Cascaded Human-Object Interaction Recognition

Tianfei Zhou[1]*, Wenguan Wang[2]*, Siyuan Qi[3], Haibin Ling[4], Jianbing Shen[1]†
[1]Inception Institute of Artificial Intelligence, UAE [2]ETH Zurich, Switzerland [3]Google, USA [4]Stony Brook University, USA
{ztfei.debug,wenguanwang.ai}@gmail.com
https://github.com/tfzhou/C-HOI

Abstract

Rapid progress has been witnessed for human-object interaction (HOI) recognition, but most existing models are confined to single-stage reasoning pipelines. Considering the intrinsic complexity of the task, we introduce a cascade architecture for a multi-stage, coarse-to-fine HOI understanding. At each stage, an instance localization network progressively refines HOI proposals and feeds them into an interaction recognition network. Each of the two networks is also connected to its predecessor at the previous stage, enabling cross-stage information propagation. The interaction recognition network has two crucial parts: a relation ranking module for high-quality HOI proposal selection and a triple-stream classifier for relation prediction. With our carefully-designed human-centric relation features, these two modules work collaboratively towards effective interaction understanding. Further beyond relation detection on a bounding-box level, we make our framework flexible to perform fine-grained pixel-wise relation segmentation; this provides a new glimpse into better relation modeling. Our approach reached the 1^{st} place in the ICCV2019 Person in Context Challenge, on both relation detection and segmentation tasks. It also shows promising results on V-COCO.

1. Introduction

Human-object interaction (HOI) recognition aims to identify meaningful ⟨*human, verb, object*⟩ triplets from images, such as ⟨*human, eat, carrot*⟩ in Fig. 1. It plays a crucial role in many vision tasks, *e.g.*, visual question answering [36, 29, 54], human-centric understanding [46, 47, 56], image generation [24], and activity recognition [41, 48, 9, 37, 35], to name a few representative ones.

Though great advances have been made recently, the task is still far from being solved. One of the main challenges comes from its intrinsic complexity: a successful HOI recognition model must accurately 1) localize and recognize each interacted entity (*human*, *object*), and 2) predict the interaction classes (*verb*). Both subtasks are difficult, leading to HOI recognition itself a highly complex problem. With a broader view of other computer vision and machine learning related fields, coarse-to-fine and cascade inference have been shown to deal well with complex problems [26, 11, 12, 45]. The central idea is to leverage sequences of increasingly fine approximations to control the complexity of learning and inference. This motivates us to propose a cascade HOI recognition model, which builds up multiple stages of neural network inference in an annealing-style. For the two subtasks of instance localization and interaction recognition, this model arranges them in a successive manner within each single stage, and carries out cascade, cross-stage inference for each. Above designs result in a multi-task, coarse-to-fine inference framework, which enables asymptotically improved HOI representation learning. This also distinctively differentiates our method from previous efforts, which rely on single-stage architectures.

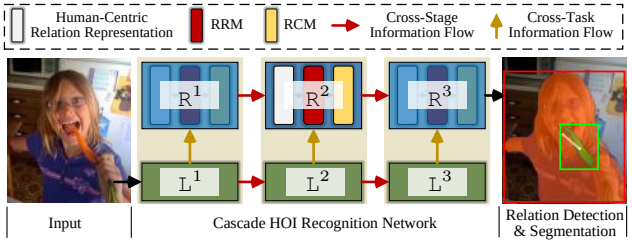

Figure 1: **Illustration of our cascade HOI recognition network,** which is able to handle both object-level relation detection and pixel-wise relation segmentation tasks. Given an input image, our model performs coarse-to-fine inference over both instance localization ($\mathtt{L}^1 \sim \mathtt{L}^3$) and interaction recognition ($\mathtt{R}^1 \sim \mathtt{R}^3$).

As shown in Fig. 1, our model consists of an instance localization network and an interaction recognition network, both working in a cascade manner. Through the instance localization network, the model step-by-step increases the selectiveness of the instance proposals. With such progressively refined HOI candidates, as well as the useful relation representation from the preceding stage, better ac-

*The first two authors contribute equally to this work.
†Corresponding author: *Jianbing Shen*.

tion predictions can be achieved by current-stage interaction recognition network. Moreover, in the interaction recognition network, both human semantics and facial patterns are mined to boost relation reasoning, as these cues are tied to underlying purposes of human actions. With such human-centric features, a relation ranking module (RRM) is proposed to rank all the possible *human-object* pairs. Only the top-ranked, high-quality candidates are fed into a relation classification module (RCM) for final *verb* prediction.

More essentially, previous HOI literature mainly address *relation detection*, *i.e.*, recognizing HOIs at a bounding-box level. In addition to addressing this classic setting, we take a further step towards more fine-grained HOI understanding, *i.e.*, identifying the relations between interacted entities at the pixel level (see Fig. 1). Studying such *relation segmentation* setting not only further demonstrates the efficacy and flexibility of our cascade framework, but allows us to explore more powerful relation representations. This is because bounding box based representations only encode coarse object information with noisy backgrounds, while pixel-wise mask based features may capture more detailed and precise cues. We empirically study the effectiveness of bounding box and pixel-wise mask based relation representations as well as their hybrids. Our results suggest that the pixel-mask representation is indeed more powerful.

Our model reached the 1^{st} place in **ICCV-2019 Person in Context Challenge**[1] (PIC_{19} Challenge), on both *Human-Object Interaction in the Wild (HOIW)* and *Person in Context (PIC)* tracks, where HOIW addresses relation detection, while PIC focuses on relation segmentation. Besides, it also obtains promising results on V-COCO [20].

This paper makes three major contributions. **First**, we formulate HOI recognition as a coarse-to-fine inference procedure with a novel cascade architecture. **Second**, we introduce several techniques to learn rich features that represent the semantics of HOIs. **Third**, for the first time, we study the feature representations of HOI and find pixel-mask to be more powerful than the traditional bounding-box representation. We expect such a study could inspire more future efforts towards pixel-level HOI understanding.

2. Related Work

Human-object interaction recognition has a rich study history in computer vision. Early methods [51, 50, 6] mainly exploited human-object contextual information in structured models, such as Bayesian inference [18, 19], and compositional framework [8].

With the recent renaissance of neural networks in computer vision, deep learning based solutions are now dominant in this field. For instance, in [16], a multi-branch architecture was explored to address human, object, and relation representation learning. Some researchers revisited the classic graph model and solved this task in a neural message passing framework [38]. For learning more effective human feature representations, pose cues have been widely adopted in recent leading approaches [31, 21, 43, 10, 55]. Some other efforts addressed long-tail distribution and zero-shot problems with external knowledge [17, 25, 57, 42]. All these models use single-stage pipelines for inference (Fig. 2 (a)), and they can potentially benefit from the general architecture we propose here: a multi-stage pipeline that performs coarse-to-fine inference as shown in Fig. 2(b).

