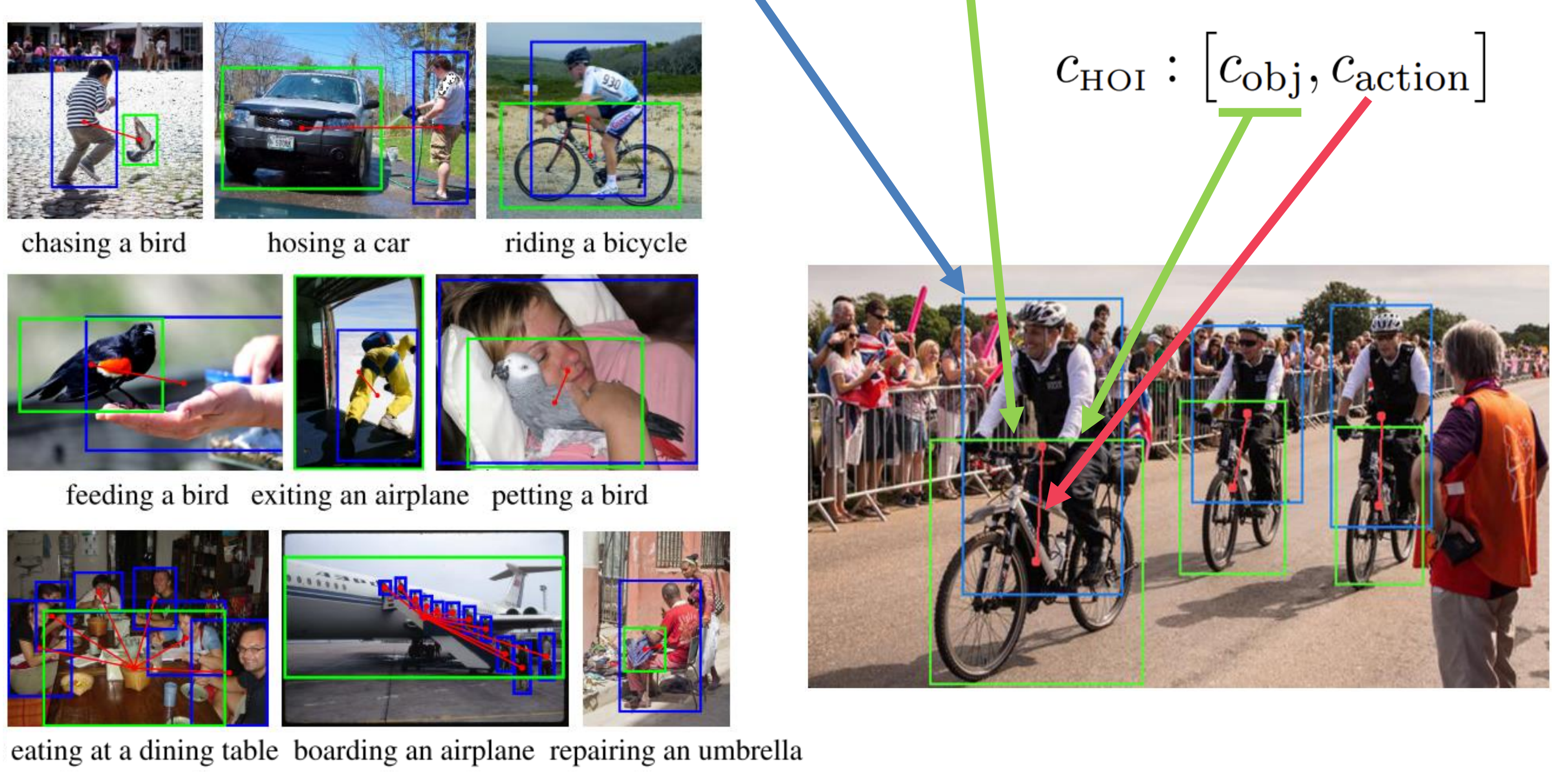
Parallel Queries for Human-Object Interaction Detection

Junwen Chen, Keiji Yanai
Department of Informatics
The University of Electro-Communications
Tokyo, Japan

HOI Detection

$$\left\{\left[x_1^{\text{human}}, y_1^{\text{human}}, x_2^{\text{human}}, y_2^{\text{human}}\right], \left[x_1^{\text{obj}}, y_1^{\text{obj}}, x_2^{\text{obj}}, y_2^{\text{obj}}\right], c_{\text{HOI}}\right\}$$

