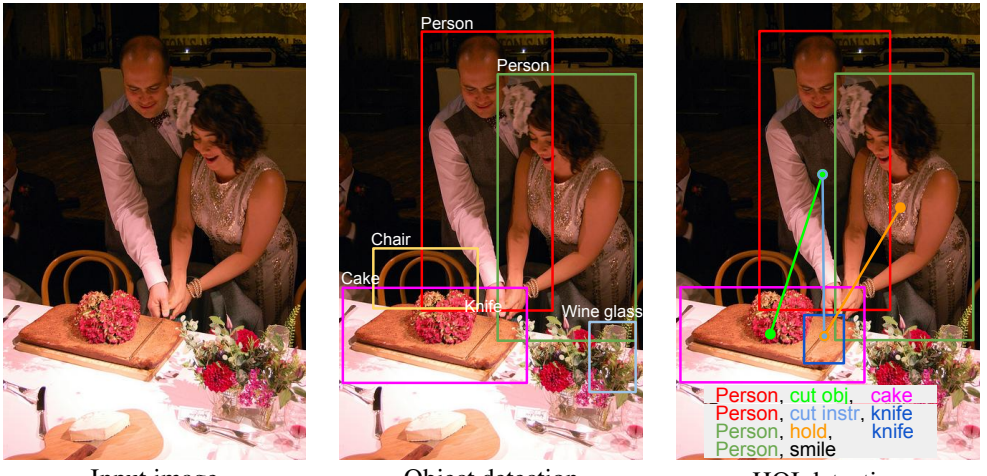
iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection

Chen Gao
chengao@vt.edu
Yuliang Zou
ylzou@vt.edu
Jia-Bin Huang
jbhuang@vt.edu

Virginia Tech
Virginia, USA



Input image Object detection HOI detection

Figure 1: **Human-object interaction detection.** Given an input image (*left*) and the detected object instances in the image (*middle*), our method detects and recognizes the interactions between each person and the objects they are interacting with (*right*).

Abstract

Recent years have witnessed rapid progress in detecting and recognizing individual object instances. To understand the situation in a scene, however, computers need to recognize how humans interact with surrounding objects. In this paper, we tackle the challenging task of detecting human-object interactions (HOI). Our core idea is that the appearance of a person or an object instance contains informative cues on which relevant parts of an image to attend to for facilitating interaction prediction. To exploit this cue, we propose an instance-centric attention module that learns to dynamically highlight regions in an image conditioned on the appearance of each instance. Such an attention-based network allows us to selectively aggregate features relevant for recognizing HOIs. We validate the efficacy of the proposed network on the Verb in COCO and HICO-DET datasets and show that our approach compares favorably with the state-of-the-arts.

1 Introduction

Over the past few years, there has been rapid progress in visual recognition tasks, including object detection [8, 10, 27, 34], segmentation [6, 11, 18, 28], and action recognition [6, 9, 13, 30, 41]. However, understanding a scene requires not only detecting individual object instances but also recognizing the visual relationship between object pairs. One particularly important class of visual relationship detection is detecting and recognizing how each person interacts with the surrounding objects. This task, known as Human-Object Interactions (HOI) detection [4, 14, 15, 16], aims to localize a person, an object, as well as identify the interaction between the person and the object. In Figure 1, we show an example of the HOI detection problem. Given an input image and the detected instances from an object detector, we aim to identify all the triplets $\langle$ `human, verb, object` $\rangle$.

Why HOI? Detecting and recognizing HOI is an essential step towards a deeper understanding of the scene. Instead of "What is where?" (i.e., localizing object instances in an image), the goal of HOI detection is to answer the question "What is happening?". Studying the HOI detection problem also provides important cues for other related high-level vision tasks, such as pose estimation [2, 40], image captioning [24, 39], and image retrieval [21].

