UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection

Bumsoo Kim[1⋆], Taeho Choi[1⋆], Jaewoo Kang[1†], and Hyunwoo J. Kim[1†]

Korea University, Seoul 02841, Republic of Korea
{meliketoy,major1965,kangj,hyunwoojkim}@korea.ac.kr

Abstract. Recent advances in deep neural networks have achieved significant progress in detecting individual objects from an image. However, object detection is not sufficient to fully understand a visual scene. Towards a deeper visual understanding, the interactions between objects, especially humans and objects are essential. Most prior works have obtained this information with a bottom-up approach, where the objects are first detected and the interactions are predicted sequentially by pairing the objects. This is a major bottleneck in HOI detection inference time. To tackle this problem, we propose *UnionDet*, a one-stage meta-architecture for HOI detection powered by a novel union-level detector that eliminates this additional inference stage by directly capturing the region of interaction. Our one-stage detector for human-object interaction shows a significant reduction in interaction prediction time ($4\times \sim 14\times$) while outperforming state-of-the-art methods on two public datasets: V-COCO and HICO-DET.

Keywords: visual relationships, real-time detection, human-object interaction detection, object detection

1 Introduction

Recent advances in deep neural networks have achieved significant progress in detecting and recognizing individual objects from an image. However, to understand a scene, we need a deeper visual understanding that transcends the level of individual object detection. To understand what is happening in the image, not only do we have to accurately detect individual objects, but we also have to properly predict the interactions between the detected objects. Among the interactions, in this paper, we focus on *human-object interaction (HOI) detection* that involves the localization and classification of interactions between humans and surrounding objects. HOI detection has been formally defined in [10] as the task to detect $\langle human, verb, object \rangle$ triplets within an image.

The main challenge of HOI detection boils down to a simple question: "*How can we localize* ***interactions****?*". When asked to localize the area of *"A person rides a horse."*, a human can naturally spot the tight area that covers both the

⋆ equal contribution, †corresponding author

person and the horse he/she is riding. This is the *union region* of the interacting objects that have been considered as a representation of visual relationships from previous works [21], and have been widely utilized in HOI detection [8, 11, 15, 17, 29, 30, 35]. Ironically, no detector in the literature has been studied to *directly* capture the union region.

All the previous HOI detectors, therefore, incorporated a multi-stage and sequential pipeline that detects the individual objects first and 'associate' them to obtain the union region. This approach is far from intuitive and it makes HOI detectors inefficient. The sequential pipeline of object detection and interaction prediction makes end-to-end training impossible and creates a huge bottleneck in inference time for HOI detection. In standard object detection, one-stage detectors [19, 20, 26, 32, 33] were able to speed up two-stage detectors by eliminating the second stage while yielding a competitive performance. Yet in HOI detection, previous multi-stage models mainly focused on performance (e.g., average precision) leaving the large gap between high-performance and real-time detection unexplored. In this work, our goal is to fill the gap between the performance and inference time of HOI detection with a fast, single-stage model.

To this end, we propose ***UnionDet:*** a one-stage meta-architecture powered by a novel *union-level* detector that captures the union region of human-object interaction. Instead of associating the object detection results by feeding each object pair into a separate neural network afterward, we directly detect interacting $\langle human, object \rangle$ pairs with our novel union-level detection framework. This eliminates the need for heavy neural network inference after object detection and enables our model to detect interactions with minimal additional time on top of existing object detectors. Though the union-level detection sounds intuitive, detecting the union region is much more challenging than instance-level detection. In this paper, we study new challenges in union-level detection and address them by new techniques: (i) union anchor labeling, (ii) target object classification loss and (iii) union foreground focal loss. Based on these new methods, our proposed one-stage HOI detector achieves a $4\times \sim 14\times$ speed-up in additional inference time for interaction prediction while surpassing state-of-the-art performance on two HOI detection benchmark datasets: V-COCO (*Verbs in COCO*) and HICO-DET. The main **contributions** of our paper are threefold:

- We study new technical challenges in union-level detection, including bias toward human regions, inaccuracy of standard IoU-based matching and union regions containing multiple interactions and more than two objects.
- We propose a novel *union-level* detector that directly detects the interaction region. We study a new set of training techniques to address the new challenges of union-level detection.
- We propose a meta-architecture ***UnionDet*** equipped with our *union-level* detector. It is a single-stage HOI detector achieving $4\times \sim 14\times$ speed-up in interaction prediction and the *state-of-the-art* performance in two public datasets.

2 Related Work

2.1 One-Stage Object Detection

One-stage object detectors based on deep neural networks have formerly been proposed for faster detection [20, 26, 28]. These detectors have achieved a significant speed-up but they often come with a considerable loss of accuracy. One known problem is the class imbalance problem. Since one-stage object detectors densely sample anchor boxes, foreground anchor boxes are relatively much rarer than background anchor boxes (or negative samples), unlike two-stage methods that classify only a few anchor boxes after RPN. One common technique to resolve the class imbalance is hard negative mining which samples a few hard anchor boxes for training [27]. Later, RetinaNet [19] introduced focal loss to address the issue in a fundamental way by modifying the loss function to reduce the effect of easy negatives. Including these efforts, various techniques have been proposed to enhance one-stage object detection frameworks [32, 33]. Recently, YOLACT [2] has expanded the capacity of one-stage networks to perform instance segmentation.