[1] http://picdataset.com/challenge/leaderboard/pic2019

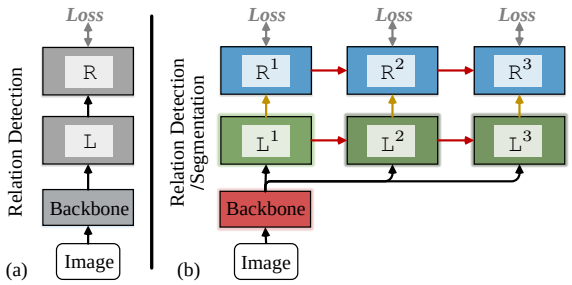

Figure 2: (a) Previous HOI recognition models largely rely on a single-stage architecture and only concern relation detection. (b) Our proposed HOI recognition model carries out instance localization and interaction recognition in a unified cascade architecture, and addresses both relation detection and segmentation. Note that the loss for the instance localization part is omitted for clarity.

Object detection has gained remarkable progress recently, benefiting from the availability of large-scale datasets (*e.g.*, MS-COCO [33]) and strong representation power of deep neural networks. Mainstream methods are often categorized into two-stage [40, 22, 5, 3] or single-stage [39, 34, 30, 28] paradigms. Recently, some multi-stage pipelines have been explored for coarse-to-fine object detection [2, 5]. Similarly, we revisit the general idea of cascade inference in HOI recognition, where both instance localization and relation recognition are coupled for step-by-step HOI reasoning.

3. Our Algorithm

3.1. Cascade Network Architecture

To identify $\langle$*human, verb, object*$\rangle$ triplets in images, our method carries out progressive refinement on instance localization and relation recognition at multiple stages (see Fig. 2 (b)). At each stage t, the multi-tasking is achieved by two networks: an instance localization network $\mathtt{L}^t$ generates *human* and *object* proposals, and an instance recognition network $\mathtt{R}^t$ identifies the action (*i.e.*, *verb*) for each human-object pair sampled from the proposals, as shown in Fig. 3 (a). Our cascade network is organized as follows:

$$\text{Instance Localization (§3.2):}\ \ \mathcal{O}^t = \mathtt{L}^t(\mathcal{O}^{t-1}),$$
$$\text{Human-Object Pair Sampling:}\ \ (h, o) \sim \mathcal{O}^t \times \mathcal{O}^t,$$
$$\text{Interaction Recognition (§3.3):}\ \ \boldsymbol{s}^t = \mathtt{R}^t(\boldsymbol{X}^t, \boldsymbol{X}^{t-1}).$$

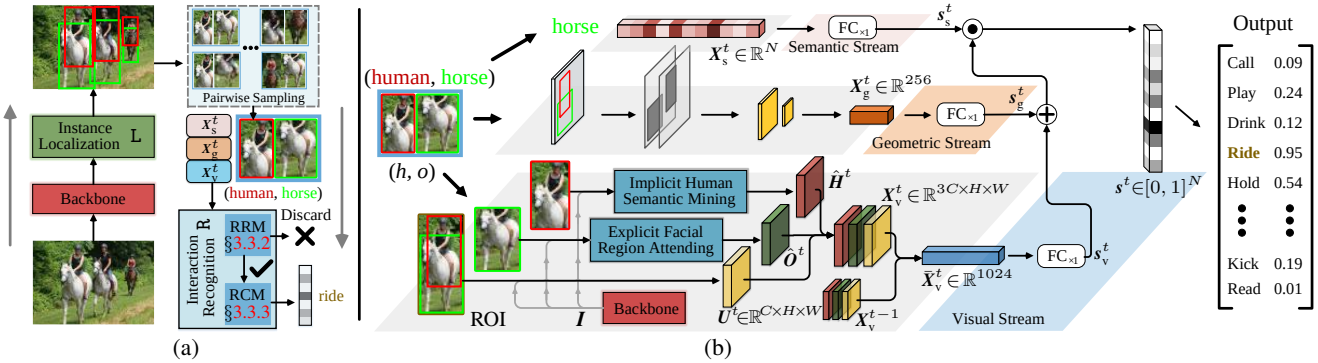

Figure 3: (a) Pipeline of our cascade network for identifying a triplet of ⟨*human*, *verb*, *object*⟩ from an input image. (b) Illustration of our triple-stream relation classification module (RCM) that achieves HOI recognition based on our human-centric relation representation.

At stage t, $\mathtt{L}^t$ takes the detection results $\mathcal{O}^{t-1}$ from $\mathtt{L}^{t-1}$ as inputs and outputs refined results $\mathcal{O}^t$. Then, a human-object pair (h, o) is sampled from $\mathcal{O}^t \times \mathcal{O}^t$. Finally, $\mathtt{R}^t$ uses the relation features $\boldsymbol{X}^t$ and $\boldsymbol{X}^{t-1}$ of (h, o) at current and previous stages to estimate a *verb* score vector $\boldsymbol{s}^t$. More details about the relation feature are given in §3.3.1. Notably, the instance localization $\mathtt{L}^t$ and interaction recognition $\mathtt{R}^t$ networks work closely at each stage, and $\mathtt{R}^t$ can benefit from the improved localization results $\mathcal{O}^t$ of $\mathtt{L}^t$ and give better interaction predictions.

Next we will describe in detail our instance localization network in §3.2 and interaction recognition network in §3.3.

3.2. Instance Localization Network

The instance localization network $\mathtt{L}$ outputs a set of human and object regions, from which human-object pair candidates are sampled and fed into the interaction recognition network $\mathtt{R}$ for relation classification. It is built on a cascade of detectors, *i.e.*, at stage t, $\mathtt{L}^t$ refines an object region $o^{t-1} \in \mathcal{O}^{t-1}$ detected from the preceding stage by:

$$\boldsymbol{Y}^t = \mathtt{P}(\boldsymbol{I}, o^{t-1}), \tag{1}$$

$$o^t = \mathtt{D}^t(\boldsymbol{Y}^t), \tag{2}$$

where $\boldsymbol{I}$ is the CNN feature of the backbone network, shared by different stages. $\boldsymbol{Y}^t \in \mathbb{R}^{C\times H\times W}$ indicates the box feature derived from $\boldsymbol{I}$ and the input RoI. $\mathtt{P}$ and $\mathtt{D}^t$ represent RoIAlign [22] and a box regression head, respectively.

Similar to previous cascade object detectors [2, 5], at each stage, $\mathtt{L}^t$ is trained with a certain interaction over union (IoU) threshold, and its output is re-sampled to train the next detector $\mathtt{L}^{t+1}$ with a higher IoU threshold. In this way, we gradually increase the quality of training data for deeper stages in the cascade, thus boosting the selectiveness against hard negative examples. At each stage, the instance localization loss $\mathcal{L}^t_{\text{LOC}}$ is the same as Faster R-CNN [40].

