Human Object Interaction

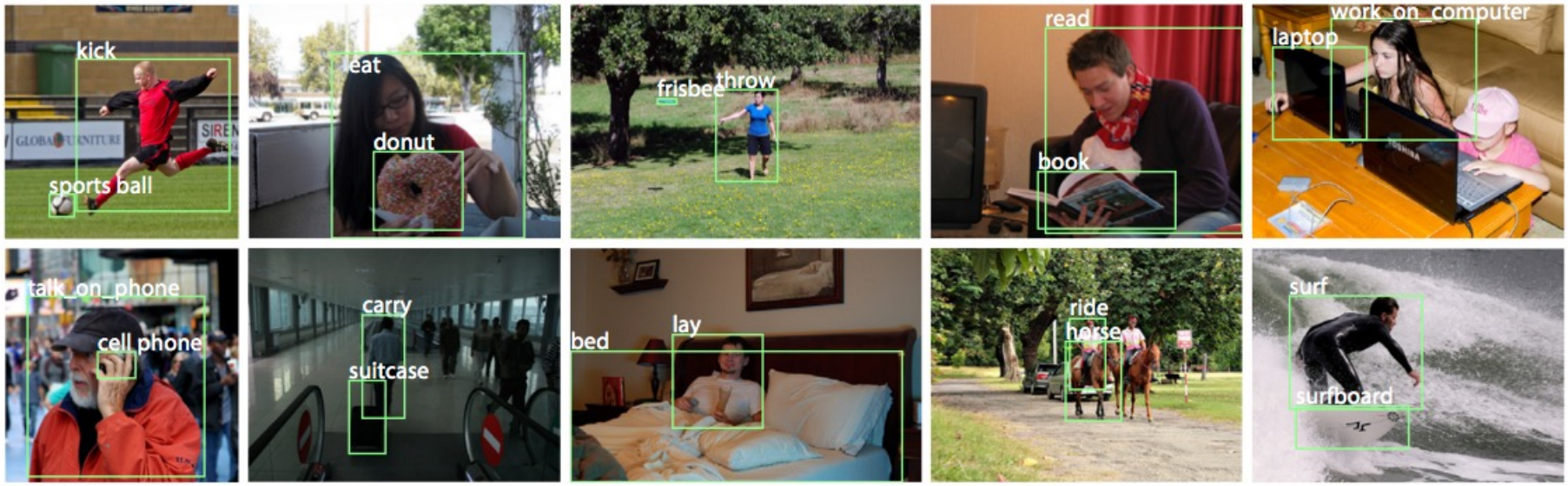
육현준

ETRI
한국전자통신연구원

HOI

Human Object Interaction

- Detection + activity recognition



- ⟨human, verb, object⟩

HOI

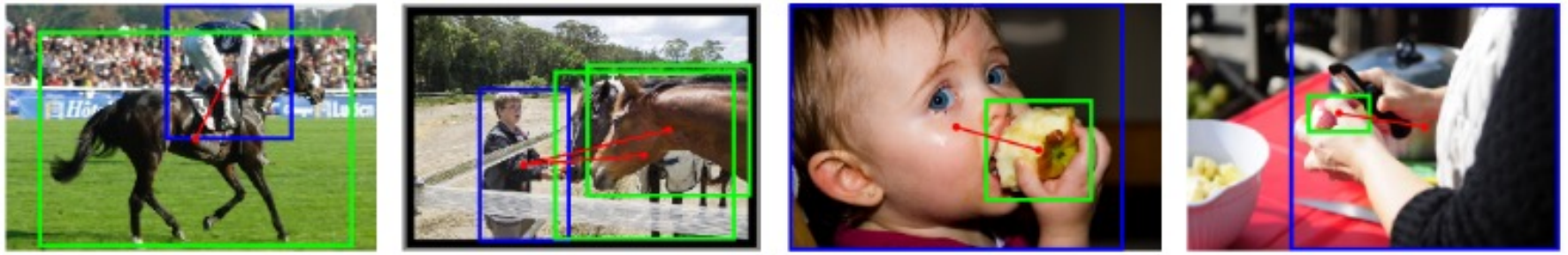
Human Object Interaction

- video

HOI

Dataset

- HICO-DET
- V-COCO
- TUHOI



Sample annotations in the HICO-DET benchmark

http://websites.umich.edu/~ywchao/hico/hoi-det-ui/demo_20171121.html

HOI

Dataset

- HICO-DET
  - annotation

Sample annotations for the description "**A person riding a bicycle.**"

HOI

평가지표

- AP-role
  - AP of the triplet <human, verb, object>
- AP-agent
  - AP of the pair <human, verb>

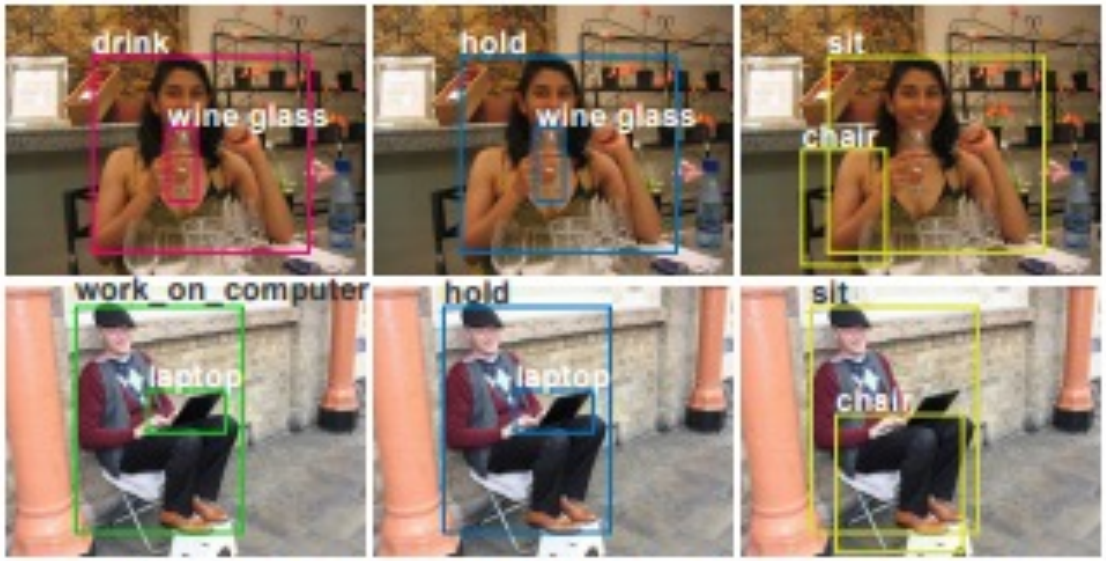


Figure 7. Results of InteractNet on test images. An individual person can take multiple actions and affect multiple objects.