2.2 Human-Object Interactions

Human-Object Interaction (HOI) detection has been initially proposed in [10]. Later, human-object detectors have been improved using human body parts [6], human appearance [9], instance appearance [8] and spatial relationship of human-object pairs [8, 15, 17]. Especially, InteractNet [9] extended an existing object detector by introducing an action-specific density map to localize target objects based on the appearance of a detected human. Note that interaction detection based on visual cues from individual boxes often suffers from the lack of contextual information. So iCAN [8] proposed an instance-centric attention module that extracts contextual features complementary to the features from the localized objects/humans. GPNN [25] proposes a Graph Parsing Neural Network for HOI recognition—a general framework that explicitly represents HOI structures with graphs and automatically infers the optimal graph structures. Deep Contextual Attention [30] leverages contextual information by a contextual attention framework in HOI. Recent works in HOI have also explored external knowledge to improve the performance of HOI detection. Since the performance of HOI detection is dependent on how well we recognize the appearance of human actions, human pose information extracted from external models [3, 5, 13, 7, 16] shows meaningful improvement in performance [17, 11, 29, 35]. Interactiveness Knowledge has also been implemented in previous works [17] by adding an additional inference stage where the model learns the probability of interactiveness by combining multiple HOI training datasets. Linguistic priors and knowledge graphs are also utilized to improve HOI detection performance. These sources are either used directly as an additional feature [24, 31, 12] or features to cluster the objects by their functions [1]. However, all the previous methods are multi-stage detectors focusing on accuracy and they are not suitable for real-time applications.

3 Method

We now introduce our method to detect human-object-interaction. To be specific, the goal is to capture $\langle human, verb, object \rangle$ triplets from an image without any external knowledge. The standard HOI detection benchmarks (e.g., V-COCO and HICO-DET) require the localization and classification of interactions. In this paper, we propose a one-stage HOI detector powered by our *union-level* detector, which directly detects the union region of an interacting pair. Since standard benchmarks require the instance-level localization of humans and objects, we *parallelly* combine the union-level detector and an instance-level detector, which allows more accurate instance-level localization. We name this meta-architecture **UnionDet** shown in Figure 2. Our UnionDet is compatible with any one-stage object detectors such as SSD [20], RetinaNet [18], and STDN [34]. For a fair comparison with baseline HOI detectors, in this paper, we implement our model based on RetinaNet with ResNet50-FPN [14, 18] since it's performance is comparably similar to Faster-RCNN—the dominant backbone network in previous works on HOI in literature.

We discuss new challenges in *union-level* detection and how to address them by the components in our union-level detector, which is the union branch in UnionDet in Figure 2. We explain how to modify a standard instance-level detector in UnionDet for HOI detection and lastly, the details of training and inference are provided.

3.1 Challenges in Union-Level Detection

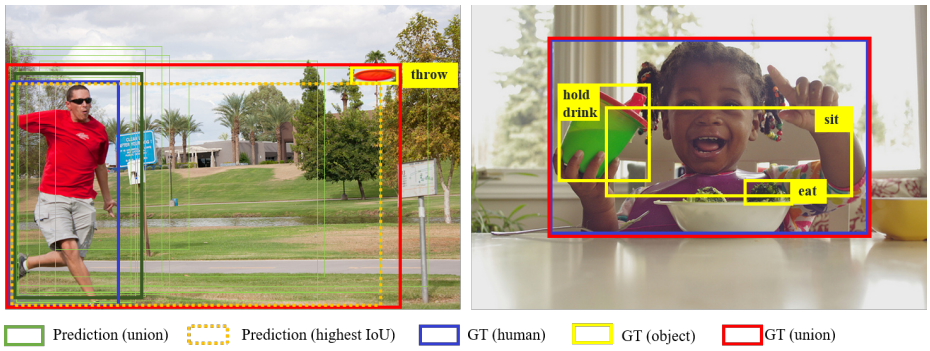

Fig. 1. Technical Challenges in Union Detection. (Left) The box with the highest confidence for the ground-truth action 'throw' is highlighted in bold, and the region with the highest IoU with the ground-truth union region is dotted. As you can see, the highest confidence is biased towards the human region, and despite the high IoU, the dotted region failed to capture the target object (frisbee). (Right) Issues of Overlapping Union regions in *One vs Many* relations. Even though the 3 different $\langle human, object \rangle$ pairs represent a different sets of interactions, all 3 pairs have an identical union region.

The union region of a pair of objects is an intuitive representation of visual relationships [21]. Union-level detection looks similar to instance-level detection. But standard object detectors are not directly applicable due to the following technical challenges. ***First***, a naive union-level detection often suffers from the large bias towards human regions since every union region of HOI has a human. The left figure in Figure 1 shows that union predictions (green bboxes) by a vanilla detector are densely distributed around a human. ***Second***, the standard IoU is not an accurate metric for union bounding box matching. For instance, when one union region has two remote objects, a high IoU with the union region does not ensure that both human and target object is enclosed by the predicted region (see the left figure in Figure 1). ***Lastly***, one union region (or anchor box) may contain multiple interactions and more than two objects. These are often observed especially when a human bounding box contains multiple interacting objects. In the following explanation of Union Branch that performs union-level detection, we show in detail how we address these issues.