3.3. Interaction Recognition Network

As shown in Fig. 3 (a), the interaction recognition network $\mathtt{R}$ comprises a relation ranking module (RRM, §3.3.2) and a relation classification module (RCM, §3.3.3). Both RRM and RCM rely on our elaborately designed human-centric relationship representation (§3.3.1).

3.3.1 Human-Centric Relation Representation

At each stage t, for each human-object pair $(h^t, o^t) \in \mathcal{O}^t \times \mathcal{O}^t$, three types of features, *i.e.*, *semantic* feature $\boldsymbol{X}^t_{\text{s}}$, *geometric* feature $\boldsymbol{X}^t_{\text{g}}$ and *visual* feature $\boldsymbol{X}^t_{\text{v}}$, are considered for a thorough relation representation, as shown in Fig. 3(b). In the following paragraphs, the superscript 't' is omitted for conciseness unless necessary.

Semantic feature $\boldsymbol{X}_{\text{s}}$. It captures our prior knowledge of *object affordances* [14] (*e.g.*, a phone affords calling). We build $\boldsymbol{X}_{\text{s}} \in \mathbb{R}^N$ as the frequency of label co-occurrence between object and action categories [52], where N denotes the number of pre-defined actions in a HOI dataset.

Geometric feature $\boldsymbol{X}_{\text{g}}$. It characterizes the spatial relationship between human and object. Similar to [4, 13], we first adopt a two-channel mask representation strategy, obtaining a $(2, 64, 64)$-d feature tensor for the two entities. Then two conv+pooling operations followed by a fully connected (FC) layer are applied on the tensor to get $\boldsymbol{X}_{\text{g}} \in \mathbb{R}^{256}$.

Visual feature $\boldsymbol{X}_{\text{v}}$. Compared with $\boldsymbol{X}_{\text{s}}$ and $\boldsymbol{X}_{\text{g}}$, the visual feature is of greater significance and has profound effects for human beings to recognize subtle interactions. For each human-object pair (h, o), we have three features $\boldsymbol{H} \in \mathbb{R}^{C\times H\times W}$, $\boldsymbol{O} \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{U} \in \mathbb{R}^{C\times H\times W}$ from the human, object and their union regions correspondingly:

$$\boldsymbol{H} = \mathtt{P}(\boldsymbol{I}, h), \quad \boldsymbol{O} = \mathtt{P}(\boldsymbol{I}, o), \quad \boldsymbol{U} = \mathtt{P}(\boldsymbol{I}, (h, o)). \tag{3}$$

Here $\boldsymbol{H}$, $\boldsymbol{O}$ and $\boldsymbol{U}$ are specific instances of the RoIAlign feature $\boldsymbol{Y}$ in Eqs. (1,2), which are renamed to make it clear that they come from different regions.

To better capture the underlying semantics in HOI, we introduce two feature-enhancement mechanisms: *implicit human semantic mining* to improve the human feature $\boldsymbol{H}$ and *explicit facial region attending* to enhance the object feature $\boldsymbol{O}$. Then we have the visual feature as:

$$\boldsymbol{X}_{\text{v}} = [\bar{\boldsymbol{H}}, \bar{\boldsymbol{O}}, \boldsymbol{U}] \in \mathbb{R}^{3C\times H\times W}, \tag{4}$$

where $\bar{\boldsymbol{H}}$ and $\bar{\boldsymbol{O}}$ denote the enhanced human and object features, respectively, and $[\cdot]$ is the concatenation operation. Next we detail our two feature-enhancement mechanisms.

1) *Implicit Human Semantic Mining.* To reason about human-object interactions, it is essential to understand **how** humans interact with the world, *i.e.*, which human parts are involved for an action. Different from current leading methods resorting to expensive human pose annotations [43, 10, 31], we propose to implicitly learn human parts and their mutual interactions.

For each pixel (position) i inside the human region (feature) $\boldsymbol{H}$, we define its *semantic context* as the pixels that belong to the same semantic human part category of i. We use such semantic context to enhance our human representation, as it captures the relations within and among parts. Such enhancement would require a human part label map. Here, we compute a semantic similarity map as a surrogate to expedite computation. Specifically, for each pixel i we compute a semantic similarity map $A^i \in [0,1]^{H\times W}$, where each element $a^i_j \in A^i$ stores the 'relation' between the *latent* part categories of pixel i and j:

$$a^i_j = \frac{1}{z_i} exp(\boldsymbol{h}_i^\top \boldsymbol{h}_j), \tag{5}$$

where $\boldsymbol{h}_i \in \mathbb{R}^C$ and $\boldsymbol{h}_j \in \mathbb{R}^C$ are the feature vectors of pixels i and j in $\boldsymbol{H}$, respectively. z_i is a normalization term: $z_i = \sum_j exp(\boldsymbol{h}_i^\top \boldsymbol{h}_j)$. Here A^i can be considered as a soft label map for the semantic human part of i.

Then for a pixel i, we collect information from its semantic context according to A^i:

$$\boldsymbol{c}_i = \sum\nolimits_{j=1}^{H\times W} a^i_j \boldsymbol{h}_j \in \mathbb{R}^C. \tag{6}$$

After assembling all the semantic context information for all the parts (pixels) within $\boldsymbol{H}$, we get a semantic context enhanced feature $\boldsymbol{C} \in \mathbb{R}^{C\times H\times W}$, which is used to compute an improved human representation $\bar{\boldsymbol{H}}$:

$$\bar{\boldsymbol{H}} = \boldsymbol{H} + \boldsymbol{C} \in \mathbb{R}^{C\times H\times W}. \tag{7}$$

2) *Explicit Facial Region Attending.* Human face is vital for HOI understanding, as it conveys rich information closely tied to underlying attention and intention [27] of humans. There are many interactions that directly involve human face. For example, humans use *eyes* to *watch TV*, use *mouth* to *eat food*, and so on. Besides, face-related interactions are typically fine-grained and combined with heavy occlusions on the interacted objects, *e.g.*, *call a phone*, *play a phone*, posing great difficulties for HOI models. To address the above issues, we propose another feature-enhancement mechanism, called explicit facial region attending. This mechanism enriches the object representation $\boldsymbol{O}$ via two attention mechanisms:

