation-aware Frame Part B		√	√
CPM		√	√
$\text{mAP}_{role}(\%)$	51.4	53.0	54.2

prove its practicality (i.e., ⟨ human, surf, surfboard ⟩, ⟨ human, talk on, phone ⟩ and ⟨ human, ride, horse ⟩). The first and last columns of Figure 6 show input images and detection results respectively, the second to fourth columns visualize the most focused area from each head feature.

- Q1: In Relation-aware Frame, how H+O head features better than DLA head features to build I head for predicting I points? Meanwhile, how concatenation of all point head features better than DLA head features to build D head for predicting displacements?

The contexts and locations from the human and object can provide crucial information to locate and classify the interaction. As we can see from the Figure 6, DLA head features in the second column do not always pay attention to the contexts and locations of human and object, so it may not be able to effectively infer where and what the interaction occurs.

Nevertheless, H and O head features in third column imply attention on human and object, which are conducive for I point prediction. In short, in Relation-aware Frame Part A, building I head upon H and O head features helps to predict I points with better utilization of human and object contexts and locations. Similarly, as shown in the forth column of Figure 6, concatenating I head features further strengthens the attention on interactions. Thus, we build displacement head (D head) upon concatenation of all point heads in Relation-aware Frame Part B, so that contexts and locations of H, O and I points can be implicitly used for predicting displacements from I point to H and O points.

3) Analysis of IIM and CPM: We give more visual evidence in this subsection to illustrate effects of IIM and CPM.

- Q2: How do IIM and CPM effect HOI detection?

Not like region-based feature enhancements, IIM and CPM are region-independent relation reasoning modules. To better predict category and location of an interaction, the IIM intensifies response between human and different objects by building a channel from human features to different object features. The CPM learns interactive correlations among humans, objects and interactions so that these targets with various categories and locations can be better matched to each other. We take several examples of interaction point and displacements predictions before point grouping in Figure 7 to visualize the effects of IIM and CPM. The first column of each example shows the ground-truth annotation. The second column visualizes interaction point and displacements predictions before point grouping using baseline method. The third column represents above predictions with IIM and CPM. In the first two rows, with IIM and CPM, our framework intuitively makes better predictions of interaction points and displacements compared with baseline, which verifies the necessity of relation reasoning. It is worth noting that due to the inaccurate prediction of interaction points and displacements, the baseline framework in the third and fourth rows may lead to miss-match between the humans and objects. However, our framework provides predictions that are close to the ground-truth annotations, which helps to produce correct HOI detection in the point grouping procedure.

D. Qualitative Examples

Several examples of detection are given in Figure 8. We highlight the detected humans, objects and associated interactions for better view. We first compare our results with baseline in the first two rows to demonstrate our improvements. Each subplot displays one detected $\langle$ human, verb, object $\rangle$ triplet for easy observation, including the location of human and object, as well as the interaction between the above two instances. We can see that our method is capable of detecting various HOIs with higher scores. In addition, the rest of rows show that our method is able to detect multiple people taking different interactions with diversified objects. Summarizing all above results validates that our proposed method can learn HOI-specific representations well from complex environments to provide high quality HOI detection.

E. Limitations

Although we have demonstrated good performance, our model is not perfect. We discuss the limitations of our method below, with some incorrect detection examples in Figure 9. First, our relation reasoning modules are built after generating H and O head feature. Therefore, category and location prediction of H and O points can not benefit from the rich relation and interactive information captured by our proposed modules. As shown in the first subplot of Figure 9, the incorrect detection of $phone$ leads to a wrong predicted HOI label of $talk_on$. Miss-localization of bed in the second subplot also causes failure case of HOI detection. We believe that employing better backbone networks and further exploiting relation reasoning in procedure of producing H and O head feature may be helpful to reduce this type of errors. Second, our method is not preeminent in detecting rare HOIs (HOI categories with less than 10 training instances). In the third and fourth subplots of Figure 9, our method does not return HOI predictions corresponding to the annotations since $sign$ and dry are rare in training data. This phenomenon illustrates that improving rare HOI detection is still an interesting open problem. Mitigating this problem will be the content of our next research.