3.2 Union-level Detector: Union Branch

Union Branch performs union-level detection which is the essence of our proposed meta-architecture, *UnionDet*. As in Figure 2, Union Branch consists of three sub-branches that share the backbone Feature Pyramid Network. Out of the three sub-branches, the Action Classification sub-branch and the Union Box Regression sub-branch are the main sub-branches that contribute to the inference stage. Action Classification sub-branch performs multi-class classification for the interactions that are related to the union region, and the Union Box Regression sub-branch performs action-agnostic bounding box regression to predict the final union region with multiple actions. Vanilla detection results for union regions can be obtained through these two sub-branches. However, union regions inherently accompany several technical challenges as mentioned above. To address these challenges, Union Branch is trained by new techniques: i) union anchor labeling ii) target object classification loss iii) foreground focal loss. This provides accurate union-level detections even in various distances, see Figure 4.

Union Anchor Labeling. The standard IoU is not suitable for union-level detections. Especially, when objects are small and remote, an anchor box may fail to include either the subject or the object in interaction even when the ground-truth union bounding box and the anchor box has a high (e.g., 0.9) IoU. To address this, we propose a new labeling function to match union-level labels to anchor boxes. Union Branch detects the union regions based on the set of anchors A generated from the backbone Feature Pyramid Network. During the forward propagation of Union Branch, each anchor $a_j \in A$ obtains a multi-label action prediction $\breve{a}_j^{act} \in \mathbb{R}^T$, target object class prediction $\breve{a}_j^{cls} \in \mathbb{R}^K$, and a location prediction $\breve{a}_j^{loc} \in \mathbb{R}^4$. T and K denote the number of interactions and the number of target object categories, respectively. $U_{ij} \in \{0, 1\}$ indicates whether the i_{th} union-level ground-truth label $\breve{g}_i$ matches the j_{th} anchor a_j or

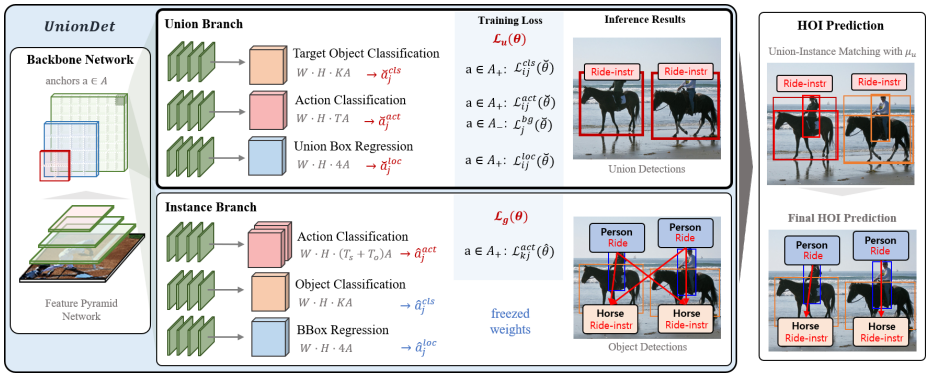

Fig. 2. The overall architecture of UnionDet. Our UnionDet is generally compatible with one-stage object detectors. The feature pyramid obtained from the backbone network is simultaneously fed to Union Branch and Instance Branch. While Union Branch directly captures the region of interaction, Instance Branch performs traditional object detection and action classification for more fine-grained HOI detection results.

not. We propose a new anchor labeling function, which is used during training. Let $\mathbb{1}(\cdot)$ as an indicator function. Given human box $\breve{h}_i$ and object box $\breve{o}_i$ of i-th ground truth union box $\breve{g}_i$, U_{ij} is calculated as:

$$U_{ij} = \mathbb{1}(\text{IoU}(a_j, \breve{g}_i^{loc}) > t_u) \cdot \mathbb{1}\left(\frac{a_j \cap \breve{h}_i^{loc}}{\breve{h}_i^{loc}} > t_h\right) \cdot \mathbb{1}\left(\frac{a_j \cap \breve{o}_i^{loc}}{\breve{o}_i^{loc}} > t_o\right), \quad (1)$$

where t_u, t_h, t_o indicate thresholds for union IoU, human inclusion ratio, and object inclusion ratio. They are set to 0.5 in our experiments. If multiple union-level ground truths are matched, the union with the largest IoU is associated with the anchor box so that an anchor box has at most one ground truth.

After labeling each anchor according to Eq.1, we can build a basic loss function to train the Union Branch. Based on the positive anchor set $A_+ \subseteq A$ where $\{a_j | \sum_i U_{ij} = 1\}$ and the negative anchor samples $A_- \subseteq A$ where $\{a_j | \sum_i U_{ij} = 0\}$, the loss function $\mathcal{L}_u(\breve{\theta})$ is written as

$$\mathcal{L}_u(\breve{\theta}) = \sum_{a_j \in A_+} \sum_{\breve{g}_i \in \breve{\mathcal{G}}} U_{ij} \left[\mathcal{L}_{ij}^{act}(\breve{\theta}) + \mathcal{L}_{ij}^{loc}(\breve{\theta})\right] + \sum_{a_j \in A_-} \mathcal{L}_j^{bg}(\breve{\theta}), \quad (2)$$

where $\breve{\mathcal{G}}$ denotes the ground truth union box set and $\breve{\theta}$ denotes the Union Branch model parameters. $\mathcal{L}_{ij}^{act}(\breve{\theta}) = FL(\breve{a}_j^{act}, \breve{g}_i^{act}, \breve{\theta})$, $\mathcal{L}_{ij}^{loc}(\breve{\theta}) = smooth_{L1}(\breve{a}_j^{loc}, \breve{g}_i^{loc}, \breve{\theta})$, $\mathcal{L}_j^{bg} = FL(\breve{a}_j^{act}, \vec{0}, \breve{\theta})$, where FL and $smooth_{L1}$ each denotes focal loss [19] and Smooth L1 loss, respectively. After training the Union Branch with Eq.2, a vanilla prediction of union regions can be obtained. However, it suffers from 1) the prediction being biased toward the human region, and 2) the noisy learning caused when multiple union regions overlap over each other.

