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

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

Virginia Tech
Virginia, USA

In this supplementary document, we provide additional experimental results to complement the main paper. First, we provide more implementation details of the proposed model. Second, we provide per-class performance on V-COCO dataset. Third, we analyze several different types of error caused by the proposed iCAN model.

1 Implementation Details

Network training. To augment the positive training data, we apply spatial jitter to the human box and object box in the ground truth triplet to generate additional 15 positive training triplets. We construct negative training examples by pairing all the Detectron [1] detected humans and objects that are *not* annotated in the ground truth labels. The losses for both human and object stream are only computed on the 16 positive triplets. The loss for the spatial configuration branch is computed on both the 16 positive triplets and negative triplets.

Evaluation on the V-COCO dataset. The V-COCO dataset assumes one human can only performs one action on one object. Consequently, similar to [2, 3], rather than scoring every potential triplet, we report the object box with the maximum action score $S^a_{h,o}$ for each human-action pair. That is, we compute:

$$b_{o^*} = \arg\max_{b_o} \; s_o \cdot s^a_{h,o} \cdot s^a_{sp}, \tag{1}$$

where s_o is the class score from Detectron [1]. After selecting the best object b_o for human b_h and action a, we have the finial score $S^a_{h,o} = s_h \cdot s_o \cdot s^a_{h,o} \cdot s^a_{sp}$ for detected triplet $\langle b_h, a, h_o \rangle$. To address action classes that have no interaction with target objects, we use the human stream to compute the scores on human boxes.

© 2018. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

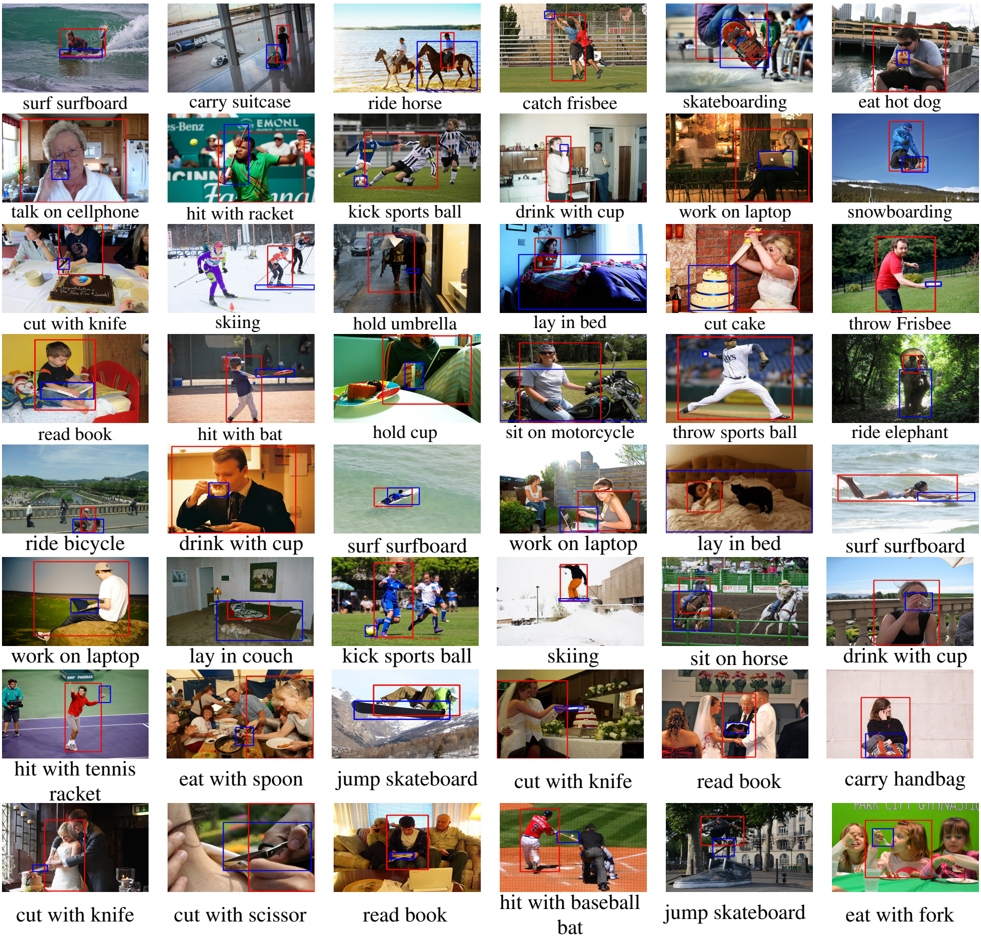
Table 1: Detailed results on V-COCO *test* dataset.