HOI

Failure cases

False Positive

HOI

- Sequential
- Parallel
- Transformer

HOI

Sequential

- branch

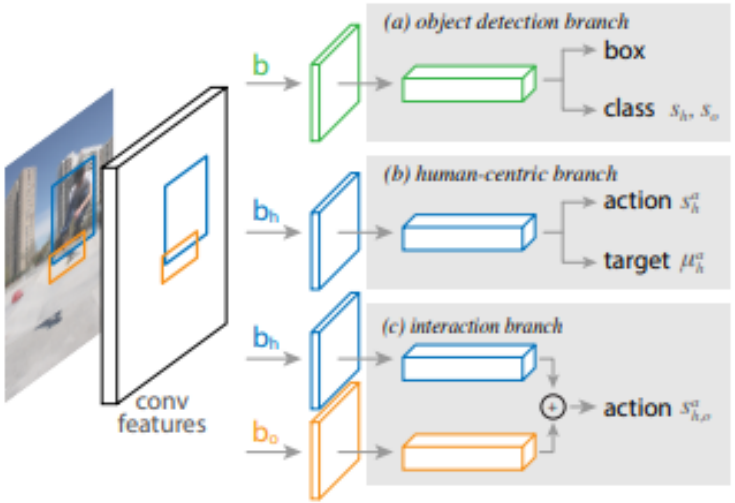


Figure 3. **Model Architecture.** Our model consists of (a) an *object detection* branch, (b) a *human-centric* branch, and (c) an optional *interaction* branch. The person features and their layers are shared between the human-centric and interaction branches (blue boxes).

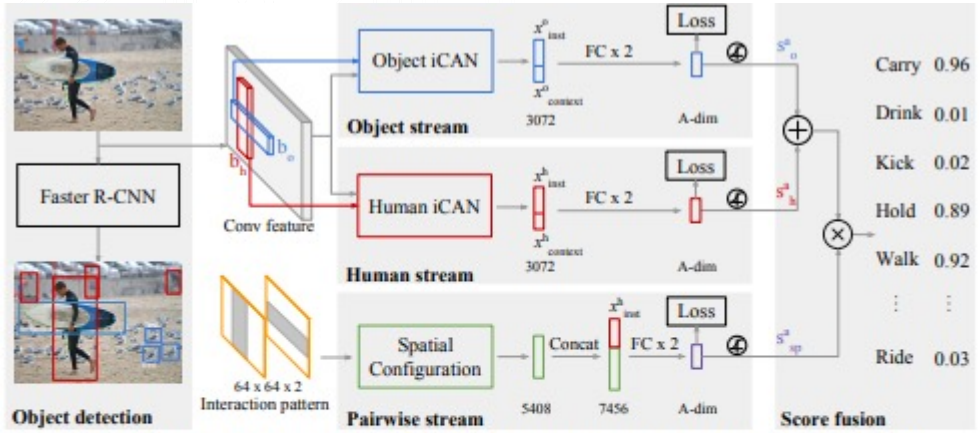


Figure 3: **Overview of the proposed model.** The proposed model consists of following three major streams: (1) a *human stream* for detecting interaction based on human appearance; (2) an *object steam* that predicts the interaction based on object appearance; (3) a *pairwise stream* for encoding the 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.

<interactNet>

<iCAN>

HOI

Sequential

- InteractNet

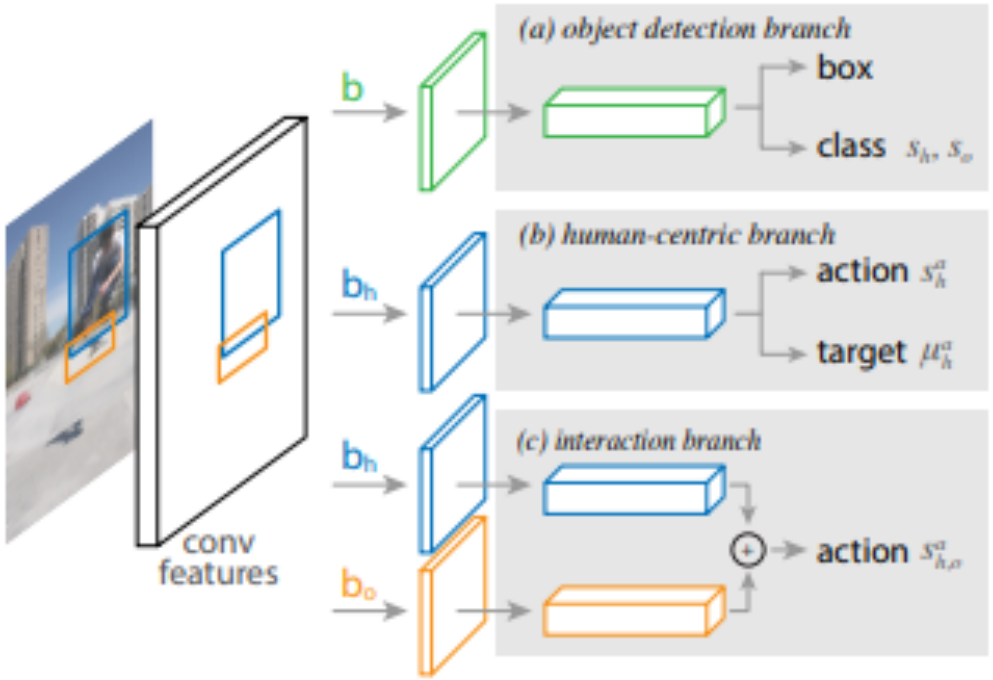
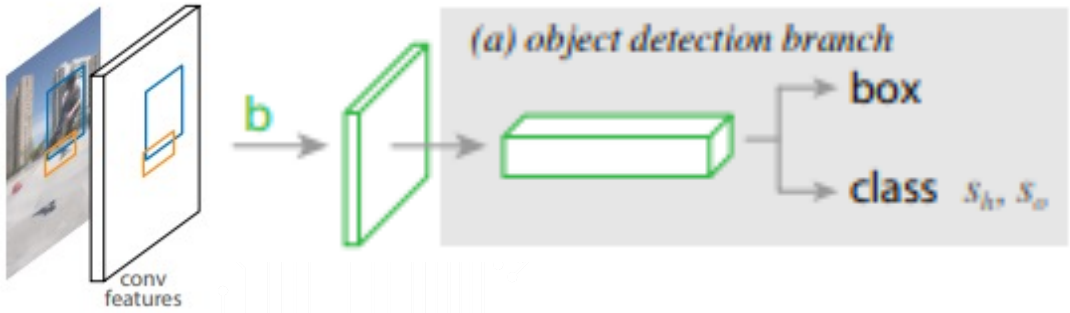


Figure 3. **Model Architecture.** Our model consists of (a) an *object detection* branch, (b) a *human-centric* branch, and (c) an optional *interaction* branch. The person features and their layers are shared between the human-centric and interaction branches (blue boxes).