Target Object Classification Loss. To address the first issue where the union prediction is biased toward the human region with a vanilla union-level detector, we design a pretext task, 'target object classification' from the detected union region. This encourages the union-level detector to focus more on target objects and helps the union-level detector to capture the region that encloses the target object. We add the target object classification loss to Eq.2 and the loss function $\mathcal{L}_u$ of Union Branch is given as

$$\mathcal{L}_u(\breve{\theta}) = \sum_{a_j \in A_+} \sum_{\breve{g}_i \in \breve{\mathcal{G}}} U_{ij} \Big[\mathcal{L}_{ij}^{act}(\breve{\theta}) + \mathcal{L}_{ij}^{loc}(\breve{\theta}) + \mathcal{L}_{ij}^{cls}(\breve{\theta}) \Big] + \sum_{a_j \in A_-} \mathcal{L}_j^{bg}(\breve{\theta}), \tag{3}$$

where $\mathcal{L}_{ij}^{cls}(\breve{\theta}) = BCE(\breve{a}_j^{cls}, \breve{g}_i^{cls}, \breve{\theta})$ is the Binary Cross Entropy loss. Though we do not use the target classification score at inference in the final HOI score function, we observed that learning to classify the target objects during training improves the union region detection as well as overall performance (see, Table 3).

Union Foreground Focal Loss. Union regions often overlap over each other when a single person interacts with multiple surrounding objects. The right subfigure in Fig.1 shows an extreme example of overlapping union regions where different interaction pairs have the exactly same union region. In such cases where large portion of union regions overlap with each other, applying vanilla focal loss $\mathcal{L}_{ij}^{act}(\breve{\theta})$ as in Eq.2 and Eq.3 might mistakenly give negative loss to the overlapped union actions (more detailed explanation of such cases will be dealt in our supplement). To address this issue, we deployed a variation of focal loss where we selectively calculate losses for only positive labels for foreground regions. This is implemented by simply multiplying $\breve{g}_i^{act}$ to $\mathcal{L}_{ij}^{act}$, thus our final loss function is written as:

$$\mathcal{L}_u(\breve{\theta}) = \sum_{a_j \in A_+} \sum_{\breve{g}_i \in \breve{\mathcal{G}}} U_{ij} \Big[\breve{g}_i^{act} \cdot \mathcal{L}_{ij}^{act}(\breve{\theta}) + \mathcal{L}_{ij}^{loc}(\breve{\theta}) + \mathcal{L}_{ij}^{cls}(\breve{\theta}) \Big] + \sum_{a_j \in A_-} \mathcal{L}_j^{bg}(\breve{\theta}). \tag{4}$$

3.3 Instance-level Detector: Instance Branch

HOI detection benchmarks require the localization of instances in interactions. For more accurate instance localization, we added Instance Branch to our architecture, see Fig. 2. The Instance Branch parallelly performs instance-level HOI detection: object classification, bbox regression, and action (or *verb*) classification.

Object Detection. The instance-level detector was built based on a standard anchor-based single-stage object detector that performs object classification and bounding box regression. For training, we adopt the focal loss [19] to handle the class imbalance problem between the foreground and background anchors. The object detector is frozen for the V-COCO dataset and fine-tuned for the HICO-DET dataset. More discussion is available in the supplement.

Action Classification. The instance-level detector was extended by another sub-branch for action classification. We treat the action of subjects T_s and objects T_o as different types of actions. So, the action classification sub-branch predicts $(T_s + T_o)$ action types at every anchor. This helps to recognize the direction of interactions and can be combined with the interaction prediction from the Union Branch. For action classification, we only calculate the loss at the positive anchor boxes where an object is located at. This leads to more efficient loss calculation and improvement accuracy.

Training Loss $\mathcal{L}_g$ to Learn Instance-level Actions. Instance Branch and Union Branch share the anchors A generated from the backbone Feature Pyramid Network. The set of instance-level ground-truth annotations for an input image is denoted as $\hat{\mathcal{G}}$. The ground-truth label $\hat{g}_i \in \hat{\mathcal{G}}$ at anchor box i consists of target class label $\hat{g}_i^{cls} \in \{0,1\}^K$, multi-label action types $\hat{g}_i^{act} \in \{0,1\}^{(T_s+T_o)}$ and a location $\hat{g}_i^{loc} \in \mathbb{R}^4$, i.e., $\hat{g}_i = (\hat{g}_i^{cls}, \hat{g}_i^{act}, \hat{g}_i^{loc}) \in \hat{\mathcal{G}}$. During the forward propagation of Instance Branch, each anchor $a_j \in A$ obtains a multi-label action prediction $\hat{a}_j^{act} \in \mathbb{R}^{T_s+T_o}$ and object class prediction $\hat{a}_j^{cls} \in \mathbb{R}^K$ after sigmoid activation, and a location prediction $\hat{a}_j^{loc} = \{x, y, w, h\}$ after bounding box regression. $I_{ij} \in \{0,1\}$ indicates whether object $\hat{g}_i$ matches anchor a_j or not, i.e., $I_{ij} = \mathbb{1}(IoU(a_j, \hat{g}_i) > t)$. We used threshold $t = 0.5$ in the experiments. The Object Classification and BBox Regression sub-branches are fixed with pre-trained weights of object detectors [19], and only the Action Classification sub-branch is trained. Given parameters of Action Classification sub-branch $\hat{\theta}$, the loss for the Instance Branch $\mathcal{L}_g(\hat{\theta})$ will be $\mathcal{L}_g(\hat{\theta}) = \mathcal{L}_{ij}^{act}(\hat{\theta}) = BCE(\hat{a}_j^{act}, \hat{g}_i^{act}, \hat{\theta})$.