Why attention? Driven by the progress in object detection [18, 34], several recent efforts have been devoted to detecting HOI in images [4, 14, 16, 35]. Most existing approaches infer interactions using appearance features of a person and an object as well as their spatial relationship. In addition to using only appearance features from a person, recent action recognition algorithms exploit contextual cues from an image. As shown in Figure 2, examples of encoding context include selecting a secondary box [13], using the union of the human and object bounding boxes [29], extracting features around human pose keypoints [6], or exploiting global context from the whole image [31]. While incorporating context generally helps improve performance, these hand-designed attention regions may not always be relevant for recognizing actions/interactions. For examples, attending to human poses may help identify actions like 'ride' and 'throw', attending to the point of interaction may help recognize actions involving hand-object interaction such as 'drinking from with cup' and 'eat with spoon', and attending to the background may help distinguish between 'hit with tennis racket' and 'hit with baseball ball bat'. To address this limitation, recent works leverage end-to-end trainable attention modules for action recognition [9] or image classification [20]. These methods, however, are designed for *image-level* classification tasks.

Our work. In this paper, we propose an end-to-end trainable *instance-centric* attention module that learns to highlight informative regions using the appearance of a person or an object instance. Our intuition is that the appearance of an instance (either human or an object) provides cues on where in the image we should pay attention to. For example, to better determine whether a person is carrying an object, one should direct its attention to the region around the person's hands. On the other hand, given a bicycle in an image, attending to the pose of the person nearby helps to disambiguate the potential interactions involved with object instance (e.g., riding or carrying a bike). The proposed instance-centric attention network (iCAN) dynamically produces an attentional map for each detected person or object instance highlighting regions relevant to the task. We validate the efficacy of our network design on two large public benchmarks on HOI detection: Verbs in COCO (V-COCO) [16] and Humans Interacting with Common Objects (HICO-DET) [4] datasets. Our results show that the proposed iCAN compares favorably against the state-of-the-art with around 10% relative improvement on V-COCO and 29% on HICO-DET with respect to the existing best-performing methods.

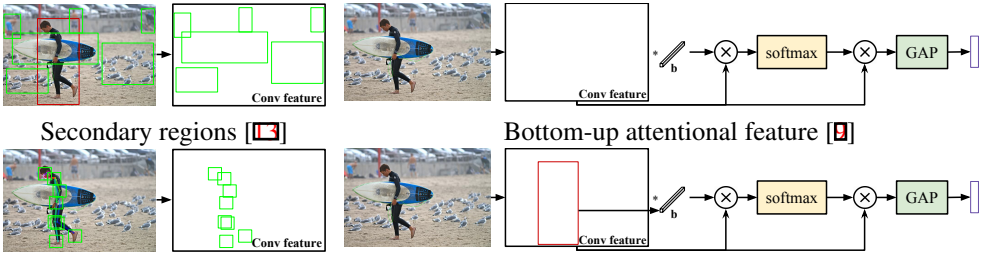


Figure 2: **Examples of contextual features.** Different ways of capturing contextual cues from an image in addition to using the bounding boxes of persons and objects.

Our contributions. We make the following four contributions.

- We introduce an instance-centric attention module that allows the network to dynamically highlight informative regions for improving HOI detection.
- We establish new state-of-the-art performance on two large-scale HOI benchmark datasets.
- We conduct detailed ablation study and error analysis to identify the relative contributions of the individual components and quantify different types of errors.
- We release our source code and pre-trained models to facilitate future research. [1]

2 Related Work

Object detection. Object detection [8, 10, 11, 27, 34] is an essential building block for scene understanding. Our work uses the off-the-shelf Faster R-CNN [12, 34] to localize persons and object instances. Given the detected instances, our method aims to recognize interactions (if any) between all pairs of person and object instances.

Visual relationship detection. A number of recent work addresses the problem of detecting visual relationship [1, 7, 19, 22, 23, 32, 43, 44] and generating scene graph [25, 39, 42]. Several papers leverage some forms of language prior [29, 33] to help overcome the problem of large numbers of the relationship subject-predicate-object triplets and limited data samples. Our work focuses on one particular class of visual relationship detection problems: detecting human-object interactions. HOI detection poses additional challenges over visual relationship detection. With human as a subject, the interactions (i.e., the predicate) with objects are a lot more fine-grained and diverse than other generic objects.