HOI

Sequential

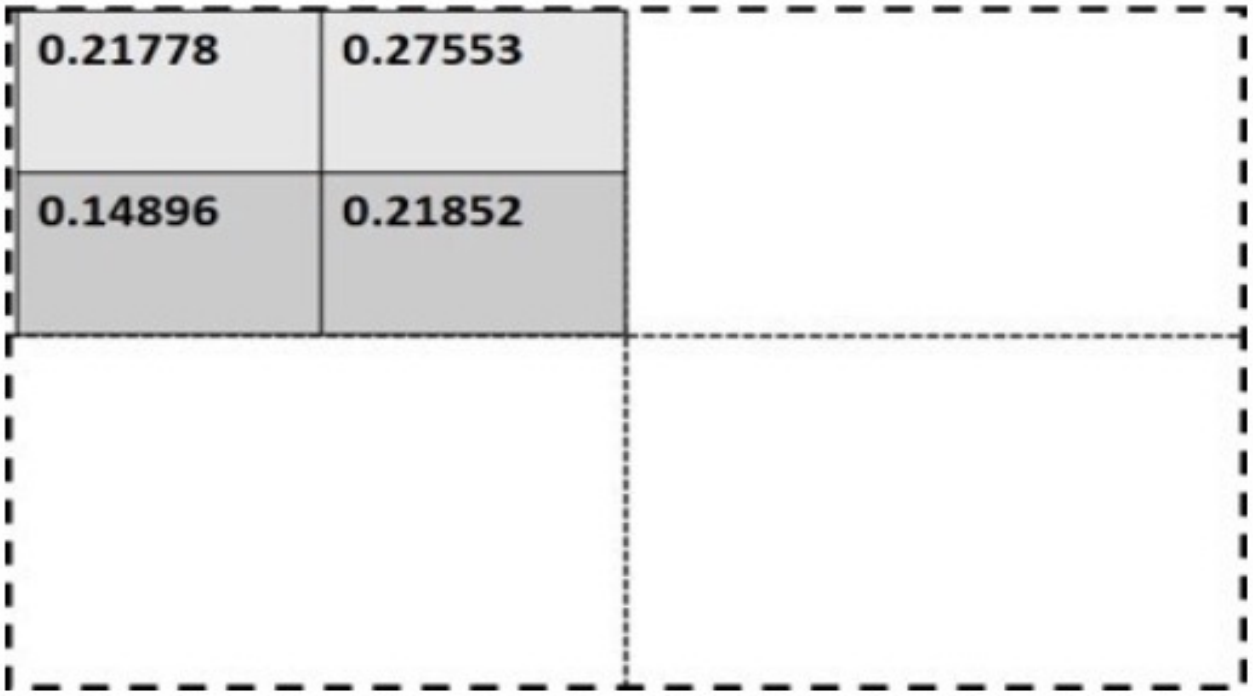
- InteractNet – object detection branch



HOI

Sequential

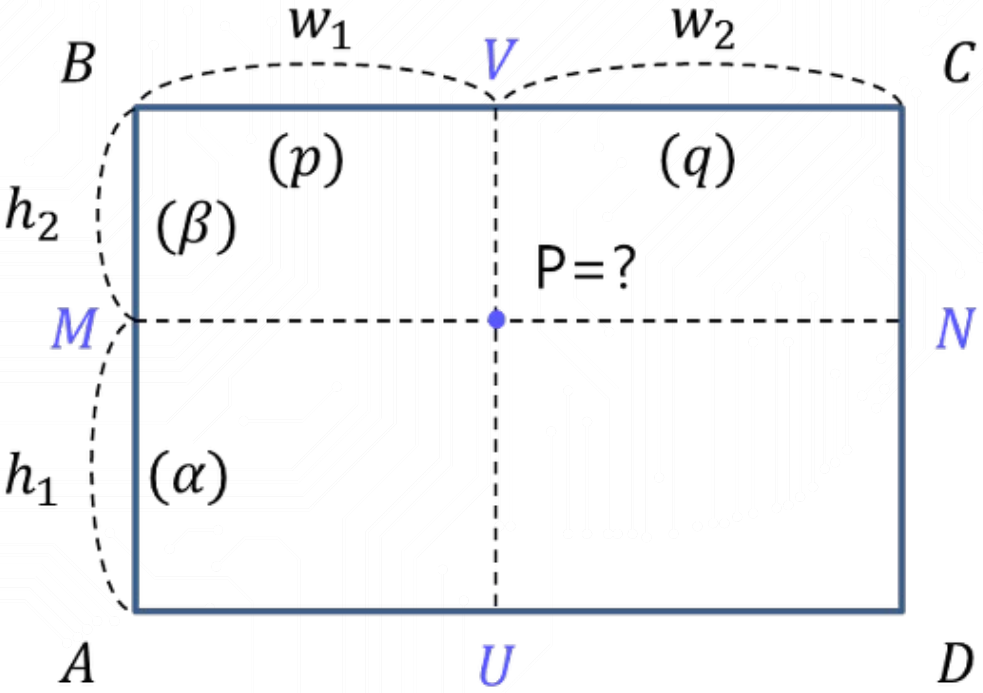
- InteractNet – object detection branch
  - RoiAlign



HOI

Sequential

- InteractNet – object detection branch
  - RoiAlign
    - Bilinear interpolation

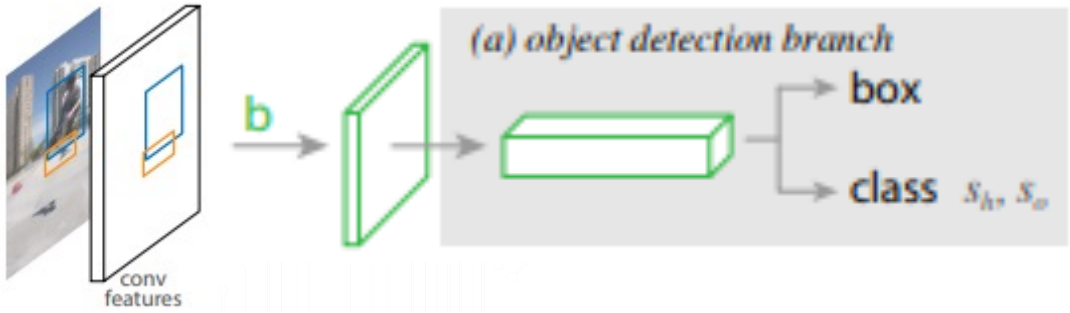


$$P=q(\beta A+\alpha B)+p(\beta D+\alpha C)$$
$$=q\beta A+q\alpha B+p\beta D+p\alpha C$$