3.4 Training UnionDet

The two branches of UnionDet shown in Figure 2 (i.e., the Union Branch and Instance Branch) are trained jointly. Our overall loss is the sum of the losses of both branches, $\mathcal{L}_u$ and $\mathcal{L}_g$, where $\breve{\theta}$ is the parameters for Union Branch and $\hat{\theta}$ is the parameters for Instance Branch ($\theta = \breve{\theta} \cup \hat{\theta}$). The final loss becomes $\mathcal{L}(\theta) = \mathcal{L}_u(\breve{\theta}) + \mathcal{L}_g(\hat{\theta})$. For focal loss, we use $\alpha = 0.25$, $\gamma = 2.0$ as in [19]. Our model is trained with an Adam optimizer with a learning rate of 1e-5.

3.5 HOI Detection Inference

UnionDet at inference time parallelly performs the inference of Union Branch and Instance Branch and then seeks the highly-likely triplets using a summary score combining predictions from the subnetworks. Instance Branch performs object detection and action classification per anchor box. Non-maximum suppression with its object classification scores was performed. Union Branch directly detects the union region that covers the $\langle human, verb, object \rangle$ triplet. For Union Branch, non-maximum suppression was applied with union-level action classification scores. Instead of applying class-wise NMS as in ordinary object detection, we treated different action classes altogether to handle multi-label predictions of union regions.

Union-Instance Matching. As mentioned in section 3.2, IoU is not an accurate measure for union regions, especially in the case where the target object of the interaction is remote and small. To search for a solid union region that covers the given human box b_h and object box b_o, we search for the union box b_u with our proposed *union-instance matching score* defined as

$$\mu_u = \frac{\text{IoU}(\lceil b_h \cup b_o \rfloor, b_u)}{2} + \frac{1}{2}\sqrt{\frac{(b_h \cap b_u)}{b_h} \cdot \frac{(b_o \cap b_u)}{b_o}}, \tag{5}$$

where $\frac{b_1}{b_2}$ is the ratio of the areas of two bounding boxes b_1 and b_2 and $\lceil \cdot \rfloor$ stands for the tightest bounding box that covers the area. We use this union-instance matching score to calculate the HOI score instead of the standard IoU.

HOI Score. The detections from Union Branch and Instance Branch are integrated. This further improves the accuracy of the final HOI detection. Our HOI score function combines union-level action score s_u^a from Union Branch with the human category score s_h, human action score s_h^a, object class score s_o, instance-level action score s_o^a from Instance Branch. For each $\langle human, object \rangle$ pair, we first identify the best union area with the highest union-instance matching score μ_u in Eq. (5) and then calculate the HOI score $S_{h,o}^a$ as

$$S_{h,o}^a = (s_h \cdot s_h^a + s_o \cdot s_o^a) \cdot (1 + \mu_u \cdot s_u^a). \tag{6}$$

When the action classes do not involve target objects, or no union region is predicted, the score will be $S_{h,o}^a = s_h \cdot s_h^a$ and $S_{h,o}^a = s_h \cdot s_h^a + s_o \cdot s_o^a$, respectively.

The calculation of Eq.(6) has in principle $O(n^3)$ complexity when the number of detections is n. However, our framework calculates the final triplet scores without any additional neural network inference after Union and Instance Branches. The calculation time of Eq. (6) is negligible ($< 1ms$). The end-to-end inference time of our model is marginally increased ($\sim 9ms$) compared to the vanilla object detector (RetinaNet with ResNet50-FPN) thanks to the parallel architecture.

4 Experiments

In this section, we demonstrate the effectiveness of UnionDet in HOI detection. We first describe the two public datasets that we use as our benchmark: V-COCO and HICO-DET. Next, we perform various qualitative and quantitative analysis to show that our union-level detector successfully addresses the proposed technical challenges and captures quality union regions, leading to a fast and accurate one-stage HOI detector.

Datasets. To validate the performance of our model, we evaluate our model on two public benchmark datasets: the V-COCO (*Verbs in COCO*) dataset and HICO-DET dataset. ***V-COCO*** is a subset of COCO and has 5,400 `trainval`

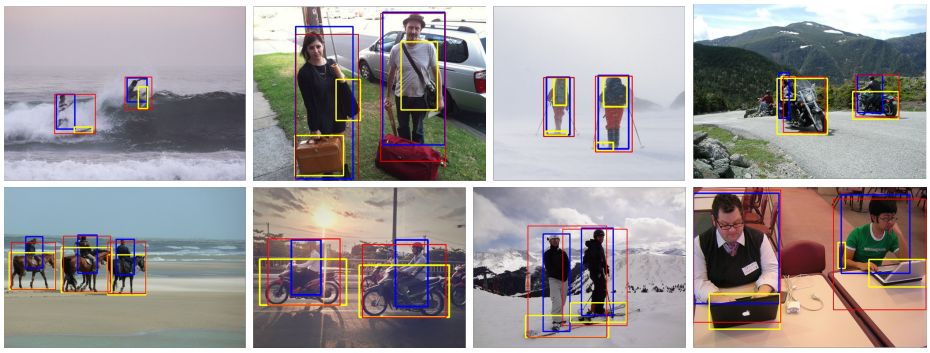

Fig. 3. Union-level detections (red) by Union Branch successfully group correct pairs of humans (blue), and target objects (yellow) among confusing cases caused by multiple triplets with the same action and target object types in an image. Best viewed in color.