Attention. Extensive efforts have been made to incorporate attention in action recognition [6, 13] and human-object interaction tasks [29, 31]. These methods often use hand-designed attention regions to extract contextual features. Very recently, end-to-end trainable attention-based methods have been proposed to improve the performance of action recognition [9] or image classification [20]. However, these methods are designed for *image-level* classification task. Our work builds upon the recent advances of attention-based techniques and extends them to address *instance-level* HOI recognition tasks.

Human-object interactions. Detecting HOI provides a deeper understanding of the situation in a scene. Gupta and Malik [16] first tackle the HOI detection problem — detecting

[1] Project webpage: https://gaochen315.github.io/iCAN/

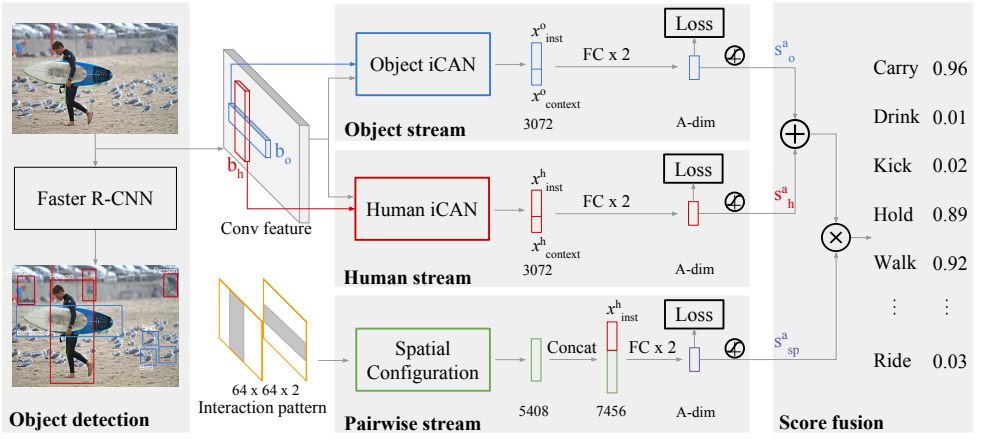


Figure 3: **Overview of the proposed model.** The proposed model consists of following three major streams: (1) a *human stream* for detecting interaction based on human appearance; (2) an *object steam* that predicts the interaction based on object appearance; (3) a *pairwise stream* for encoding the spatial layouts between the human and object bounding boxes. Given the detected object instances by the off-the-shelf Faster R-CNN, we generate the HOI hypothesis using all the human-object pairs. The action scores from individual streams are then fused to produce the final prediction as shown on the right.

people doing actions and the object instances they are interacting with. Associating objects in a scene with various semantic roles with the action leads to a finer-grained understanding of the current state of activity. Very recently, Gkioxari et al. [14] extend the method in [16] by introducing an action-specific density map over target object locations based on the appearance of a detected person. Significantly improved results have also been shown by replacing feature backbone with ResNet-50 [17] and the Feature Pyramid Network [27]. In addition to using object instance appearances, Chao et al. [4] also encode the relative spatial relationship between a person and the object with a CNN. Our work builds upon these recent advances in HOI detection, but with a key differentiator. Existing work recognizes interactions based on individual cues (either human appearance, object appearance, or spatial relationship between a human-object pair). Our key observation is that such predictions inevitably suffer from the lack of contextual information. The proposed instance-centric attention module to extract contextual features complementary to the appearance features of the localized regions (e.g., humans/object boxes) to facilitate HOI detection.

3 Instance-Centric Attention Network

In this section, we present our Instance-centric Attention Network for HOI detection (Figure 3). We start with an overview of our approach (Section 3.1) and then introduce the instance-centric attention module (Section 3.2). Next, we outline the details of the three main streams for feature extraction (Section 3.3): the human stream, the object stream, and the pairwise stream. Finally, we describe our inference procedure (Section 3.4).