- *Face-aware Attention.* For a human-object pair (h, o), we detect the facial region using an off-the-shelf face detector [7]. Then we get an RoIAlign feature $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ from the detected facial region as the face representation. An attention score $\alpha \in [0,1]$ is learned for interpreting the importance of the facial region for the object o:

$$\alpha = \sigma(\text{FC}_{\times 2}([\boldsymbol{F}, \boldsymbol{O}])), \tag{8}$$

  where σ is the *sigmoid* function, and $\text{FC}_{\times 2}$ stands for two stacked FC layers.
- *Face-agnostic Attention.* The face-aware enhancement addresses the relevance between human face and object. To mine the potential relations between object and other human regions, we propose a face-agnostic attention. We first remove the facial region from the human h, by setting the pixel values in the face region to zero. Then we get the corresponding RoIAlign feature $\bar{\boldsymbol{F}} \in \mathbb{R}^{C\times H\times W}$ from the face-removed human regions. Finally, we calculate an importance score $\bar{\alpha} \in [0,1]$ between $\bar{\boldsymbol{F}}$ and $\boldsymbol{O}$:

$$\bar{\alpha} = \sigma(\text{FC}_{\times 2}([\bar{\boldsymbol{F}}, \boldsymbol{O}])). \tag{9}$$

Considering Eqs.(8,9), the object feature $\boldsymbol{O}$ is enhanced by:

$$\bar{\boldsymbol{O}} = \boldsymbol{O} + \alpha \boldsymbol{F} + \bar{\alpha} \bar{\boldsymbol{F}} \in \mathbb{R}^{C\times H\times W}. \tag{10}$$

In our cascade framework, for a human-object pair $(h, o) \in \mathcal{O}^t \times \mathcal{O}^t$ at stage t, we update its visual feature $\boldsymbol{X}^t_{\text{v}} \in \mathbb{R}^{3C\times H\times W}$ by considering the one $\boldsymbol{X}^{t-1}_{\text{v}} \in \mathbb{R}^{3C\times H\times W}$ in prior stage:

$$\bar{\boldsymbol{X}}^t_{\text{v}} = \text{FC}_{\times 2}(\boldsymbol{X}^t_{\text{v}} + \boldsymbol{X}^{t-1}_{\text{v}}) \in \mathbb{R}^{1024}. \tag{11}$$

We do not update semantic $\boldsymbol{X}_{\text{s}}$ and geometric $\boldsymbol{X}_{\text{g}}$ features.

3.3.2 Relation Ranking Module

Once obtaining the features $\{\boldsymbol{X}_{\text{s}}, \boldsymbol{X}_{\text{g}}, \bar{\boldsymbol{X}}_{\text{v}}\}$ of a human-object pair, we can directly predict its action label. However, a big issue here is how to sample human-object pairs. Given the proposals detected from the localization network, previous HOI methods typically pair all humans and objects, leading to large computational overhead. As a matter of fact, human beings interact with the world following some regularity rather than in a pure chaotic way [1]. By leveraging such regularity, we propose a human-object relation ranking module (RRM) to select high-quality HOI candidates for further relation recognition. This also helps decrease the difficulty in relation classification and erase the serious class imbalance, as the samples for 'non-interaction' class are much more than the ones of any other interaction classes.

RRM is built upon an insight that, although some human-object relations are miss annotated in HOI datasets, the annotated human-object pairs tend to be more relevant (*i.e.*, higher ranking score) than those without any HOI relation labelling. Given the detection results $\mathcal{O}$ of the instance localization network $\mathtt{L}$ (§3.2), we denote the set of all the possible human-object pairs as: $\mathcal{P} = \{P = (h, o) \in \mathcal{O} \times \mathcal{O}\}$. $\mathcal{P}$ can be further divided into two subsets: $\mathcal{P} = \hat{\mathcal{P}} \cup \check{\mathcal{P}}$, where $\hat{\mathcal{P}}$ and $\check{\mathcal{P}}$ indicate the sets of annotated and un-annotated

human-object pairs, respectively. The goal of RRM is to learn a ranking function $\mathtt{g} : \mathbb{R}^{1024+256} \rightarrow \mathbb{R}$ that fulfills the following constraint:

$$\forall \hat{P} \succ \check{P} : \mathtt{g}(\hat{P}) > \mathtt{g}(\check{P}), \quad \text{where } \hat{P} \in \hat{\mathcal{P}}, \check{P} \in \check{\mathcal{P}}. \tag{12}$$

Here $\hat{P} \succ \check{P}$ means $\hat{P}$ has a higher ranking than $\check{P}$. $\mathtt{g}(P)$ gives the ranking score of P:

$$\mathtt{g}(P) = \sigma(\mathrm{FC}_{\times 1}(\bar{\boldsymbol{X}}_\mathrm{v}, \boldsymbol{X}_\mathrm{g})) \in [0, 1]. \tag{13}$$

In RRM, the learning of $\mathtt{g}$ is achieved by minimizing the following pairwise ranking hinge loss:

$$\mathcal{L}_\mathrm{RRM} = \sum\nolimits_{\hat{P} \in \hat{\mathcal{P}}} \sum\nolimits_{\check{P} \in \check{\mathcal{P}}} \max(0, \mathtt{g}(\check{P}) - \mathtt{g}(\hat{P}) + \epsilon)), \tag{14}$$

where the margin ϵ is empirically set as 0.2. This loss penalizes the situation that assigning an un-annotated pair $\check{P}$ with a higher ranking score, compared to a labeled pair $\hat{P}$.

3.3.3 Relation Classification Module

Through RRM, only a few top-ranked, high-quality human-object pairs are preserved and fed into a triple-stream [53], relation classification module (RCM) for final HOI recognition. For a HOI candidate (h, o), the semantic $\boldsymbol{X}_\mathrm{s}$, geometric $\boldsymbol{X}_\mathrm{g}$ and visual $\bar{\boldsymbol{X}}_\mathrm{v}$ features, are separately fed into a corresponding stream in RCM for estimating a HOI action score vector independently:

$$\begin{aligned} \text{semantic stream: } & \boldsymbol{s}_\mathrm{s} = \sigma(\mathrm{FC}_{\times 1}(\boldsymbol{X}_\mathrm{s})) \in [0, 1]^N, \\ \text{geometric stream: } & \boldsymbol{s}_\mathrm{g} = \sigma(\mathrm{FC}_{\times 1}(\boldsymbol{X}_\mathrm{g})) \in [0, 1]^N, \\ \text{visual stream: } & \boldsymbol{s}_\mathrm{v} = \sigma(\mathrm{FC}_{\times 1}(\bar{\boldsymbol{X}}_\mathrm{v})) \in [0, 1]^N, \end{aligned} \tag{15}$$

where $\boldsymbol{s}_\mathrm{s}$, $\boldsymbol{s}_\mathrm{g}$ and $\boldsymbol{s}_\mathrm{v}$ are the score vectors from semantic, geometric and visual streams, respectively, and N is the number of pre-defined actions in HOI. Note that here follows a multi-label classification setting.