		InteractNet [2]	iCAN	iCAN (early fusion)
		ResNet-50-FPN	ResNet-50	ResNet-50
carry		33.1	34.4	32.0
catch		42.5	46.7	47.6
drink		33.8	27.8	32.2
hold		26.4	24.8	29.1
jump		45.1	52.0	51.5
kick		69.4	63.7	66.9
lay		21.0	23.4	22.4
look		20.2	16.8	26.5
read		23.9	23.1	30.7
ride		55.2	63.9	61.9
sit		19.9	27.1	26.0
skateboard		75.5	83.8	79.4
ski		36.5	42.5	41.7
snowboard		63.9	71.6	74.4
surf		65.7	79.5	77.2
talk-on-phone		31.8	51.0	52.8
throw		40.4	42.2	40.6
work-on-computer		57.3	62.4	56.3
cut	(object)	23.0	36.8	34.8
	(instrument)	36.4	36.8	37.2
eat	(object)	32.4	37.8	37.7
	(instrument)	2.0	6.6	8.3
hit	(object)	62.3	42.4	46.1
	(instrument)	43.3	75.1	74.1
mean AP role		40.0	44.7	45.3

Evaluation on the HICO-DET dataset. For HICO-DET, we have a list of pre-defined HOI categories of interest. Therefore, for each detected human and object pair, we compute the score for each of the related HOI category, e.g. for object motorcycle, the related HOI categories are *hold*, *inspect*, *jump*, *hop on*, *park*, *push*, *race*, *ride*, *sit on*, *straddle*, *turn*, *walk*, *wash*, or *no interaction*. We set a threshold of 0.6 to filter out most of false positive object detections. For each bounding box pair, we predict the action score $S^a_{h,o}$.

2 Additional Results

Per-class role mAP. We show the detailed AP_{role} for each action class in Table 1. The proposed instance-centric network performs well on actions that have a distinctive scene such as *surf* (79.5%) and *snowboard* (71.6%). We also achieve high AP_{role} for classes that have a distinctive objects associated with the action, e.g., , *hit instrument* (75.1%) and *work on the computer* (62.4%).



surf surfboard | carry suitcase | ride horse | catch frisbee | skateboarding | eat hot dog

talk on cellphone | hit with racket | kick sports ball | drink with cup | work on laptop | snowboarding

cut with knife | skiing | hold umbrella | lay in bed | cut cake | throw Frisbee

read book | hit with bat | hold cup | sit on motorcycle | throw sports ball | ride elephant

ride bicycle | drink with cup | surf surfboard | work on laptop | lay in bed | surf surfboard

work on laptop | lay in couch | kick sports ball | skiing | sit on horse | drink with cup

hit with tennis racket | eat with spoon | jump skateboard | cut with knife | read book | carry handbag

cut with knife | cut with scissor | read book | hit with baseball bat | jump skateboard | eat with fork

Figure 1: **Detection results on V-COCO *test* set.** Our model can detect various forms of human-object interactions in everyday photos.

Visual examples of HOI detections. Here we show additional qualitative results of HOI detection in Figure 1 and Figure 2.

3 Error Analysis

Inspired by [3, 4], we diagnose errors in HOI detection task for a better understanding of the network's weakness. For each action class, we consider the top *num-inst* detections [3], where *num-inst* is the number of ground truth instances for a specific class. Similar to [3], we mainly analyze the following six error types:

1. **incorrect label**: when the person is detected around a ground truth person box but is labeled incorrectly for a certain action.

2. **bck**: when the detected person has an IoU less than 0.1 with any of the ground truth persons.

3. **person misloc**: when the detected person has an IoU between 0.1 and 0.5 with a ground truth person.