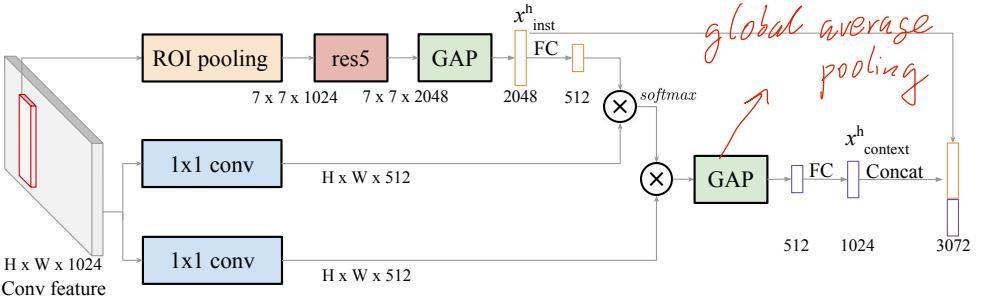


Figure 4: **iCAN module.** Given the convolutional features of the image (shown in gray) and a human/object bounding box (shown in red), the iCAN module extracts the appearance features of the instance $x^{\text{h}}_{\text{inst}}$ (for human) or $x^{\text{o}}_{\text{inst}}$ (for object) as well as the features from the our instance-centric attentional map. For computing the attentional map, we measure the similarity in the embedding space with a bottleneck of 512 channels [37, 38]. Specifically, we embed the image feature using a 1×1 convolution and the instance appearance feature $x^{\text{h}}_{\text{inst}}$ with a fully connected layer. Here `res5` denotes the fifth residual block, `GAP` denotes a global average pooling layer, and `FC` denotes a fully connected layer.

3.1 Algorithm overview

Our approach to human-object interaction detection consists of two main steps: 1) object detection and 2) HOI prediction. First, given an input image we use Faster R-CNN [34] from Detectron [12] to detect all the person/object instances. We denote b_h as the detected bounding box for a person and b_o for an object instance. We use s_h and s_o to denote the confidence scores for a detected person and an object, respectively. Second, we evaluate all the human-object bounding box pairs through the proposed instance-centric attention network to predict the interaction score. Figure 3 shows an overview of the model.

Inference. We predict HOI scores in a similar manner to the existing methods [14, 16]. For each human-object bounding box pair (b_h, b_o), we predict the score $S^a_{h,o}$ for each action $a \in \{1, \cdots, A\}$, where A denotes the total number of possible actions. The score $S^a_{h,o}$ depends on (1) the confidence for the individual object detections (s_h and s_o), (2) the interaction prediction based on the appearance of the person s^a_h and the object s^a_h, and (3) the score prediction based on the spatial relationship between the person and the object s^a_{sp}. Specifically, our HOI score $S^a_{h,o}$ for the human-object bounding box pair (b_h, b_o) has the form:

$$S^a_{h,o} = s_h \cdot s_o \cdot (s^a_h + s^a_o) \cdot s^a_{sp}. \quad (1)$$

For some of the action classes that do not involve any objects (e.g., walk, smile), we use the action score s^a_h from the human stream only. For those actions, our final scores are $s_h \cdot s^a_h$.

Training. As a person can concurrently perform different actions to one or multiple target objects, e.g., a person can 'hit with' and 'hold' a tennis racket at the same time, HOI detection is thus *multi-label classification* problem, where each interaction class is independent and not mutually exclusive. We apply a binary sigmoid classifier for each action category, and then minimize the cross-entropy loss between action score s^a_h, s^a_o, or s^a_{sp} and the ground-truth action label *for each action category*. In the following, we introduce our instance-centric attention module for extracting informative features from an image and then describe a multi-stream network architecture for computing the action scores s^a_h, s^a_o, and s^a_{sp}.