During training, for each stream, the binary cross-entropy loss is used to evaluate the discrepancy between the output score and truth target. The total loss $\mathcal{L}_\mathrm{RCM}$ is the sum of the ones from streams. During inference, the final prediction is obtained by:

$$\boldsymbol{s} = (\boldsymbol{s}_\mathrm{v} + \boldsymbol{s}_\mathrm{g}) \odot \boldsymbol{s}_\mathrm{s}, \tag{16}$$

where $\odot$ denotes the Hadamard product.

3.4. Relation Segmentation

So far, we strictly follow the classic *relation detection* setting in HOI recognition [16, 31, 10, 49], *i.e.*, identify the interaction entities by bounding boxes. Now we focus on how to adapt our cascade framework to *relation segmentation*, which addresses more fine-grained HOI understanding by representing each entity at the pixel level.

Inspired by [2], for the instance localization network $\mathtt{L}^t$ at each stage t, an instance segmentation head $\mathtt{S}^t$ is added and the whole workflow (Eqs. (1, 2)) is changed as:

$$\begin{aligned} \text{Instance Detection: } & \boldsymbol{Y}^t = \mathtt{P}(\boldsymbol{I}, o^{t-1}), o^t = \mathtt{D}^t(\boldsymbol{Y}^t), \\ \text{Instance Segmentation: } & \bar{\boldsymbol{Y}}^t = \mathtt{P}(\boldsymbol{I}, o^t), \quad \bar{o}^t = \mathtt{S}^t(\bar{\boldsymbol{Y}}^t, \bar{\boldsymbol{Y}}^{t-1}), \end{aligned} \tag{17}$$

where $\bar{o}^t \in \bar{\mathcal{O}}^t$ indicates a generated object instance mask. Then, in our relation recognition network (§3.3), the human-object pair (h, o) is sampled from the object masks $\bar{\mathcal{O}}^t$ and associated with finer features: $\boldsymbol{H}$, $\boldsymbol{O}$ and $\boldsymbol{U}$ by pixel-wise RoI. In addition, the generation of geometric feature $\boldsymbol{X}_\mathrm{g}$ is based on pixel-level masks. The binary cross-entropy loss $\mathcal{L}^t_\mathrm{SEG}$ is used for training $\mathtt{S}$.

3.5. Implementation Details

Training Loss. Since all the modules mentioned above are differentiable, our cascade architecture can be trained in an end-to-end manner. In the *relation detection* setting, the entire loss is computed as:

$$\mathcal{L} = \sum\nolimits_{t=1}^{T} \beta^t \mathcal{L}^t_\mathrm{LOC} + \gamma^t (\mathcal{L}^t_\mathrm{RRM} + \mathcal{L}^t_\mathrm{RCM}). \tag{18}$$

Here, $\mathcal{L}^t_\mathrm{LOC}$ is the localization loss at stage t (§3.3). $\mathcal{L}^t_\mathrm{RRM}$ and $\mathcal{L}^t_\mathrm{RCM}$ are the losses of RRM (§3.3.2) and RCM (§3.3.3), respectively. The coefficients β_t and γ_t are used to balance the contributions of different stages and tasks. There are three stages used in our method ($T = 3$), and we set $\beta = \gamma = [1, 0.5, 0.25]$. In the *relation segmentation* setting, the instance segmentation head $\mathtt{S}^t$ is injected into the network (§3.4). The corresponding instance segmentation loss $\mathcal{L}^t_\mathrm{SEG}$ is further added in Eq. (18), with coefficients $[1, 0.5, 0.25]$.

Cascade Inference. During inference, the object proposals generated by the instance localization network in different stages are merged together. We remove the ones whose confidence scores are smaller than 0.3. Then, all the possible human-object pairs, generated from the remaining proposals, are fed into RRM for relation ranking. After that, we only select the top 64 pairs as candidates and feed them into RCM for final relation classification. The last-stage output of RCM is used as the final action score.

4. Experiments

Experiments are conducted on three datasets, *i.e.*, HOIW, PIC and V-COCO [20]. The former two are from the PIC_{19} Challenge, and the last one is a gold standard benchmark.

Training Settings: Unless specially noted, we adopt the following training settings for all the experiments. We use ResNet-50 [23] as the backbone. The training includes two phases: 1) training the instance localization network; and then 2) jointly training the instance localization and interaction recognition networks. In the first phase, the network is initialized using the weights pre-trained on COCO [33]. The three stages are trained using gradually increased IoU thresholds $\mu = \{0.5, 0.6, 0.7\}$ [2, 5]. Training images are resized to a maximum scale of 1333×800, without changing the aspect ratio. We apply horizontal flipping for data augmentation and train the network for 12 epochs with batch

Challenge	Team	mAP_{rel}
PIC_{19} Challenge (*HOIW Track*)	Ours	**66.04**
	GMVM	60.26
	FINet	56.93
	F2INet	49.13
	TIN [31]	48.64

Table 1: **Relation detection results on HOIW `test` set in PIC_{19} Challenge** (§4.1).

Challenge	Team	R@100 mIoU: 0.25	R@100 mIoU: 0.50	R@100 mIoU: 0.75	Mean
PIC_{19} Challenge (*PIC Track*)	Ours	**60.17**	**55.11**	**42.29**	**52.52**
	HTC+iCAN	56.21	52.32	37.49	48.67
	RelNet	53.17	49.26	32.44	44.96
	XNet	38.42	33.15	17.29	29.62

Table 2: **Relation segmentation results on PIC `test` set in PIC_{19} Challenge.** Please see §4.1 for details.

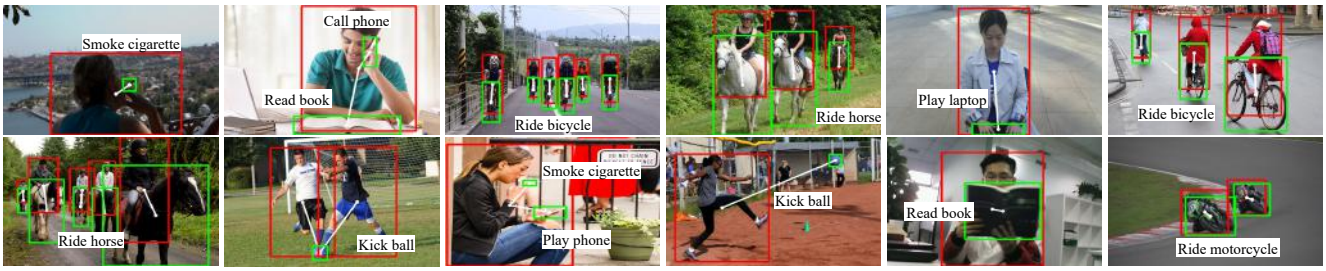

Figure 4: **Visual results for relation detection, on HOIW `test` set in PIC_{19} Challenge** (§4.1).

size 16 and initial learning rate 0.02, which is reduced by 10 at epoch 8 and 11. In the second phase, we adopt the image-centric training strategy [15], *i.e.*, using pairwise samples from one image to make up a mini-batch. For each mini-batch, we sample at most 128 HOI proposals with a ratio of 1:3 of positive to negative samples to jointly train RRM and RCM. At each stage, the same IoU threshold μ is used to determine positive HOI proposals so that the training data for the interaction recognition network closely match the detection quality. Besides, ground-truth HOIs are also used at each stage for training. The second phase is trained with learning rate 0.02 and batch-size 8 for 7 epochs.