V. Conclusion

Recent HOI feature learning techniques focus on region-based relation modeling or feature enhancements, which are computationally expensive and hardly applicable to end-to-end detectors in real applications. Therefore, we make a first try to explore novel techniques of region-independent relation reasoning, which opens up a new direction to improve HOI detection. Specifically, a Relation-aware Frame is proposed to provide a coherent structure for HOI inference, upon which an Interaction Intensifier Module and a Correlation Parsing Module are further developed to reason HOI-specific representations without instance regions. Extensive experiments show that our proposed RR-Net outperforms existing methods with sizable gaps, validating its efficacy in detecting HOIs. In future studies, we will further explore relation reasoning techniques and pay more attention to improve our method in detecting rare HOIs.

Acknowledgment

This paper was partially supported by the IER foundation (No. HT-JD-CXY-201904) and Shenzhen Municipal Development and Reform Commission (Disciplinary Development Program for Data Science and Intelligent Computing). Special acknowledgements are given to AOTO-PKUSZ Joint Research Center of Artificial Intelligence on Scene Cognition technology Innovation for its support.

References

[1] Y. Ji, Y. Yang, F. Shen, H. T. Shen, and X. Li, "A survey of human action analysis in hri applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 2114–2128, 2020.

[2] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

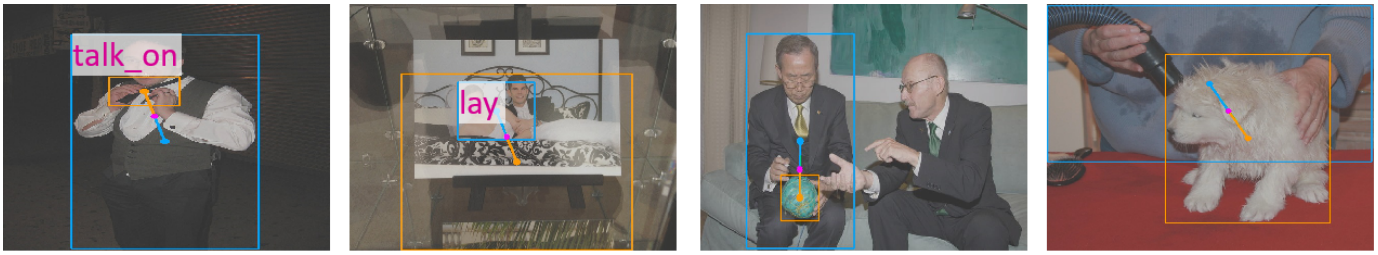

Fig. 9: Examples of incorrect detection from our method. The first two subplots show incorrect detection related to falsely recognizing objects (i.e., hallucination of *phone* and miss-localization of *bed*). The third and fourth subplots show incorrect detection of rare HOIs (no recall from the method).

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[4] D. Yang, Y. Zou, J. Zhang, and G. Li, "C-rpns: Promoting object detection in real world via a cascade structure of region proposal networks," *Neurocomputing*, vol. 367, pp. 20 – 30, 2019.

[5] D. Zhang, W. Zeng, J. Yao, and J. Han, "Weakly supervised object detection using proposal-and semantic-level relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[6] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[7] L. Zhou, J. Zhao, J. Li, L. Yuan, and J. Feng, "Object relation detection based on one-shot learning," *arXiv preprint arXiv:1807.05857*, 2018.

[8] Y. Jiao, H. Yao, and C. Xu, "Pen: Pose-embedding network for pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1150–1162, 2021.

[9] Q. Wang, P. Zhang, H. Xiong, and J. Zhao, "Face. evolve: A high-performance face recognition library," *arXiv preprint arXiv:2107.08621*, 2021.

[10] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3684–3692.

[11] X. Zhang, Y. Chen, B. Zhu, J. Wang, and M. Tang, "Part-aware context network for human parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8971–8980.

[12] Z. Tan, B. Liu, Q. Chu, H. Zhong, Y. Wu, W. Li, and N. Yu, "Real time video object segmentation in compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 175–188, 2021.

[13] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2334–2343.

[14] S. Liu, Y. Li, and G. Hua, "Human pose estimation in video via structured space learning and halfway temporal evaluation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 2029–2038, 2019.

[15] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng, "Multiple-human parsing in the wild," *arXiv preprint arXiv:1705.07206*, 2017.

[16] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, "Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 792–800.

[17] J. Zhao, J. Li, H. Liu, S. Yan, and J. Feng, "Fine-grained multi-human parsing," *International Journal of Computer Vision*, vol. 128, no. 8, pp. 2185–2203, 2020.

[18] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *international conference on learning representations*, 2015.