3.2 Instance-centric attention module

In this section, we introduce the instance-centric attention module for extracting contextual features from an image. Figure 4 shows the detailed procedure using human as an instance for clarity. Using an object as an instance is straightforward.

We first extract instance-level appearance feature x^h_{inst} using the standard process, e.g., applying ROI pooling, a residual block, followed by global average pooling. Next, our goal is to dynamically generate an attention map conditioned on the object instance of interest. To do so, we embed both the instance-level appearance feature x^h_{inst} and the convolutional feature map onto a 512-dimensional space and measure the similarity in this embedding space using vector dot product. We can then obtain the instance-centric attentional map by applying softmax. The attentional map highlights relevant regions in an image that may be helpful for recognizing HOI associated with the given human/object instance. Using the attentional map, we can extract the contextual feature x^h_{context} by computing the weighted average of the convolutional features. The final output of our iCAN module is a concatenation of instance-level appearance feature x^h_{inst} and the attention-based contextual feature x^h_{context}.

Our iCAN module offers several advantages over existing approaches. First, unlike hand-designed contextual features based on pose [6], entire image [31], or secondary regions [13], our attention map are automatically learned and jointly trained with the rest of the networks for improving the performance. Second, when compared with attention modules designed for image-level classification, our *instance-centric* attention maps provides greater flexibility as it allows attending to different regions in an image depending on different object instances.

3.3 Multi-stream network

As shown in Figure 3, our network uses three streams to compute the action scores based on human appearance s^a_h, object appearance s^a_o, and their spatial relationship s^a_{sp}.

Human/object stream. For human and object stream, we extract both 1) the instance-level appearance feature x^h_{inst} for a person or x^o_{inst} for an object and 2) the contextual features x^h_{context} (or x^o_{context}) based on the attentional map following the steps outlined in Section 3.2 and Figure 4. With the two feature vectors, we then concatenate them and pass it through two fully connected layers to produce the action scores s^a_h and s^a_o. The score s^a_h from the human stream also allows us to detect actions that do not involve any objects, e.g., walk, smile.

Pairwise stream. While the human and object appearance features contain strong cues for recognizing the interaction, using appearance features alone often leads to plausible but incorrect predictions. To encode the spatial relationship between the person and object, we adopt the two-channel binary image representation in [4] to characterize the interaction patterns. Specifically, we take the union of these two boxes as the reference box and construct a binary image with two channels within it. The first channel has value 1 within the human bounding box and value 0 elsewhere; the second channel has value 1 within the object bounding box and value 0 elsewhere. We then use a CNN to extract spatial features from this two-channel binary image. However, we found that this feature by itself cannot produce accurate action prediction due to the coarse spatial information (only two bounding boxes). To address this, we concatenate the spatial feature with the human appearance feature x^h_{inst}. Our intuition is that the appearance of the person can greatly help disambiguate different actions with similar spatial layouts, e.g., riding vs. walking a bicycle.

3.4 Efficient inference

Following Gkioxari et al. [14] we compute the scores for the triplets in a cascade fashion. We first compute the scores from the human and the object stream action classification head, for each box b_h and b_o, respectively. This first step has a complexity of $O(n)$ for n human/object instances. The second step involves computing scores of all possible human-object pairs. While the second step has a complexity of $O(n^2)$, computing the scores $S^a_{h,o}$, however, is very efficient as it involves summing a pair of scores from the human stream s^a_h and object stream s^a_o (which are already computed and cached in the first step).

Late vs. early fusion. We refer our approach using the pairwise summing scores method as *late fusion* (because the action scores are independently predicted from the human/object streams first and then summed later). We also implement a variant of iCAN with *early fusion*. Specifically, we first concatenate all the features from human iCAN, object iCAN, and the pairwise stream and use two fully connected layers to predict the action score. Unlike late fusion, the early fusion approach needs to evaluate the scores from all human-object pairs and thus have slower inference speed and does not scale well for scenes with many objects.