Reproducibility: Our model is implemented on PyTorch and trained on 8 NVIDIA Tesla V100 GPUs with a 32GB memory per-card. Testing is conducted on a single NVIDIA TITAN Xp GPU with 12 GB memory.

4.1. Results on PIC_{19} Challenge

Dataset: The PIC_{19} Challenge includes two tracks, *i.e.*, HOIW and PIC tracks, each with a standalone dataset:

- **HOIW** [32] is for human-object relation detection. It has 29,842 training and 8,794 testing images, with bounding box annotations for 11 object and 10 action categories. Since it does not provide `train`/`val` splits, in our ablation study, we randomly choose 9,999 images for `val` and the other 19,843 for `train`; for the challenge result, we use `train+val` for training.
- **PIC** is for human-object relation segmentation. It has 17,606 images (12,654 for `train`, 1,977 for `val` and 2,975 for `test`) with pixel-level annotations for 143 objects. It covers 30 relationships, including 6 geometric (*e.g.*, *next-to*) and 24 non-geometric (*e.g.*, *look*, *talk*).

Evaluation Metrics: Standard evaluation metrics in the challenges are adopted. For HOIW, the performance is evaluated by mAP_{rel}. A detected triplet $\langle$*human, verb, object*$\rangle$ is considered as a true positive if the predicted *verb* is correct and both the *human* and *object* boxes have IoUs at least 0.5 with the corresponding ground-truths. For PIC, we use Recall@100 (R@100), which is averaged over two relationship categories (*i.e.*, *geometric* and *non-geometric*) and three IoU thresholds (*i.e.*, 0.25, 0.5 and 0.75). In our ablation study, we also consider R@50 and R@20 to measure the performance under stricter conditions.

Performance on the HOIW Track: Our approach reaches the 1^{st} place for relation detection on the HOIW track. As reported in Table 1, our result is substantially better than other teams. In particular, it is **5.78%** absolutely better than the 2^{nd} (GMVM) and **9.11%** better than the 3^{rd} (FINet). Our approach also significantly outperforms one published state-of-the-art, *i.e.*, TIN [31] . Fig. 4 presents some visual results on HOIW `test`. Our model shows robust to various challenges, *e.g.*, occlusions, subtle relationships, *etc*.

Performance on the PIC Track: Our approach also reaches the 1^{st} place for relation segmentation on the PIC track. As reported in Table 2, our overall score (**52.52%**) outperforms the 2^{nd} place by **3.85%** and the 3^{rd} by **7.56%**. Fig. 5 depicts visual results of two complex scenes on PIC `test`. Our method shows outstanding performance in terms of instance segmentation as well as interaction recognition. It can identify both geometric and non-geometric relationships, and is capable of recognizing many fine-grained interactions, *e.g.*, *look human*, *hold tableware*. In this track, the instance localization network is instantiate as Eq.(17).

4.2. Results on V-COCO

Dataset: V-COCO [20] provides verb annotations for MS-COCO [33]. Proposed in 2015, it is the first large-scale dataset for HOI understanding and remains the most popular one today. It contains 10,346 images in total ($2,533/2,867/4,946$ for `train`/`val`/`test` splits). 16,199 human instances are annotated with 26 action labels,

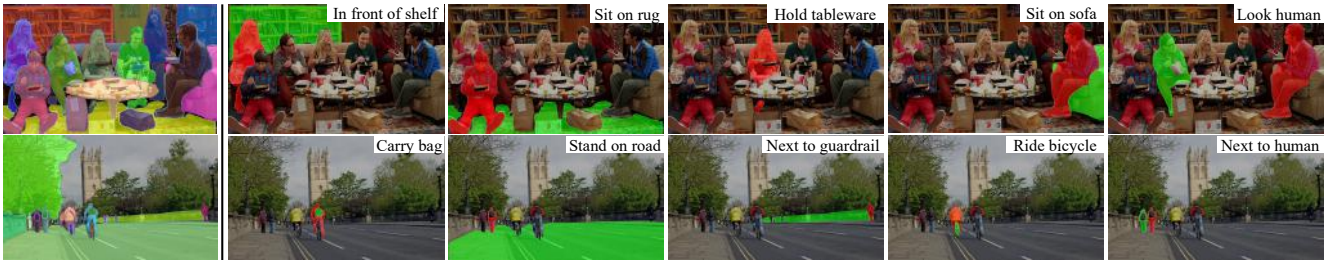

Figure 5: **Visual results for relation segmentation, on PIC `test` set in PIC$_{19}$ Challenge** (§4.1). First column: Instance segmentation results. Last five columns: Top ranked ⟨*human, verb, object*⟩ triplets. For each triplet, the *human* and *object* are shown in red and green.

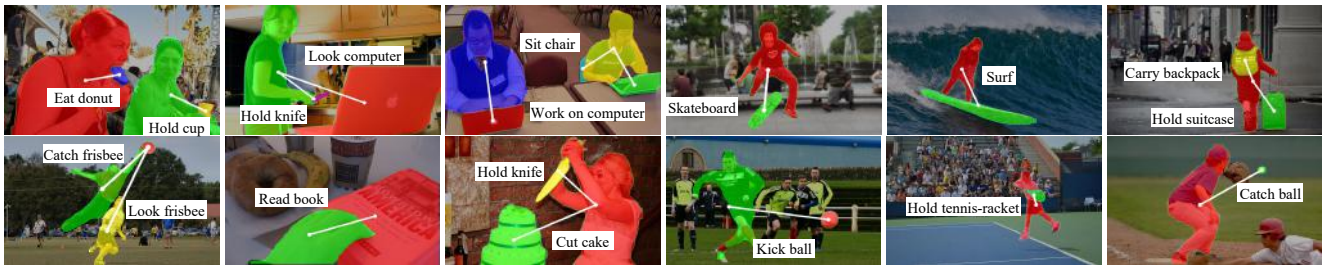

Figure 6: **Visual results for relation segmentation, on V-COCO `test` set [20].** See §4.2 for details.