[19] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 15–28, 2019.

[20] J. Kong, H. Deng, and M. Jiang, "Symmetrical enhanced fusion network for skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[21] G. Gkioxari, R. B. Girshick, P. Dollar, and K. He, "Detecting and recognizing human-object interactions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8359–8367, 2018.

[22] C. Gao, Y. Zou, and J. Huang, "ican: Instance-centric attention network for human-object interaction detection." *british machine vision conference*, p. 41, 2018.

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

[24] T. Gupta, A. Schwing, and D. Hoiem, "No-frills human-object interaction detection: Factorization, layout encodings, and training techniques," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019, pp. 9677–9685.

[25] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "Vsgnet: Spatial attention network for detecting human object interactions using graph convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[26] T. Zhou, W. Wang, S. Qi, H. Ling, and J. Shen, "Cascaded human-object interaction recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4263–4272.

[27] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, "Cascaded parsing of human-object interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[28] S. Qi, W. Wang, B. Jia, J. Shen, and S. Zhu, "Learning human-object interactions by graph parsing neural networks," *Proceedings of the European Conference on Computer Vision*, pp. 407–423, 2018.

[29] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[30] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," in *arXiv preprint arXiv:1904.07850*, 2019.

[31] A. Kolesnikov, A. Kuznetsova, C. Lampert, and V. Ferrari, "Detecting visual relationships using box attention," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[32] H. Fan, T. Zhuo, X. Yu, Y. Yang, and M. Kankanhalli, "Understanding atomic hand-object interaction with human intention," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[33] X. Zhang, Z. Wu, and Y.-G. Jiang, "Sam: Modeling scene, object and action with semantics attention modules for video recognition," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[34] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9469–9478.

[35] P. Zhou and M. Chi, "Relation parsing neural network for human-object interaction detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 843–851.

[36] W. Wang, T. Zhou, S. Qi, J. Shen, and S.-C. Zhu, "Hierarchical human semantic parsing with comprehensive part-relation modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[37] H. Wang, B. Yu, J. Li, L. Zhang, and D. Chen, "Multi-stream interaction networks for human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2021.

[38] M. Yatskar, L. Zettlemoyer, and A. Farhadi, "Situation recognition: Visual semantic role labeling for image understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5534–5542.

[39] Y. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," *workshop on applications of computer vision*, pp. 381–389, 2018.

[40] D. Yang, Y. Zou, C. Zhang, M. Cao, and J. Chen, "Rr-net: Injecting interactive semantics in human-object interaction detection," *arXiv preprint arXiv:2104.15015*, 2021.

[41] S. Zhang, C. Bauckhage, D. A. Klein, and A. B. Cremers, "Exploring human vision driven features for pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1709–1720, 2015.

[42] J.-F. Hu, W.-S. Zheng, J. Lai, S. Gong, and T. Xiang, "Exemplar-based recognition of human–object interactions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 647–660, 2016.

[43] Y.-L. Li, S. Zhou, X. Huang, L. Xu, Z. Ma, H.-S. Fang, Y. Wang, and C. Lu, "Transferable interactiveness knowledge for human-object interaction detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3585–3594.

[44] W. Tiancai, Y. Tong, D. Martin, S. K. Fahad, Z. Xiangyu, and S. Jian, "Learning human-object interaction detection using interaction points," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

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

[46] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu, "Pairwise body-part attention for recognizing human-object interactions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 51–67.

[47] Y. Liu, Q. Chen, and A. Zisserman, "Amplifying key cues for human-object-interaction detection," in *Proceedings of the European conference on computer vision*. Springer, 2020, pp. 248–265.

[48] R. Gao, B. Xiong, and K. Grauman, "Im2flow: Motion hallucination from static images for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5937–5947.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, p. 6000–6010.

[50] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

[51] M. Yin, Z. Yao, Y. Cao, X. Li, Z. Zhang, S. Lin, and H. Hu, "Disentangled non-local neural networks," in *Proceedings of the European conference on computer vision*. Springer, 2020, pp. 191–207.

[52] J. Li, P. Zhou, Y. Chen, J. Zhao, S. Roy, Y. Shuicheng, J. Feng, and T. Sim, "Task relation networks," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 932–940.

[53] F. Zhao, J. Zhao, S. Yan, and J. Feng, "Dynamic conditional networks for few-shot learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–35.