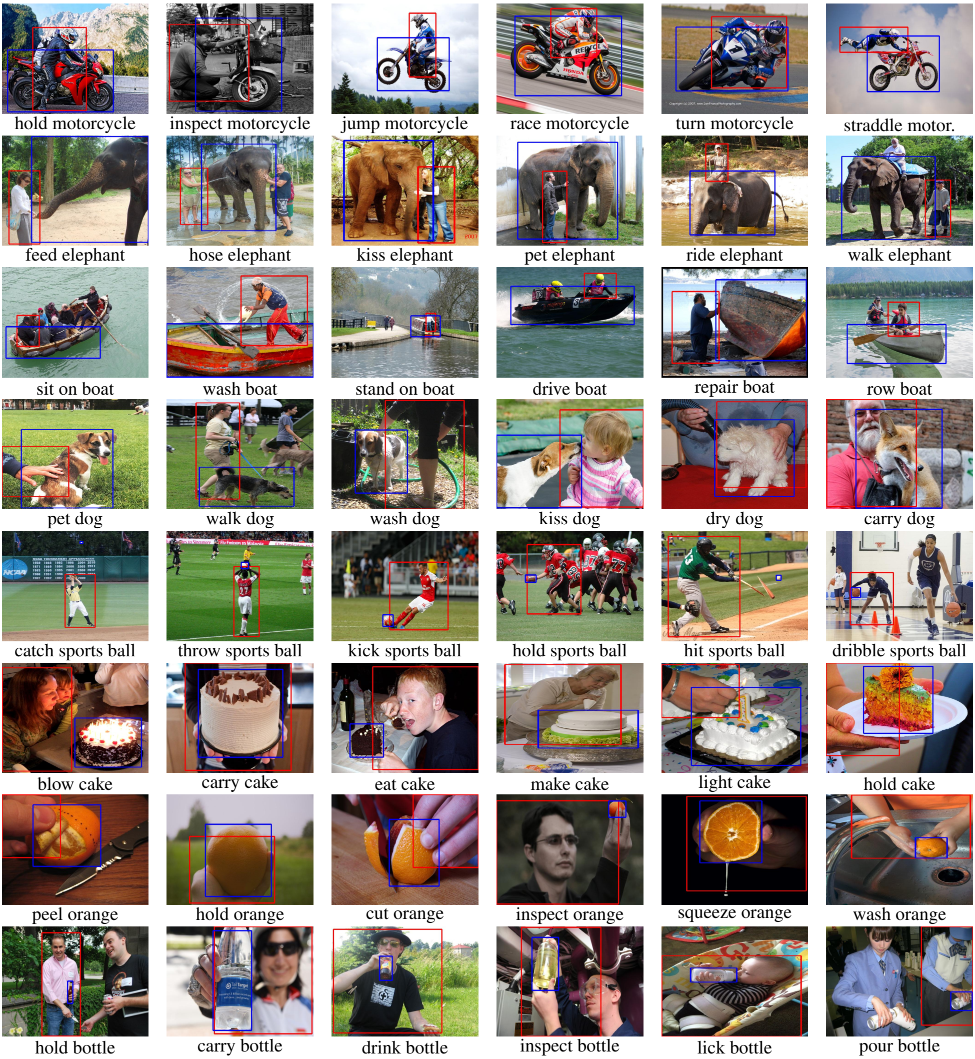
4. **object misloc**: when (1) the detected person has an IoU greater than 0.5 with the ground truth person and (2) the detected object has an IoU between 0.0 and 0.5 with the ground truth object.

5. **mis grouping**: when the detected person has IoU greater than 0.5 with the ground truth person, but the detected object is not associated with the ground truth person (i.e., IoU is 0).

6. **occlusion**: when the detected person has IoU greater than 0.5 with the ground truth person and an object is detected, but there is no ground truth object associated to this person (due to occlusion).

Figure 4 shows the distribution of incorrect detections in the top *num-inst* detections for each action class. The most dominant error for these detections is incorrect classification. It is caused by either incorrect object detection or bias. Figure 3 first row shows some incorrect classification examples. The person in the first two images are incorrectly predicted to *snowboard*. The object in the first two images are incorrectly detected as a snowboard, as a result, the network makes the wrong action detection consequently. Sometimes the action itself is ambiguous. For example, it is extremely difficult to distinguish between the action *throw* and *catch* in the third image in Figure 3. Our network gets confused and predicts both *throw* and *catch* with high confidence. The network also suffers from bias by focusing too much on the object. For example, when it observes a car, it will predict the action *ride* with high confidence, predicts the action *kick* when it observes a football, or predicts the action *eat* when it observes a pizza. Although the spatial configuration stream gives a lower probability, as shown in last three images in the first row.