Methods	Publication	Backbone	mAP$_{role}$(%)
Gupta *et. al.* [20]	Arxiv15	ResNet-50-FPN	31.8
Interact [16]	CVPR18	ResNet-50-FPN	40.0
GPNN [38]	ECCV18	ResNet-50	44.0
iCAN [13]	BMVC18	ResNet-50	45.3
Xu *et. al.* [49]	CVPR19	ResNet-50-FPN	45.9
Wang *et. al.* [44]	ICCV19	ResNet-50	47.3
RPNN [55]	ICCV19	ResNet-50	47.5
TIN [31]	CVPR19	ResNet-50	47.8
$Ours_{\text{bbox}}$	-	ResNet-50	48.3
$Ours_{\text{mask}}$	-	ResNet-50	**48.9**

Table 3: **Comparison of mAP$_{role}$ on V-COCO `test` [20]** (§4.2).

wherein three actions (*i.e.*, *cut, hit, eat*) are annotated with two types of targets (*i.e.*, instrument and direct object).

Evaluation Metrics: We use the original role mean AP (mAP$_{role}$), which is exactly same with mAP$_{rel}$ in HOIW.

Performance: Since V-COCO has both bounding box and mask annotations, we provide two variants of our methods, *i.e.*, $Ours_{\text{bbox}}$ and $Ours_{\text{mask}}$, where $Ours_{\text{bbox}}$ is trained with box annotations while $Ours_{\text{mask}}$ uses groundtruth masks. For fairness, during evaluation, the mask outputs of $Ours_{\text{mask}}$ are transformed to boxes. Table 3 summarizes the results in comparison with 8 state-of-the-arts. $Ours_{\text{bbox}}$ outperforms TIN [31] by **0.5%** and RPNN [55] by **0.8%**. $Ours_{\text{mask}}$ further improves $Ours_{\text{bbox}}$ by **0.6%**, which suggests the superiority of the mask-level representation over the box-level. We would like to note that [43] reported a 52.0% mAP$_{role}$ on V-COCO. However, It relies on an expensive pose estimator, thus it is unfair to directly compare with our method. Without the pose estimator, [43] obtains a score of 48.6%, slightly worse than $Ours_{\text{mask}}$. In Fig. 6, we illustrate HOI segmentation results of Our_{mask} on V-COCO `test` set. It precisely recognizes many fine-grained interactions, such as *look computer*, *read book*, etc.

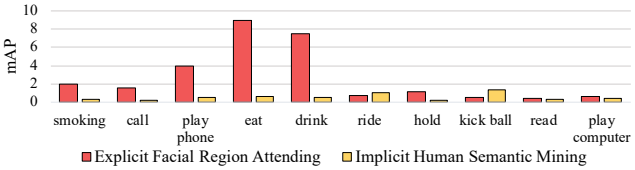

Figure 7: **Per-category performance improvement ΔmAP$_{rel}$ of the proposed attention modules on HOIW `val` set** (§4.3).

IHSM	EFRA	RRM	CAS	PIC R@20	PIC R@50	PIC R@100	HOIW mAP$_{rel}$
✗	✗	✗	✗	17.0	28.0	33.9	33.9
✓				17.9	28.6	34.3	34.4
	✓			17.6	27.5	34.6	36.7
✓	✓			18.5	28.3	35.4	37.5
✓	✓	✓		19.0	28.9	35.9	38.6
✓	✓	✓	✓	**27.8**	**38.3**	**45.3**	**43.7**

Table 4: **Ablation study of key components** in our cascade model.

Overall, our model consistently achieves promising results over different datasets as well as two different settings (*i.e.*, relation detection and segmentation), which clearly reveals its remarkable performance and strong generalization.

4.3. Ablation Study

Key Component Analysis. First, we investigate the influence of essential components in our framework, *i.e.*, implicit human semantic mining (IHSM), explicit facial region attending (EFRA), relation ranking module (RRM) and cascade network architecture (CAS). We first build a baseline model without any of these components, and then gradually add each into the baseline for investigation. As reported in Table 4, all these components can improve the performance in both PIC and HOIW datasets. **1)** IHSM and EFRA help to learn more discriminative visual features and further boost the performance (*e.g.*, **0.5%** and **2.8%** performance

T	Speed (ms)	PIC R@20	PIC R@50	PIC R@100	HOIW mAP_{rel}
1	145	19.0	28.9	35.9	38.6
2	163	25.5	36.4	43.8	42.1
3	198	**27.8**	**38.3**	**45.3**	**43.7**
4	253	**27.8**	**38.3**	45.2	**43.7**
5	314	27.6	38.1	45.2	43.4

Table 5: **Impact of the number of stages** T in our cascade model.

Backbone	Cascade	PIC R@20	PIC R@50	PIC R@100	HOIW mAP_{rel}
ResNet-50	✗	19.0	28.9	35.9	38.6
ResNet-50	✓	27.8	38.3	45.3	43.7
ResNet-101	✗	20.8	31.4	38.9	40.2
ResNet-101	✓	28.6	39.8	47.0	44.4
ResNeXt-101	✗	22.9	34.3	42.6	44.2
ResNeXt-101	✓	29.6	41.2	48.9	48.2

Table 6: **Ablation study of the cascade architecture with various backbones.**

improvements on HOIW). **2)** Fig. 7 shows the per-category performance improvement of IHSM and EFRA on HOIW `val` set. Obviously, EFRA improves the performance on face-related interactions (*e.g.*, *eat, drink, smoking, call*) and discriminates these categories from some similar ones, *e.g.*, *play phone*. In contrast, IHSM is more effective for the actions with specific poses, *e.g.*, *ride, kick ball*. **3)** RRM plays a key role in pruning negative human-object pairs, as proved by Table 4. Moreover, RRM improves the average inference speed by about 80**ms** on HOIW. **4)** Our cascade architecture substantially boosts the performance, *i.e.*, **8.8%** absolute improvement in PIC and **5.1%** in HOIW.

Cascade Architecture Analysis. We study the impact of the number of stages T used in our cascade network by varying it from 1 to 5. The IoU thresholds used for these five stages are $[0.5, 0.6, 0.7, 0.75, 0.8]$. The results in Table 5 show that the performance is significantly improved by adding a second stage, *i.e.*, **6.5%** in terms of R@20 in PIC and **3.5%** in terms of mAP_{rel} in HOIW. When further adding more than 3 stages, the performance gain is marginal. Table 5 also reports the average inference time for these variants on HOIW `val` set. The test speed decreases with adding more stages and drops quickly after using 4 or 5 stages. Considering the model complexity and performance, we choose $T = 3$ as our default setting. Table 6 reports the performance comparison of our approach with ($T = 3$) or without ($T = 1$) cascade under different backbones, *i.e.*, ResNet-50, ResNet-101 and ResNeXt-101. The results reveal that our cascade network consistently improves the performance on various backbones.