HOI

Sequential

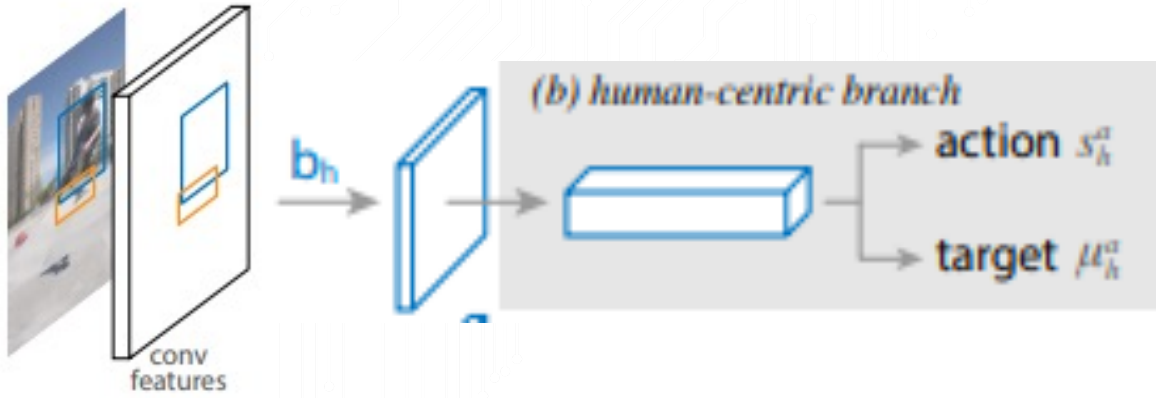
- InteractNet – object detection branch



HOI

Sequential

- InteractNet – human-centric branch
  - Action classification



HOI

Sequential

- InteractNet – human-centric branch
  - Target Localization

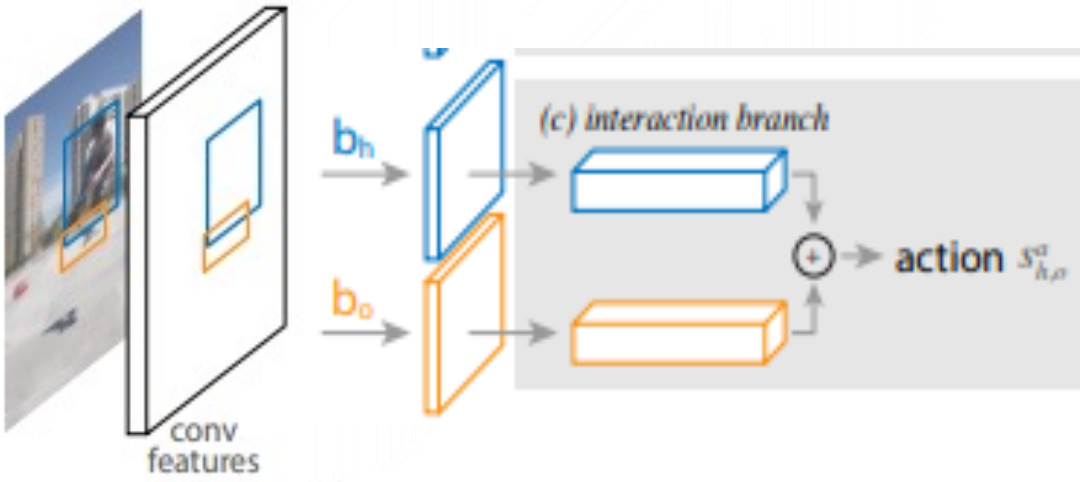
$$g^a_{h,o} = \exp(\|b_{o|h} - \mu^a_h\|^2 / 2\sigma^2)$$

$$b_{o|h} = \{\frac{x_o - x_h}{w_h}, \frac{y_o - y_h}{h_h}, \log\frac{w_o}{w_h}, \log\frac{h_o}{h_h}\}$$

HOI

Sequential

- InteractNet – interaction branch



HOI

Sequential

- InteractNet – triplet score

$$S^a_{h,o} = s_h \cdot s_o \cdot s^a_h \cdot g^a_{h,o}$$

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

HOI

parallel

- PPDM <human point, interaction point, object point>

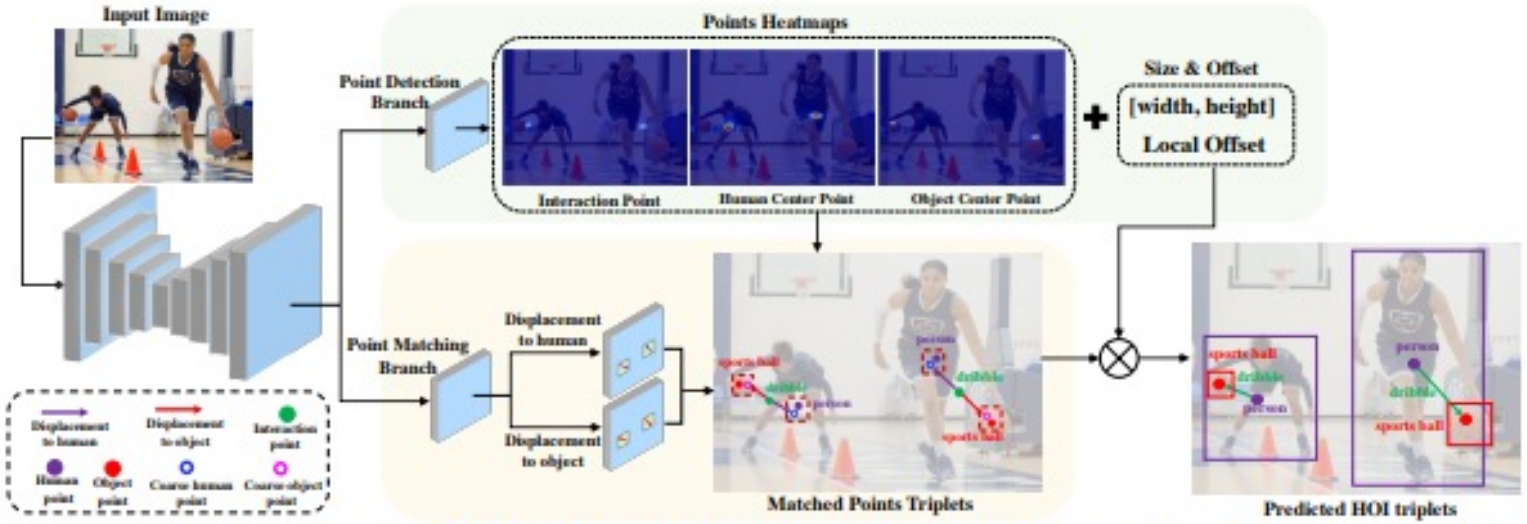
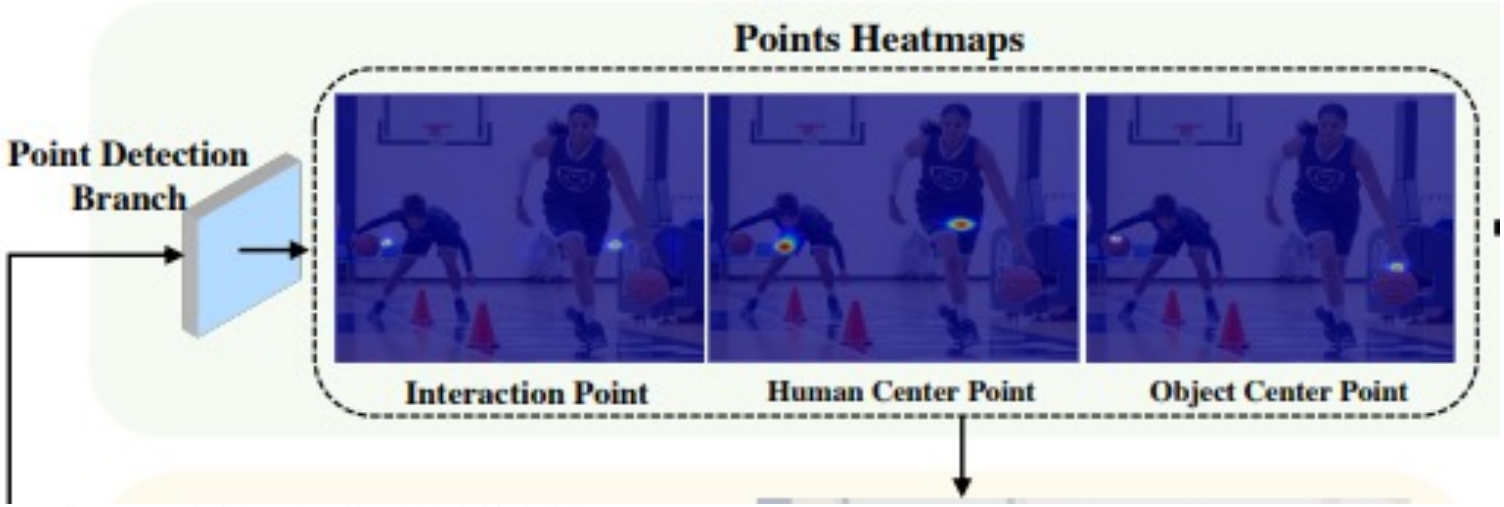


Figure 3. Overview of the proposed PPDM framework. We firstly apply a key-point heatmap prediction network, e.g. Hourglass-104 or DLA-34, to extract the appearance feature from an image. a) Point Detection Branch: Based on the extracted visual feature, we utilize three convolutional modules to predict the heatmap of the interaction points, human center points, and object center points. Additionally, to generate the final box, we regress the 2-D size and the local offset. b) Point Matching Branch: the first step of this branch is to regress the displacements from the interaction point to the human point and object point respectively. Based on the predicted points and displacements, the second step is to match each interaction point with the human point and object point to generate a set of points triplets.