images and 4,946 `test` images. For V-COCO dataset, we report the AP_{role} over $T = 29$ interactions. Including the four interaction types that do not involve target objects, V-COCO has $T_s = 26$ active actions and $T_o = 25$ passive actions. As previous works, we exclude the interaction *point* during inference time, because only 31 instances appear in the test set. ***HICO-DET*** [4] is a subset of HICO dataset and has more than 150K annotated instances of human-object pairs in 47,051 images (37,536 training and 9,515 testing) and is annotated with 600 $\langle verb, object \rangle$ interaction types. There are 80 unique object types, identical to the COCO object categories, and $T = 117$ unique verbs. In the HICO-DET dataset, we separate the 117 action classes into a_s, a_o, thus leading into a total action number of $T_s + T_o = 234$. For HICO-DET dataset, we follow the previous settings and report the mAP over three different category sets: (1) all 600 HOI categories in HICO (Full), (2) 138 HOI categories with less than 10 training instances (Rare), and (3) 462 HOI categories with 10 or more training instances (Non-Rare).

Union-level detection. Our union-level detector (Union Branch in UnionDet) directly detects union regions of HOI, see Fig. 3. Interestingly, the union-level detections are useful to disambiguate the confusing pairs with the same action and target object types (e.g., horse, or motorcycle) in an image. For example, when multiple people *ride* the same target objects as in Fig. 3, instance-level appearances are not sufficient to associate the correct pairs. Union-level detections successfully group them using the context in the union-region.

Interactions in various distances. We discussed in Sec. 3.1 that a vanilla object detector is not directly applicable to union-level detection due to the bias toward human regions. This bias gets severer especially when a human interacts with remote target objects. Fig. 4 shows that the bias is addressed by our pretext task 'Target Object Classification' and UnionDet is able to detect target

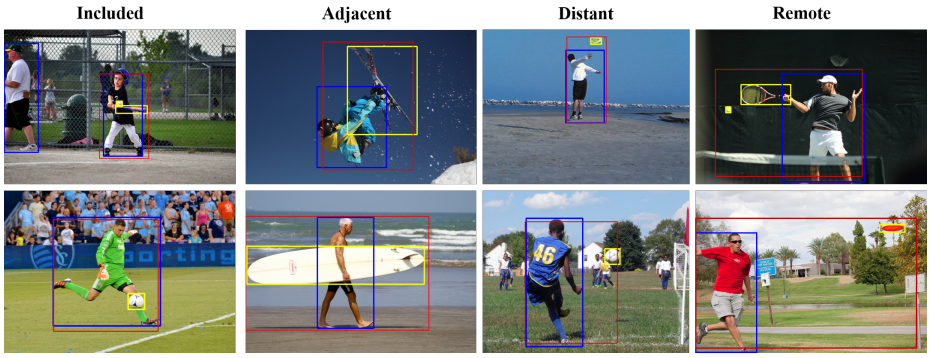

Fig. 4. Our union-level detector (Union Branch) successfully detects the union bounding boxes (red) in various distances: the interactions with included, adjacent, distant and remote target objects. Also, instance-level detections of human (blue), and target objects (yellow) by Instance Branch are visualized. Best viewed in color.

objects for various distances. We show four cases: included ($b_h \supset b_o$), adjacent ($\text{IoU}(b_h, b_o) > 0$), distant ($\text{IoU}(b_h, b_o) = 0$) and remote ($\text{IoU}(b_h, b_o) = 0$ and large distance), where b_h, and b_o are human and object bounding boxes. Especially the fourth column in Fig. 4 shows that UnionDet successfully captures the remote relation with small remote target objects (e.g., tennis ball and frisbee). Our ablation study in Table. 3 provides that the 'Target Object Classification' improves HOI detection. Qualitative results of a vanilla union-level detector without the Target Object Classification sub-branch are provided in the supplement.

HOI detection results. In Figure 5, we highlight the detected humans and objects by object detection with the blue and yellow boxes and the union region predicted by UnionDet with red boxes. The detected human-object interactions are visualized and given a pair of objects, the $\langle human, verb, object \rangle$ triplet with the highest HOI score is listed below each image. Note that our model can detect various types of interactions including one-to-one, many-to-one (multiple persons interacting with a single object), one-to-many (one person interacting with multiple objects), and many-to-many (multiple persons interacting with multiple objects) relationships.