4 Experimental Results

We evaluate the performance of our proposed iCAN model and compare with the state-of-the-art on two large-scale HOI benchmark datasets. Additional results including detailed class-wise performance and error diagnosis can be found in the supplementary material. The source code and the pre-trained models are available on our project page.

4.1 Experimental setup

Datasets. V-COCO [16] is a subset of the COCO dataset [26] that provides HOI annotations. V-COCO includes a total of 10,346 images containing 16,199 human instances. Each person is annotated with a binary label vector for 26 different actions (where each entry indicates whether the person is performing a certain action). Each person can perform multiple actions at the same time, e.g., holding a cup while sitting on a chair. HICO-DET [3] is a subset of the HICO dataset [3]. HICO-DET contains 600 HOI categories over 80 object categories (same as [26]), and provides more than 150K annotated instances of human-object pairs.

Evaluation metrics. We evaluate the HOI detection performance using the commonly used role mean average precision (role mAP) [16] for both V-COCO and HICO datasets. The goal is to detect the agent and the objects in the various roles for the action, denoted as the $\langle$ `human, verb, object` $\rangle$ triplet. A detected triplet is considered as a true positive if both the predicted human and object bounding boxes b_h and b_o have IoUs ≥ 0.5 w.r.t to the ground truth annotations and has the correct action label.

Implementation details. We use Detectron [12] with a feature backbone of ResNet-50-FPN [27] to generate human and object bounding boxes. We keep human boxes with scores s_h higher than 0.8 and object boxes with scores s_o higher than 0.4. We implement our network based on Faster R-CNN [34] with a ResNet-50 [17] feature backbone. [2] We train our network for 300K iterations on the V-COCO *trainval* set with a learning rate of 0.001, a weight decay of 0.0001, and a momentum of 0.9. Training our network on V-COCO takes 16 hours one

[2] We believe that using ResNet-50-FPN based on the Detectron framework for jointly training object detection and HOI detection could lead to improved performance.

Table 1: Performance comparison with the state-of-the-arts on the V-COCO *test* set.

Method	Feature backbone	AP_{role}
Model C of [16] (implemented by [14])	ResNet-50-FPN	31.8
InteractNet [14]	ResNet-50-FPN	40.0
BAR-CNN [22]	Inception-ResNet [36]	41.1
iCAN (ours) w/ late fusion	ResNet-50	44.7
iCAN (ours) w/ early fusion	ResNet-50	**45.3**

Table 2: Performance comparison with the state-of-the-arts on HICO-DET *test* set. The results from our model are from *late fusion*.

		Default			Known Object		
Method	Feature backbone	Full	Rare	Non Rare	Full	Rare	Non Rare
Shen et al. [35]	VGG-19	6.46	4.24	7.12	-	-	-
HO-RCNN [4]	CaffeNet	7.81	5.37	8.54	10.41	8.94	10.85
InteractNet [14]	ResNet-50-FPN	9.94	7.16	10.77	-	-	-
iCAN (ours)	ResNet-50	**12.80**	**8.53**	**14.07**	**14.70**	**10.79**	**15.87**

single P100 NVIDIA GPU. For HICO-DET, training the network on the *train* set takes 47 hours. Using a single NVIDIA P100 GPU, our method (with late score fusion) takes less than 75ms to process an image of size 480×640 (including ResNet-50 feature extraction, multi-stream network, attention-based feature extraction, and HOI recognition). We apply the *same* training and inference procedures for both V-COCO and HICO-DET datasets. Please refer to the supplementary material for additional implementation details.

4.2 Quantitative evaluation

We present the overall quantitative results in terms of AP_{role} on V-COCO in Table 1 and HICO-DET in Table 2. For V-COCO, the proposed instance-centric attention network achieves sizable performance boost over competing approaches [14, 16, 22]. For HICO-DET, we also demonstrate that our method compares favorably against existing methods [4, 14, 35]. Following the evaluation protocol [4], we report the quantitative evaluation of all, rare, and non-rare interactions with two different settings: 'Default' and 'Known Object'. Compare to [14], we achieve an absolute gain of 2.86 points over the best-performing model (InteractNet) [14] under the *full category* of the 'Default' setting. This amounts to a relative improvement of 28.8%.