Efficacy of Our Relation Representation and Score Fusion Strategy. In our method, three kinds of features, $\boldsymbol{X}_{\text{s}}$, $\boldsymbol{X}_{\text{g}}$ and $\boldsymbol{X}_{\text{v}}$, are used for capture semantic, geometric and visual information for relation modeling. Table 7 reports the performance with only considering one single feature. As seen, the visual feature is more important than the other two. In addition, we further investigate different ways to fuse the action scores from the three features, we find that the one used in Eq. (16) is the best.

Aspect	Variant	PIC R@20	PIC R@50	PIC R@100	HOIW mAP_{rel}
Relation Representation	Semantic Feature ($\boldsymbol{s}_{\text{s}}$)	14.5	20.0	23.3	26.5
	Geometric Feature ($\boldsymbol{s}_{\text{g}}$)	19.6	26.2	32.1	30.3
	Visual Feature ($\boldsymbol{s}_{\text{v}}$)	22.2	32.8	38.2	38.1
Score Fusion	$\boldsymbol{s}_{\text{v}} + \boldsymbol{s}_{\text{g}} + \boldsymbol{s}_{\text{s}}$	26.7	37.0	43.1	41.3
	$\boldsymbol{s}_{\text{v}} \odot \boldsymbol{s}_{\text{g}} \odot \boldsymbol{s}_{\text{s}}$	27.0	37.7	43.5	41.9
	$(\boldsymbol{s}_{\text{v}} + \boldsymbol{s}_{\text{g}}) \odot \boldsymbol{s}_{\text{s}}$	**27.8**	**38.3**	**45.3**	**43.7**

Table 7: **Ablation study of our relation representation and score fusion strategy.**

Relation Representation	R@20	R@50	R@100	Mean
BBox	27.1	37.9	44.8	36.6
Mask	**27.8**	**38.3**	**45.3**	**37.1**
BBox + Mask (`max`)	27.6	**38.3**	45.1	37.0
BBox + Mask (`sum`)	27.7	**38.3**	45.1	37.0

Table 8: **Comparison between mask and bbox representations.**

Exploring Better Relation Representation. Existing HOI methods typically use coarse bounding boxes to represent the entities, however, is it the best choice? To answer this, we perform experiments to explore more powerful relation representation. We evaluate the performance of our model on PIC `val` set using four different representations: a) BBox; b) Mask; c) BBox+Mask (max); and d) BBox+Mask (sum). Here, a) and b) means that we extract the features $\boldsymbol{H}$, $\boldsymbol{O}$, $\boldsymbol{U}$ by applying RoIAlign over bbox and mask regions, respectively. c) and d) are the fusion of bbox and mask features with element-wise `max` and `sum` operations, respectively. Note that the detected entities are the same for all the baselines. The results in Table 8 show that mask is superior to bbox, especially under the strictest metric R@20. The two hybrid representations are better than solely using bbox, but slightly worse than the purely mask-based. In summary, mask-based representation indeed benefits HOI recognition as it provides more precise information.

5. Conclusion

This paper introduces a cascade network architecture for coarse-to-fine HOI recognition. It consists of an instance localization network and an interaction recognition network, which are densely connected at each stage to fully exploit the superiority of multi-tasking. The interaction recognition network leverages human-centric features to learn better semantics of actions, and comprises two crucial modules for relation ranking and classification. Our model achieves the 1^{st} place on both relation detection and relation segmentation tasks in PIC_{19} Challenge, and also outperforms prior methods on a gold standard benchmark, V-COCO. Besides, we empirically demonstrate the advantages of mask over bounding box for more precise relation representation, and will go deep into this in our future research.

References

[1] Christopher Baldassano, Diane M Beck, and Li Fei-Fei. Human–object interactions are more than the sum of their parts. *Cerebral Cortex*, 27(3):2276–2288, 2017.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *arXiv preprint arXiv:1906.09756*, 2019.

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018.

[5] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019.

[6] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In *NeurIPS*, 2011.

[7] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[8] Chaitanya Desai and Deva Ramanan. Detecting actions, poses, and objects with relational phraselets. In *ECCV*, 2012.

[9] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *ICCV*, 2019.

[10] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018.

[11] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010.

[12] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient belief propagation for early vision. *IJCV*, 70(1):41–54, 2006.

[13] Chen Gao, Yuliang Zou, and Jia-Bin Huang. iCAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018.

[14] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 2014.

[15] Ross Girshick. Fast R-CNN. In *ICCV*, 2015.

[16] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *CVPR*, 2018.

[17] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *CVPR*, 2019.

[18] Abhinav Gupta and Larry S Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007.

[19] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE TPAMI*, 31(10):1775–1789, 2009.

[20] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015.

[21] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In *ICCV*, 2019.

[22] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[24] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *CVPR*, 2018.

[25] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, 2018.

[26] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[27] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.

[28] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018.

[29] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, 2019.

[30] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017.

[31] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019.

[32] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. PPDM: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020.

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[35] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, 2018.

[36] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. Learning conditioned graph structures for interpretable visual question answering. In *NeurIPS*, 2018.

[37] Bo Pang, Kaiwen Zha, Yifan Zhang, and Cewu Lu. Further understanding videos through adverbs: A new video task. In *AAAI*, 2020.

[38] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In *ECCV*, 2018.

[39] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[41] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao,

and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *ECCV*, 2018.

[42] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *WACV*, 2018.

[43] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, 2019.

[44] Tiancai Wang, Rao Muhammad Anwer, Muhammad Haris Khan, Fahad Shahbaz Khan, Yanwei Pang, Ling Shao, and Jorma Laaksonen. Deep contextual attention for human-object interaction detection. In *ICCV*, 2019.

[45] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*, 2019.

[46] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In *CVPR*, 2018.

[47] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *ICCV*, 2019.

[48] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning actor relation graphs for group activity recognition. In *CVPR*, 2019.

[49] Bingjie Xu, Yongkang Wong, Junnan Li, Qi Zhao, and Mohan S Kankanhalli. Learning to detect human-object interactions with knowledge. In *CVPR*, 2019.

[50] Bangpeng Yao and Li Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *CVPR*, 2010.

[51] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[52] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural Motifs: Scene graph parsing with global context. In *CVPR*, 2018.

[53] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *CVPR*, 2019.

[54] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In *CVPR*, 2019.

[55] Penghao Zhou and Mingmin Chi. Relation parsing neural network for human-object interaction detection. In *ICCV*, 2019.

[56] Tao Zhou, Huazhu Fu, Chen Gong, Jianbing Shen, Ling Shao, and Fatih Porikli. Multi-mutual consistency induced transfer subspace learning for human motion segmentation. In *CVPR*, 2020.

[57] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton van den Hengel. HCVRD: a benchmark for large-scale human-centered visual relationship detection. In *AAAI*, 2018.