HOI

parallel

- PPDM - Point detection branch



HOI

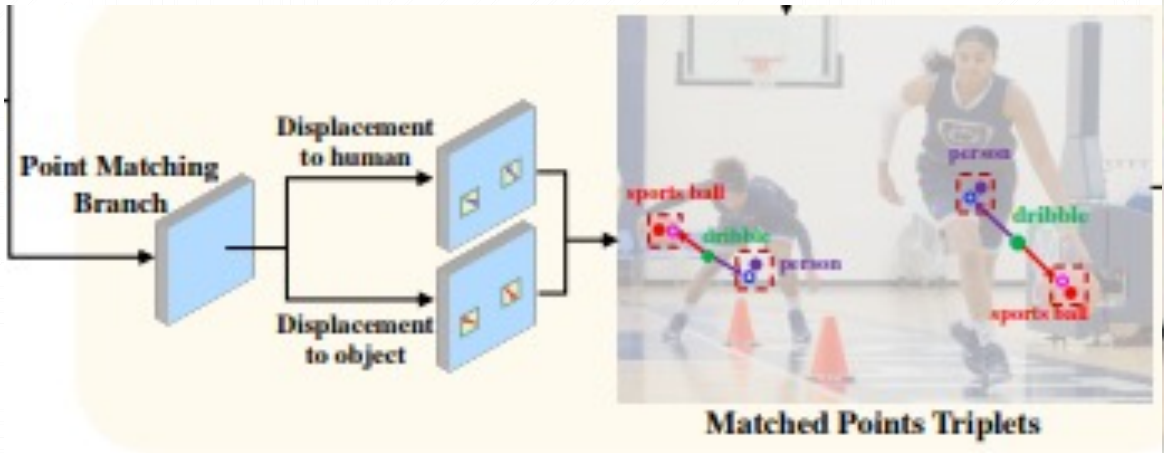
parallel

- PPDM - heatmap

HOI

parallel

- PPDM - Point matching branch



HOI

parallel

- PPDM

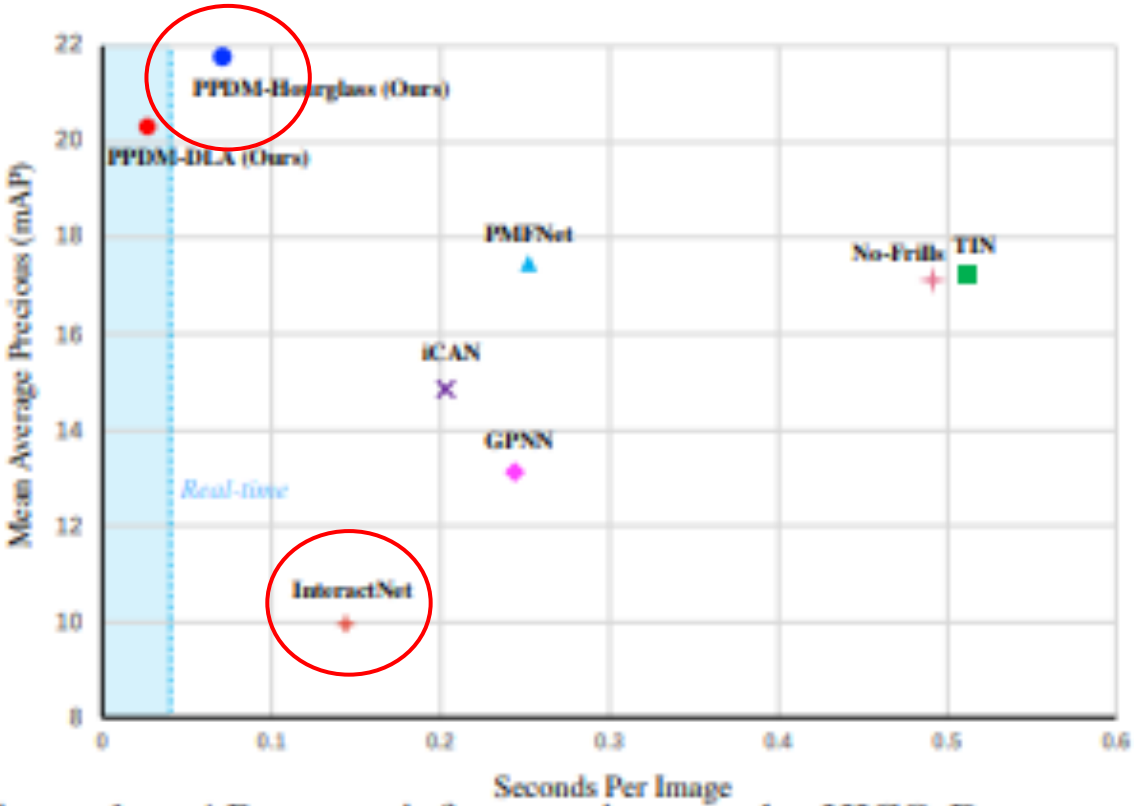


Figure 1. mAP versus inference time on the HICO-Det test set. Our PPDM-DLA outperforms the state-of-the-art methods with the inference speed of 37 fps (0.027s). It is the first real-time HOI detection method. Our PPDM-Hourglass achieves 4.27% mAP improvement over the state-of-the-arts with a faster speed.