4.3 Qualitative evaluation

HOI detection results. Here we show sample HOI detection results on the V-COCO dataset and the HICO-DET dataset. We highlight the detected human and object with red and blue bounding boxes, respectively. Figure 5 shows that our model can predict HOIs in a wide variety of different situations. Figure 6 shows that our model is capable of predicting different action with the objects from the same category. Figure 7 presents two examples of detecting a person interacting with different objects.

Attention map visualization. Figure 8 visualizes the human-centric and object-centric attention maps. The human-centric attention map often focuses on the surrounding objects that help disambiguate action prediction for the detected person. The object-centric atten-

Figure 5: **Sample HOI detections on the V-COCO *test* set.** Our model detects various forms of HOIs in everyday photos. For actions 'ride', 'eat', 'lay' and 'drink', our model detects a diverse set of objects that the persons are interacting with in different situations.

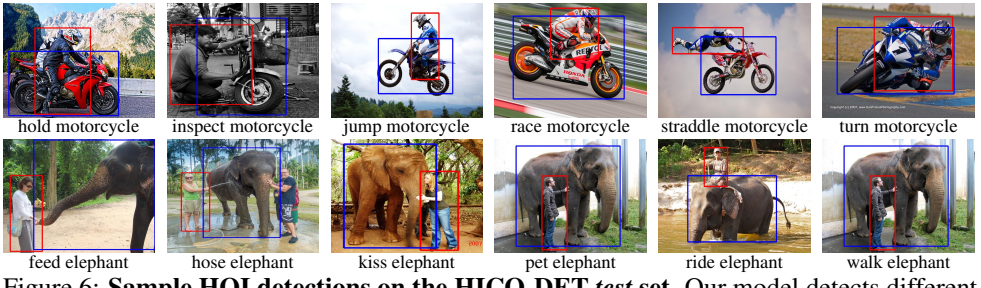


Figure 6: **Sample HOI detections on the HICO-DET *test* set.** Our model detects different types of interactions with objects from the same category.

tion map, on the other hand, highlights the informative human body part, e.g., in the first image, the attention map highlights the right hand of the person even though he was not holding anything with his right hand. We also show an example of two detected persons doing different actions. We show the cropped 100×100 patches centered at the peaks of the generated human-centric attentional maps. The highlighted regions roughly correspond to the objects they are interacting with.

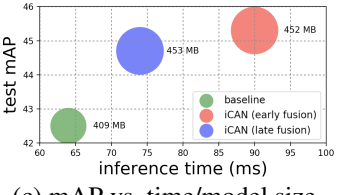
Table 3: Ablation study on the V-COCO *test* dataset

	AP_{role}
None	42.5
Full image [ ]	42.9
Bottom-up att. [ ]	<u>43.2</u>
Inst-centric att. (ours)	**44.7**

(a) Scene feature

Human	Object	AP_{role}
-	-	42.5
✓	-	<u>44.4</u>
-	✓	44.3
✓	✓	**44.7**

(b) Human/Object stream



(c) mAP vs. time/model size

4.4 Ablation study

Contextual feature. Recognizing the correct actions using only human and object appearance features remains challenging. Building upon a strong baseline that does not use any

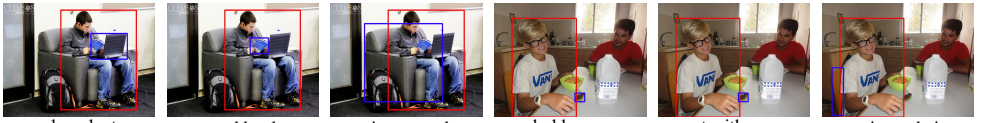


Figure 7: **Detecting multiple actions.** Our model detects an individual taking multiple actions and interacting with different objects, e.g., the person sitting on the couch is reading a book while working on a laptop.