[54] J. Li, Y. Chen, S. Xiao, J. Zhao, S. Roy, J. Feng, S. Yan, and T. Sim, "Estimation of affective level in the wild with multiple memory networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 1–8.

[55] J. Zhao12, J. Li, F. Zhao, S. Yan13, and J. Feng, "Marginalized cnn: Learning deep invariant representations," 2017.

[56] R. Girdhar and D. Ramanan, "Attentional pooling for action recognition," *neural information processing systems*, pp. 34–45, 2017.

[57] T. Wang, R. M. Anwer, M. H. Khan, F. S. Khan, Y. Pang, L. Shao, and J. Laaksonen, "Deep contextual attention for human-object interaction detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5694–5702.

[58] D. Yang, Y. Zou, J. Zhang, and G. Li, "Gid-net: Detecting human-object interaction with global and instance dependency," *Neurocomputing*, 2020.

[59] C. Gao, J. Xu, Y. Zou, and J.-B. Huang, "Drg: Dual relation graph for human-object interaction detection," in *Proceedings of the European conference on computer vision*. Springer, 2020, pp. 696–712.

[60] D. Yang and Y. Zou, "A graph-based interactive reasoning for human-object interaction detection," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 1111–1117.

[61] H. Wang, W.-s. Zheng, and L. Yingbiao, "Contextual heterogeneous graph network for human-object interaction detection," in *Proceedings of the European conference on computer vision*. Springer, 2020, pp. 248–264.

[62] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[63] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[64] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

[65] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[66] L. Shen, S. Yeung, J. Hoffman, G. Mori, and L. Fei-Fei, "Scaling human-object interaction recognition through zero-shot learning," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1568–1576.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Dongming Yang received the B.E. degree from Shan Xi University, in 2015, and the M.Sc. degree from University of Chinese Academy of Sciences, in 2018. He is currently pursuing the Ph.D. degree in computer science and engineering from Peking University. His research interests include computer vision and pattern recognition. He is now an intern in the Peng Cheng Laboratory.

Yuexian Zou received her B.Sc. degree from the University of Electronic Science and Technology in 1985 and Ph.D. degree from the University of Hong Kong in 2001, respectively. She is currently a Full Professor with Peking University and the Director of the Advanced Data and Signal Processing Laboratory in Peking University Shenzhen Graduate School. She was a recipient of the award Leading Figure for Science and Technology by Shenzhen Municipal Government in 2009 and now is the adjunct professor in Peng Cheng Laboratory. She also serves as the Deputy Director of Shenzhen Association of Artificial Intelligence (SAAI). Since 2010, she has been actively involved in teaching and research on machine learning and its applications in video and audio analysis. She conducted more than 20 research projects including NSFC and 863 projects. She has published more than 200 academic papers in famous journals and flagship conferences, issued five invention patents, and two of them have been transferred to a company. She conducts several courses for graduate students, such as machine learning and pattern recognition, digital signal processing, and array signal processing. Her research interests mainly in machine learning for signal processing and scene understanding.More information can be referred to http://web.pkusz.edu.cn/adsp/

Can Zhang Can Zhang received the B.Eng. degree in Electronic Engineering from Shanghai University, Shanghai, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Electronic and Computer Engineering (ECE), Peking University, Shenzhen, China. His current research interests include computer vision, multimedia computing and video understanding.

Meng Cao received the B.E. degree from the Huazhong University of Science and Technology (HUST), in 2018. He is currently pursuing the M.Sc. degree with the School of Electronic and Computer Engineering (ECE), Peking University. His current research interests include computer vision and computer graphics.

Jie Chen received the MSc and PhD degrees from the Harbin Institute of Technology, China, in 2002 and 2007, respectively. He joined the faculty with the Graduate School in Shenzhen, Peking University, in 2019, where he is currently an associate professor with the School of Electronic and Computer Engineering, Peking University. Since 2018, he has been working with the Peng Cheng Laboratory, China. From 2007 to 2018, he worked as a senior researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. In 2012 and 2015, he visited the Computer Vision Laboratory, University of Maryland and School of Electrical and Computer Engineering, Duke University respectively. He was a cochair of International Workshops at ACCV, CVPR, ICCV and ECCV. He was a guest editor of special issues for IEEE TPAMI, IJCV and Neurocomputing. His research interests include deep learning, computer vision, and medical image analysis. He is an Associate Editor of the Visual Computer. He is a member of the IEEE.