HOI

transformer

- HOTR

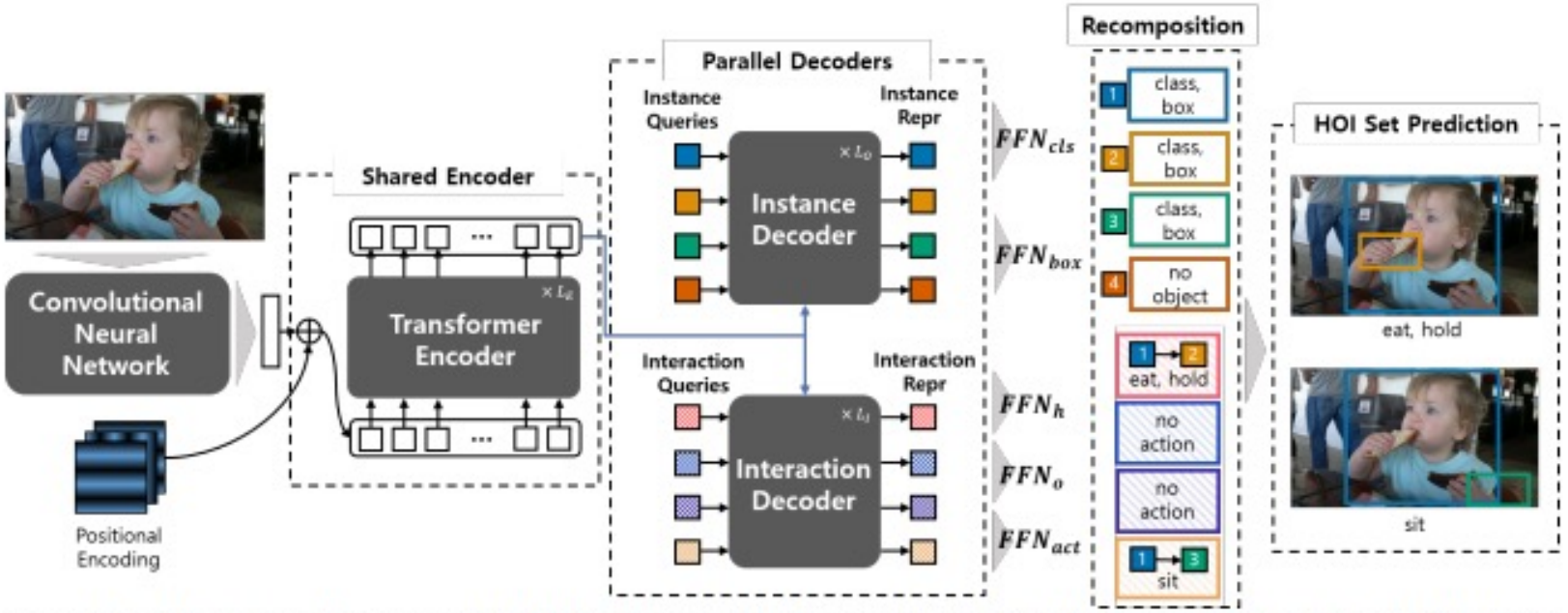
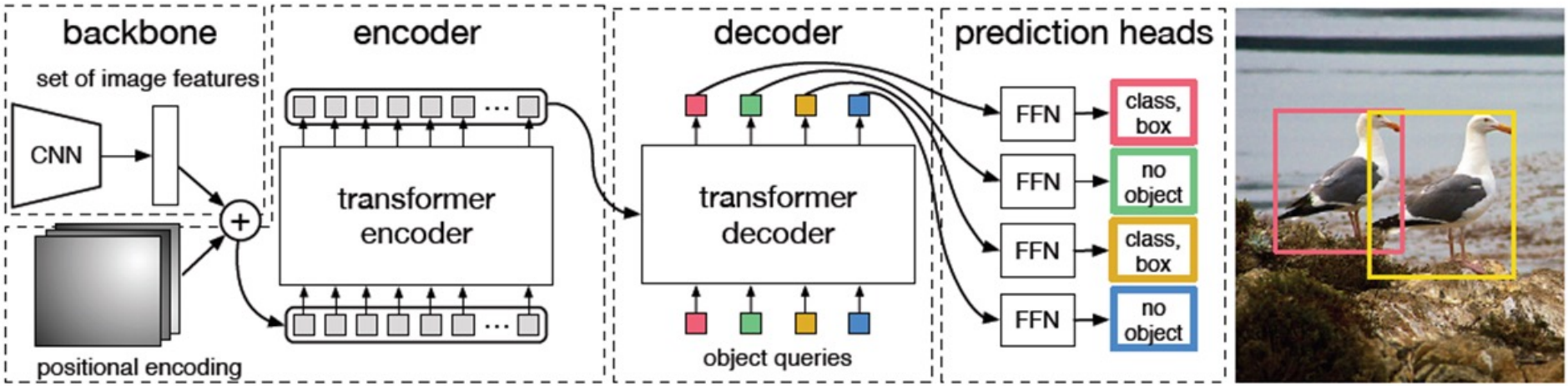


Figure 2. Overall pipeline of our proposed model. The Instance Decoder and Interaction Decoder run in parallel, and share the Encoder. In our recomposition, the interaction representations predicted by the Interaction Decoder are associated with the instance representations to predict a fixed set of HOI triplets (see Fig.3). The positional encoding is identical to [2].