Performance Analysis. We quantitatively evaluate our model on two datasets, followed by the ablation study of our proposed methods. We use the official evaluation code for computing the performance of both V-COCO and HICO-DET. In V-COCO, there are two versions of evaluation but most previous works have not explicitly stated which version was used for evaluation. We have specified the evaluation scenario if it has been referred in either the literature [35], authors' code or the reproduced code. We report our performance in both scenarios for a fair comparison with heterogeneous baselines. In both scenarios, our model out-

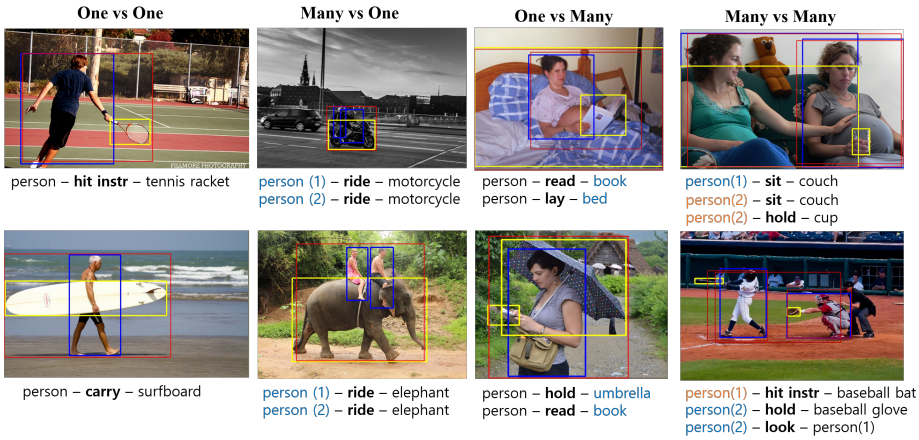

Fig. 5. The final HOI detection by our model combining the predictions of both branches of UnionDet. The columns from the left to the right show one-to-one, many-to-one (multiple persons interacting with a single object), one-to-many (one person interacting with multiple objects), and many-to-many (multiple persons interacting with multiple objects) relationships respectively.

Table 1. Comparison of performance and additional inference time on V-COCO test set. $\cdot_{\#1}$, $\cdot_{\#2}$ each refers to the performance with Scenario#1 and Scenario#2. While achieving $4\times \sim 14\times$ speed-up in additional inference time for interaction prediction, our model surpasses all state-of-the-art performances in both Scenario#1 and Scenario#2. Our model also achieves competitive performance to those that deploy external knowledge or features. Note that our model does not use any external knowledge.

Method	Feature backbone	External Resources	AP_{role}	t(ms)
*Models with **external features***				
Verb Embedding [31]	ResNet50	GloVe [23], VRD [21]	$45.9_{\#1}$	
RP_DC_D [17]	ResNet50	Pose [7, 16]	$47.8_{\#1}$	
RPNN [35]	ResNet50	Keypoint [13]	$47.5_{\#2}$	
PMFNet [29]	ResNet50-FPN	Pose [5]	52.0	
Models with original comparison				
VSRL [10]	ResNet50-FPN	✗	31.8	-
InteractNet [9]	ResNet50-FPN	✗	$40.0_{\#2}$	55
BAR-CNN [15]	ResNet50-FPN	✗	43.6	125
GPNN [25]	**ResNet152**	✗	44.0	40
iCAN [8]	ResNet50	✗	$44.7_{\#1}$	75
TIN (RC_D) [17]	ResNet50	✗	$43.2_{\#1}$	70
DCA [30]	ResNet50	✗	47.3	130
UnionDet (Ours)	ResNet50-FPN	✗	$\mathbf{47.5}_{\#1}$ $\mathbf{56.2}_{\#2}$	**9.06**

Table 2. Performance and additional inference time comparison in HICO-DET. Models with † used a heavier feature extraction backbone (i.e., ResNet152-FPN). Further experimental settings are discussed in detail in our supplement. Our model shows the fastest inference time while achieving state-of-the-art performance across the official evaluation metrics of HICO-DET.

		Default			**Known Object**			
Method	Ext src	Full	Rare	Non Rare	Full	Rare	Non Rare	*t(ms)*
*Models with **external features***								
Verb Embedding [31]	[23],[21]	14.70	13.26	15.13	-	-	-	
TIN (RP_DC_D) [17]	[7, 16]	17.03	13.42	18.11	19.17	15.51	20.26	
Functional Gen. [1]	[22]	21.96	16.43	23.62	-	-	-	
RPNN [35]	[13]	17.35	12.78	18.71	-	-	-	
PMFNet [29]	[5]	17.46	15.65	18.00	20.34	17.47	21.20	
No-Frills HOI [11]†	[3]	17.18	12.17	18.68	-	-	-	
Analogies [24]	[22]	19.4	14.6	20.9	-	-	-	
Models with original comparison								
VSRL [10]	✗	9.09	7.02	9.71	-	-	-	-
HO-RCNN [15]	✗	7.81	5.37	8.54	10.41	8.94	10.85	-
InteractNet [9]	✗	9.94	7.16	10.77	-	-	-	55
GPNN [25]†	✗	13.11	9.41	14.23	-	-	-	40
iCAN [8]	✗	14.84	10.45	16.15	16.26	11.33	17.73	75
TIN (RC_D) [17]	✗	13.75	10.12	15.45	15.34	10.98	17.02	70
DCA [30]	✗	16.25	11.16	17.75	17.73	12.78	19.21	130
Ours	✗	**17.58**	**11.72**	**19.33**	**19.76**	**14.68**	**21.27**	**9.06**

performs state-of-the-art methods [30]. Further, our model shows competitive performance compared to the baselines [31, 17, 35] that leverage heavy external features such as linguistic priors [23, 22] or human pose features [7, 16, 3, 5]. On HICO-DET, our model achieves state-of-the-art performance for both the official 'Default' setting and 'Known Object' setting. For a more comprehensive evaluation of HOI detectors, we also provide the performance of recent works leverage external knowledge [11, 29, 35, 24, 31, 17, 31, 12], although the ***models with External Knowledge*** are beyond the scope of this paper. Note that our main focus is to build a fast single-stage HOI detector from visual features.