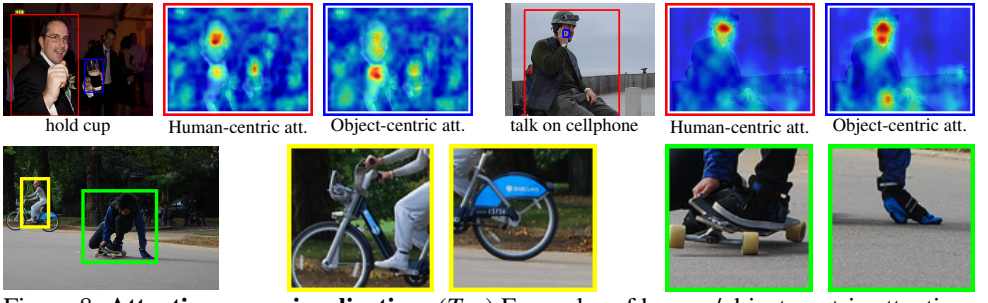


Figure 8: **Attention map visualization.** (*Top*) Examples of human/object-centric attention maps. (*Bottom*) 100×100 patches centered at the peaks of the human-centric attentional maps generated by the two persons. Our model learns to attend to objects (e.g., bicycle, skateboard) and the human poses.

contextual features, we investigate several different approaches for incorporating contextual information of the image, including bottom-up attention map [9], convolutional features from the entire image [31], and the proposed instance-centric attention map. Table 3(a) shows that incorporating contextual features generally helps improve the HOI detection performance. Our approach provides a larger boost over methods that use features without conditioning on the instance-of-interest.

Human-centric vs. object-centric. Table 3(b) validates the importance of leveraging both human-centric and object-centric attentional maps.

mAP vs. time vs. memory. Table 3(c) characterizes the variants of the proposed iCAN using their trade-off in terms of mAP, inference time, and memory usage. Our model with early fusion achieves the best performance on V-COCO dataset. However, this comes at the cost of expensive evaluation of all possible human-object pairs in an image based on their appearance features, and slower training and testing time.

5 Conclusions

In this paper, we propose an instance-centric attention module for HOI detection. Our core idea is to learn to highlight informative regions from an image using the appearance of a person and an object instance, which allow us to gather relevant contextual information facilitating HOI detection. We validate the effectiveness of our approach and show a sizable performance boost compared to the state-of-the-arts on two HOI benchmark datasets. In this work we consider class-agnostic instance-centric attention. We believe that the class-dependent instance-centric attention is a promising future direction.

Acknowledgements. This work was supported in part by NSF under Grant No. (1755785). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU.

References

[1] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *CVPR*, 2015.

[4] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2017.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4), 2017.

[6] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-CNN: Pose-based cnn features for action recognition. In *ICCV*, 2015.

[7] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *CVPR*, 2017.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *NIPS*, 2016.

[9] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *NIPS*, 2017.

[10] Ross Girshick. Fast r-cnn. In *CVPR*, 2015.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[12] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018.

[13] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *ICCV*, 2015.

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

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

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

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

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[19] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017.

[20] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. In *ICLR*, 2018.

[21] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015.

[22] Alexander Kolesnikov, Christoph H Lampert, and Vittorio Ferrari. Detecting visual relationships using box attention. *arXiv preprint arXiv:1807.02136*, 2018.

[23] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. ViP-CNN: Visual phrase guided convolutional neural network. In *CVPR*, 2017.

[24] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *CVPR*, 2017.

[25] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *ICCV*, 2017.

[26] 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.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[29] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016.

[30] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[31] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, 2016.

[32] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Weakly-supervised learning of visual relations. In *ICCV*, 2017.

[33] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive linguistic cues. In *ICCV*, 2017.

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

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

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[39] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

[40] Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Does human action recognition benefit from pose estimation? In *BMVC*, 2011.

[41] Bangpeng Yao, Aditya Khosla, and Li Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

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

[43] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. PPR-FCN: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *ICCV*, 2017.

[44] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *ICCV*, 2017.