HOI

transformer

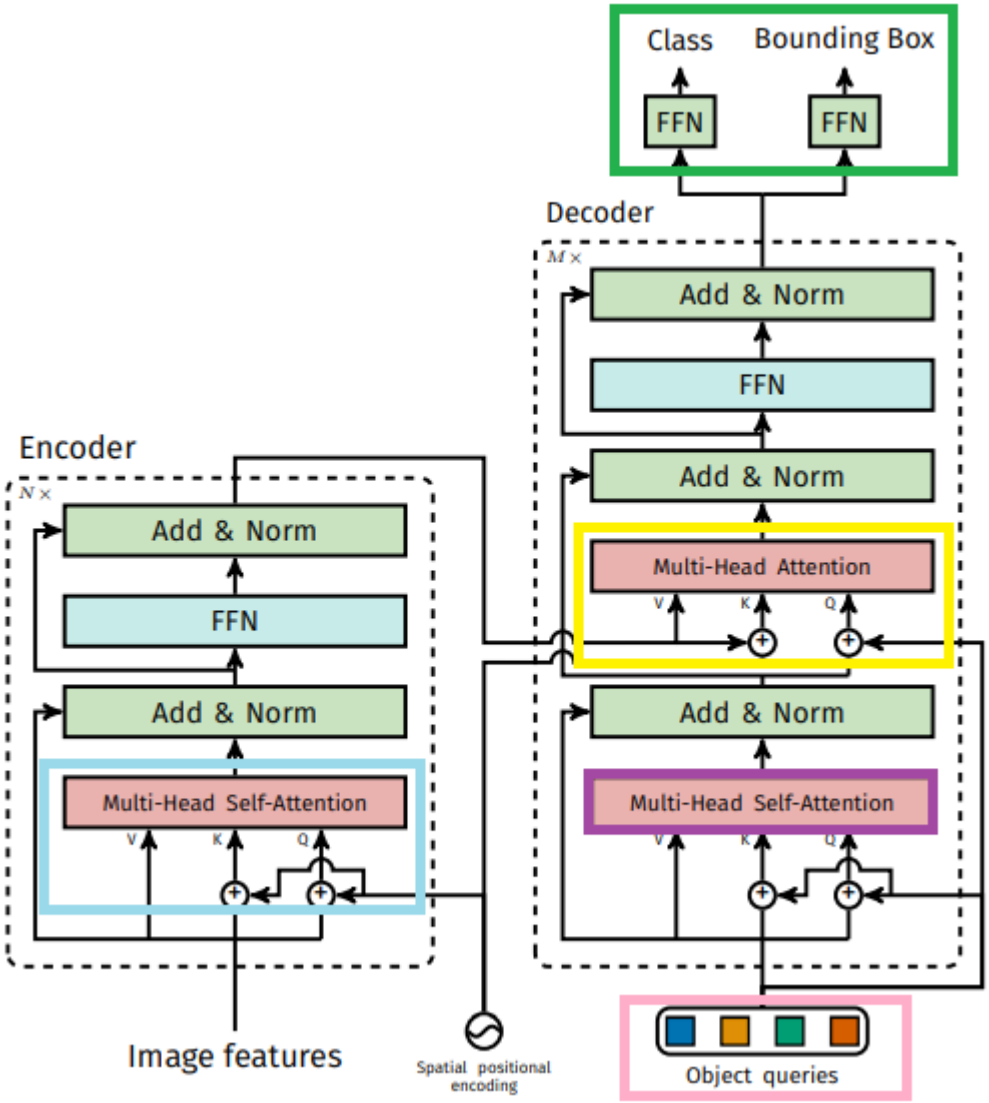
- HOTR
  - DETR



HOI

transformer

- HOTR
  - DETR



HOI

transformer

- HOTR
  - DETR
    - Positional encoding - sinusoidal

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

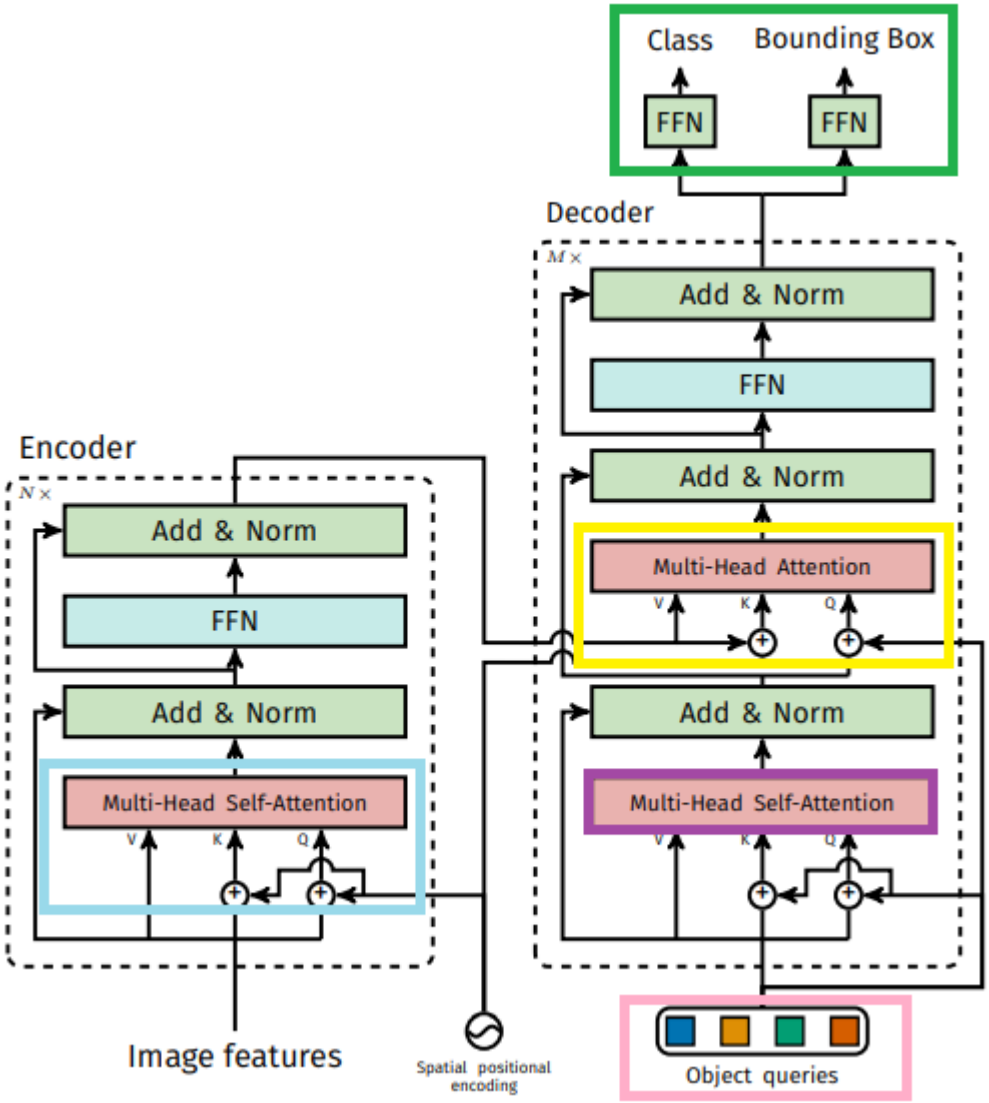
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

	Position	1	2	3	.....
Position Encoding		$\sin(1)$	$\sin(2)$	$\sin(3)$	
		$\cos(1)$	$\cos(2)$	$\cos(3)$	
		$\sin(1/10000^{2/d_{model}})$	$\sin(2/10000^{2/d_{model}})$	$\sin(3/10000^{2/d_{model}})$	.....
		$\cos(1/10000^{2/d_{model}})$	$\cos(2/10000^{2/d_{model}})$	$\cos(3/10000^{2/d_{model}})$	
		$\sin(1/10000^{4/d_{model}})$	$\sin(2/10000^{4/d_{model}})$	$\sin(3/10000^{4/d_{model}})$	
		⋮			⋱

HOI

transformer

- HOTR
  - DETR



HOI

transformer

- HOTR

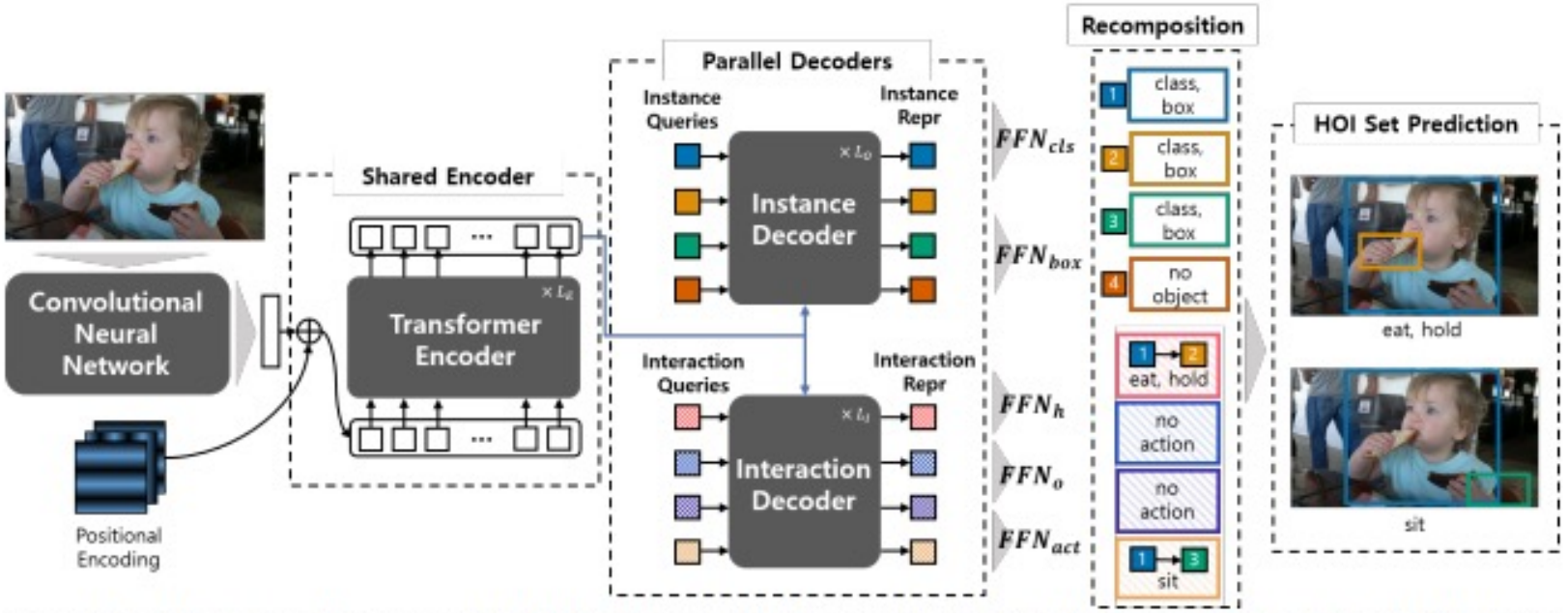


Figure 2. Overall pipeline of our proposed model. The Instance Decoder and Interaction Decoder run in parallel, and share the Encoder. In our recomposition, the interaction representations predicted by the Interaction Decoder are associated with the instance representations to predict a fixed set of HOI triplets (see Fig.3). The positional encoding is identical to [2].