Our **ablation study** in Table 3 shows that each component (foreground focal loss, target object classification loss $\mathcal{L}_{ij}^{cls}(\breve{\theta})$, union matching function μ_u) in our approach improves the overall performance of HOI detection.

Interaction Prediction Time. We measured inference time on a single Nvidia GTX1080Ti GPU. Our model achieved the fastest 'end-to-end' inference time (**77.6 ms**). However, the end-to-end inference time is not suitable for fair comparison since the end-to-end computation time of one approach may largely vary depending on the base networks or the backbone object detector. Therefore, we here compare the ***additional*** time for interaction prediction, excluding the time for object detection. The detailed analysis of end-to-end time will also be provided in the supplement. Our approach increases the minimal inference time on top of a standard object detector by eliminating the additional pair-wise neural network inference on detected object pairs, which is commonly required in previ-

Table 3. Ablation Study on V-COCO test set of our model. The first row shows the performance with only the Instance Branch. It can be observed that our proposed Union Branch plays a significant role in HOI detection. The second∼fourth row each shows the performance without the Target Object Classification loss $\mathcal{L}_{ij}^{cls}(\breve{\theta})$, μ_u substituted with standard IoU score in Eq.6, Foreground Focal Loss replaced with ordinary Focal Loss, respectively.

	UnionDet components				
Union Branch	FFL	$\mathcal{L}_{ij}^{cls}(\breve{\theta})$	μ_u	$Sce.\#1$	$Sce.\#2$
-	-	-	-	38.4	51.0
✓	✓	-	✓	44.8	53.5
✓	✓	✓	-	45.0	53.6
✓	-	✓	✓	46.9	55.6
✓	✓	✓	✓	**47.5**	**56.2**

ous works. Table 1 compares the inference time of the HOI interaction prediction excluding the time of the object detection. Note that compared to other multi-stage pipelines that have heavy network structures after the object detection phase, our model additionally requires significantly less time 9.06 ms (11.7%) compared to the base object detector. Our approach achieves **4X∼14X speed-up** compared to the baseline HOI detection models which require $40ms \sim 130ms$ per image after the object detection phase. Since most multi-stage pipelines have extra overhead for switching heavy models between different stages and saving/loading intermediate results. In a real-world application on a single GPU, the gain from our approach is much bigger.

5 Conclusions

In this paper, we present a novel one-stage human-object interaction detector. By performing action classification and union region detection in parallel with object detection, we achieved the *fastest* inference time while maintaining comparable performance with state-of-the-art methods. Also, our architecture is generally compatible with existing one-stage object detectors and end-to-end trainable. Our model enables a unified HOI detection that performs object detection and human-object interaction prediction at near real-time frame rates. Compared to heavy multi-stage HOI detectors, our model does not need to switch models across different stages and save/load intermediate results. In the real-world scenario, our model will more beneficial.

Acknowledgement

This work was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT)(No.CAP-18-03-ETRI), National Research Foundation of Korea (NRF-2017M3C4A7065887), and Samsung Electronics, Co. Ltd.

References

1. Bansal, A., Rambhatla, S.S., Shrivastava, A., Chellappa, R.: Detecting human-object interactions via functional generalization. In: AAAI. pp. 10460–10469 (2020)
2. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: real-time instance segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9157–9166 (2019)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7291–7299 (2017)
4. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 ieee winter conference on applications of computer vision (wacv). pp. 381–389. IEEE (2018)
5. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
6. Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 51–67 (2018)
7. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343 (2017)
8. Gao, C., Zou, Y., Huang, J.B.: ican: Instance-centric attention network for human-object interaction detection. arXiv preprint arXiv:1808.10437 (2018)
9. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8359–8367 (2018)
10. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
11. Gupta, T., Schwing, A., Hoiem, D.: No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9677–9685 (2019)
12. Gupta, T., Shih, K., Singh, S., Hoiem, D.: Aligned image-word representations improve inductive transfer across vision-language tasks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4213–4222 (2017)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Kolesnikov, A., Kuznetsova, A., Lampert, C., Ferrari, V.: Detecting visual relationships using box attention. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 0–0 (2019)
16. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10863–10872 (2019)
17. Li, Y.L., Zhou, S., Huang, X., Xu, L., Ma, Z., Fang, H.S., Wang, Y., Lu, C.: Transferable interactiveness knowledge for human-object interaction detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3585–3594 (2019)

18. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
20. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European conference on computer vision. pp. 21–37. Springer (2016)
21. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision. pp. 852–869. Springer (2016)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
24. Peyre, J., Laptev, I., Schmid, C., Sivic, J.: Detecting unseen visual relations using analogies. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1981–1990 (2019)
25. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 401–417 (2018)
26. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
28. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
29. Wan, B., Zhou, D., Liu, Y., Li, R., He, X.: Pose-aware multi-level feature network for human object interaction detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9469–9478 (2019)
30. Wang, T., Anwer, R.M., Khan, M.H., Khan, F.S., Pang, Y., Shao, L., Laaksonen, J.: Deep contextual attention for human-object interaction detection. arXiv preprint arXiv:1910.07721 (2019)
31. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
32. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Single-shot refinement neural network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4203–4212 (2018)
33. Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., Ling, H.: M2det: A single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9259–9266 (2019)

34. Zhou, P., Ni, B., Geng, C., Hu, J., Xu, Y.: Scale-transferrable object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 528–537 (2018)
35. Zhou, P., Chi, M.: Relation parsing neural network for human-object interaction detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 843–851 (2019)