Another common error is mis-localization. It is caused by either failing to localize human/object or match human and related object. In Figure 3, the second row shows some examples of mis-localization. The red box indicates detected person. The blue box indicates the detected object instance and the green box indicates the ground truth object. The first and second image in the second row are examples of mis-grouping. It is common in actions like *surf*, *snowboard* and *skateboard* because typically there several people performing the same action with a related object in an image. The rest examples show object mis-localization.



hold motorcycle, inspect motorcycle, jump motorcycle, race motorcycle, turn motorcycle, straddle motor.

feed elephant, hose elephant, kiss elephant, pet elephant, ride elephant, walk elephant

sit on boat, wash boat, stand on boat, drive boat, repair boat, row boat

pet dog, walk dog, wash dog, kiss dog, dry dog, carry dog

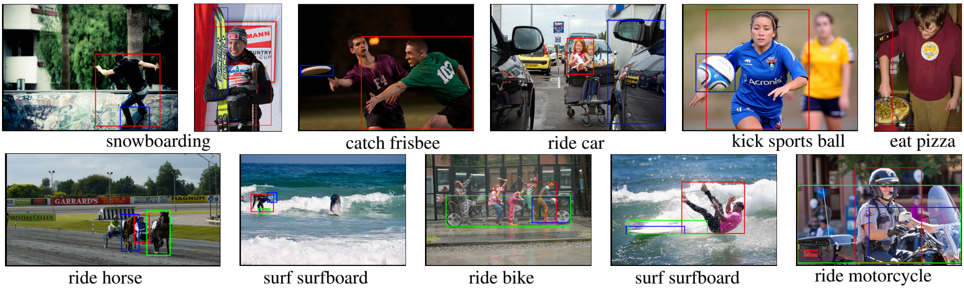
catch sports ball, throw sports ball, kick sports ball, hold sports ball, hit sports ball, dribble sports ball

blow cake, carry cake, eat cake, make cake, light cake, hold cake

peel orange, hold orange, cut orange, inspect orange, squeeze orange, wash orange

hold bottle, carry bottle, drink bottle, inspect bottle, lick bottle, pour bottle

Figure 2: **Detection results on HICO-DET *test* set.**



snowboarding, catch frisbee, ride car, kick sports ball, eat pizza

ride horse, surf surfboard, ride bike, surf surfboard, ride motorcycle

Figure 3: **Visualizations of the incorrect detections.** First row: incorrect label. Second row: mis grouping (1st and 2nd) and object mis-localization (3rd to 5th). Red box indicates the detected person, blue box means the (incorrect) detected object, green box shows the ground truth object annotation.

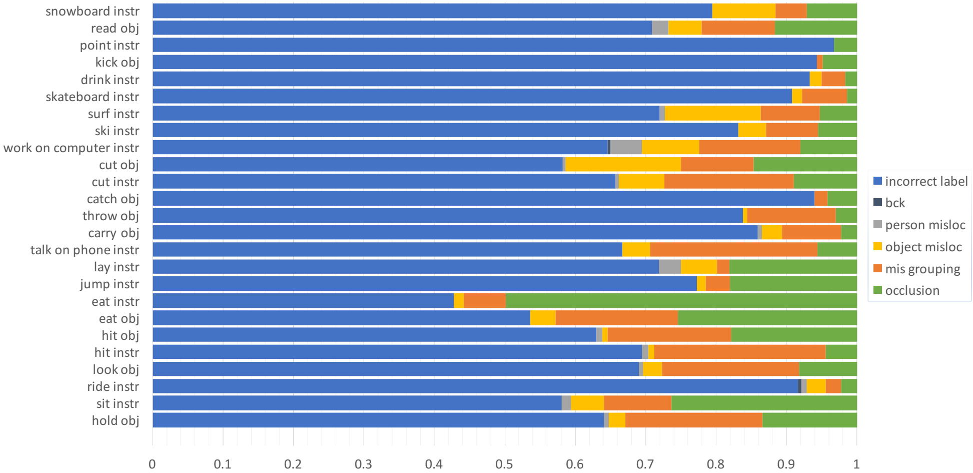


Figure 4: Distribution of the incorrect detections for each action class. **'incorrect label'** refers to when the detected person is not doing this action. **'bck'** indicates when the object detection branch totally fails to localize person (IoU with any ground truth person less than 0.1). **'person misloc'** means when the object detection branch imperfectly localizes the person (regardless of related object). **'object misloc'** refers to when the object detection branch imperfectly localizes the object (while successfully localizes the related person). **'mis-grouping'** indicates when the person is successfully localized but the network fails to match the person to the correct related object. **'occlusion'** means we associate an object instance with a correct detected person, while the object is not annotated in the ground truth due to occlusions.

References

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

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

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

[4] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In *ECCV*, 2012.