HOI

transformer

- HOTR

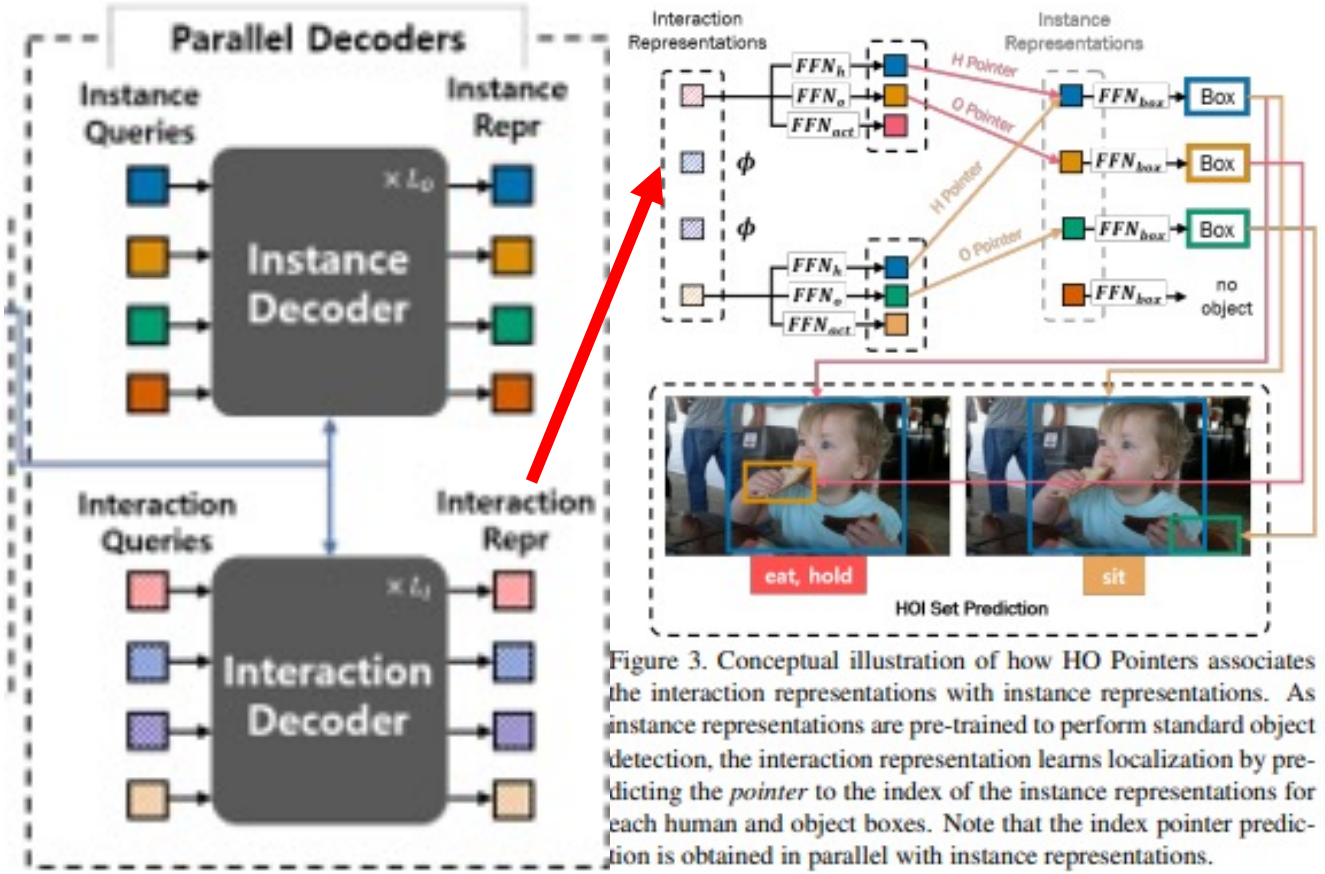


Figure 3. Conceptual illustration of how HO Pointers associates the interaction representations with instance representations. As instance representations are pre-trained to perform standard object detection, the interaction representation learns localization by predicting the *pointer* to the index of the instance representations for each human and object boxes. Note that the index pointer prediction is obtained in parallel with instance representations.

HOI

transformer

- HOTR

Parallel

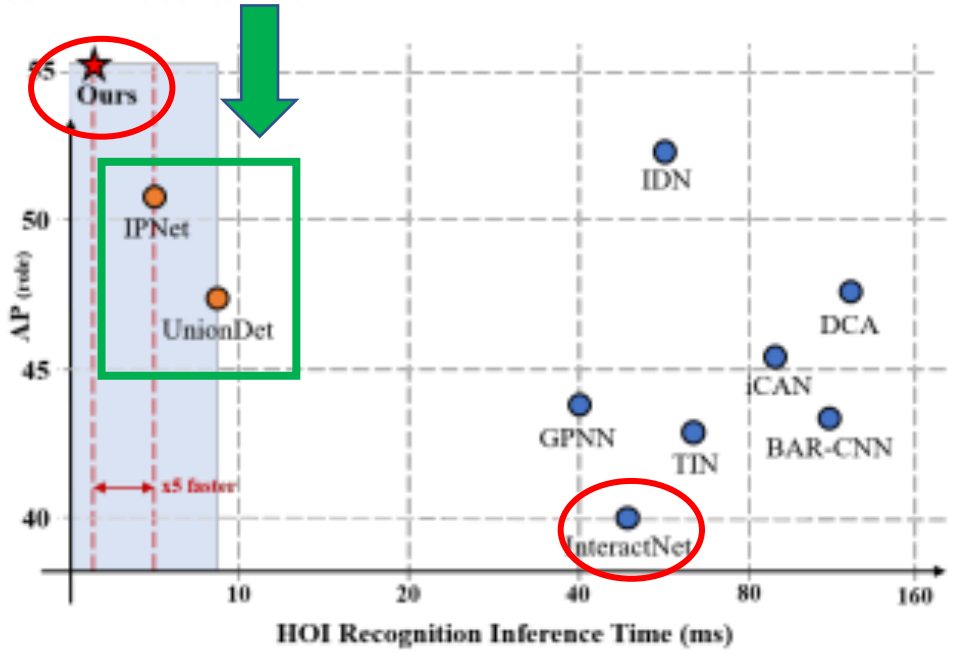


Figure 1. Time vs. Performance analysis for HOI detectors on V-COCO dataset. HOI recognition inference time is measured by subtracting the object detection time from the end-to-end inference time. Blue circle represents sequential HOI detectors, orange circle represents parallel HOI detectors and red star represents ours. Our method achieves an HOI recognition inference time of 0.9ms, being significantly faster than the parallel HOI detectors such as IPNet [30] or UnionDet [12] (the comparison between parallel HOI detectors is highlighted in blue).

참고문헌

G. Gkioxari, R. Girshick, P. Dollár and K. He, "Detecting and Recognizing Human-Object Interactions," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8359-8367, doi: 10.1109/CVPR.2018.00872.

http://chengao.vision/iCAN/

Gao, Chen, Yuliang Zou, and Jia-Bin Huang. "ican: Instance-centric attention network for human-object interaction detection." *arXiv preprint arXiv:1808.10437* (2018).

Y. Liao, et al., "PPDM: Parallel Point Detection and Matching for Real-Time Human-Object Interaction Detection," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020 pp. 479-487.

Kim, Bumsoo, et al. "Hotr: End-to-end human-object interaction detection with transformers." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.