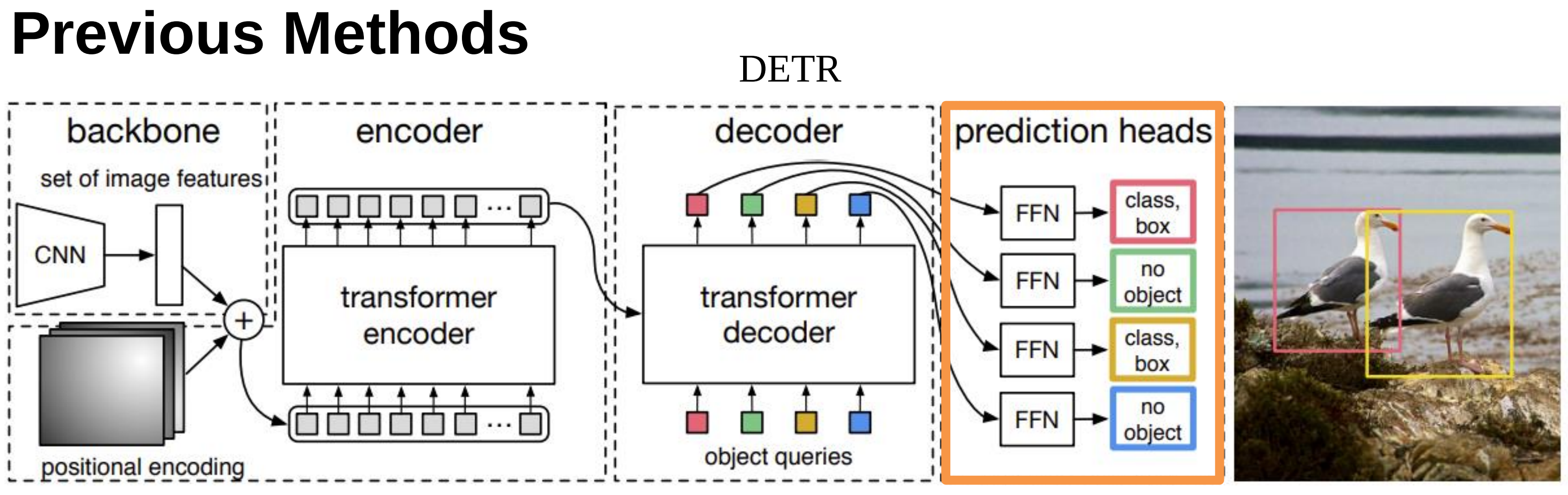
$$c_{\text{HOI}} : \left[c_{\text{obj}}, c_{\text{action}}\right]$$

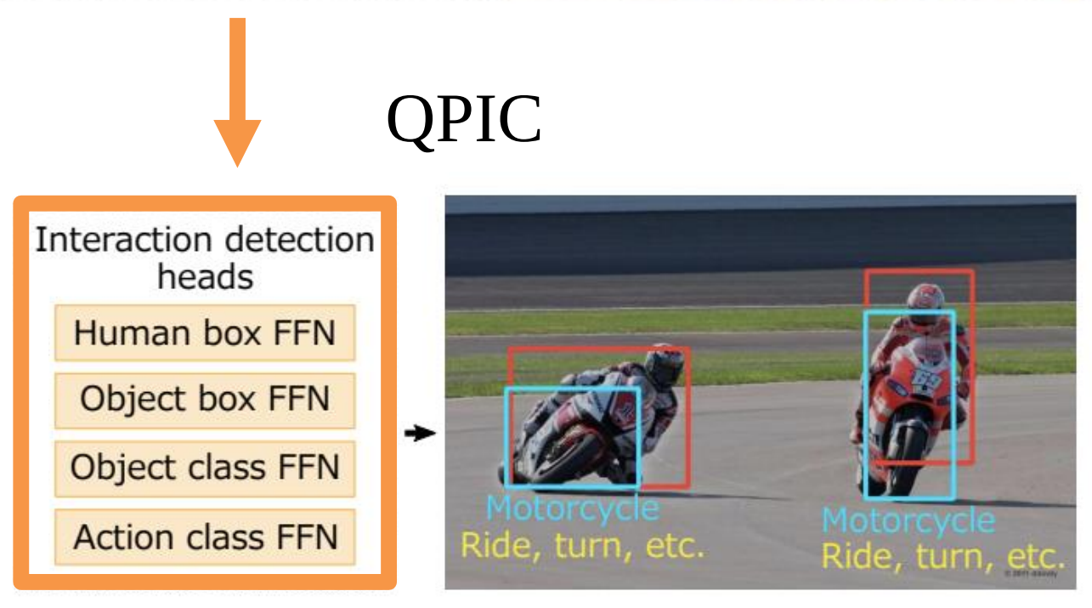


- Human-object interaction (HOI) detection has recently received increasing attention as a field with great potential applications
- HICO-DET is the most widely used dataset for HOI detection
  - Training set: 38,118 images, Test set: 9,658 images
  - HOI class: 600 classes consisting of 117 verbs and 80 objects

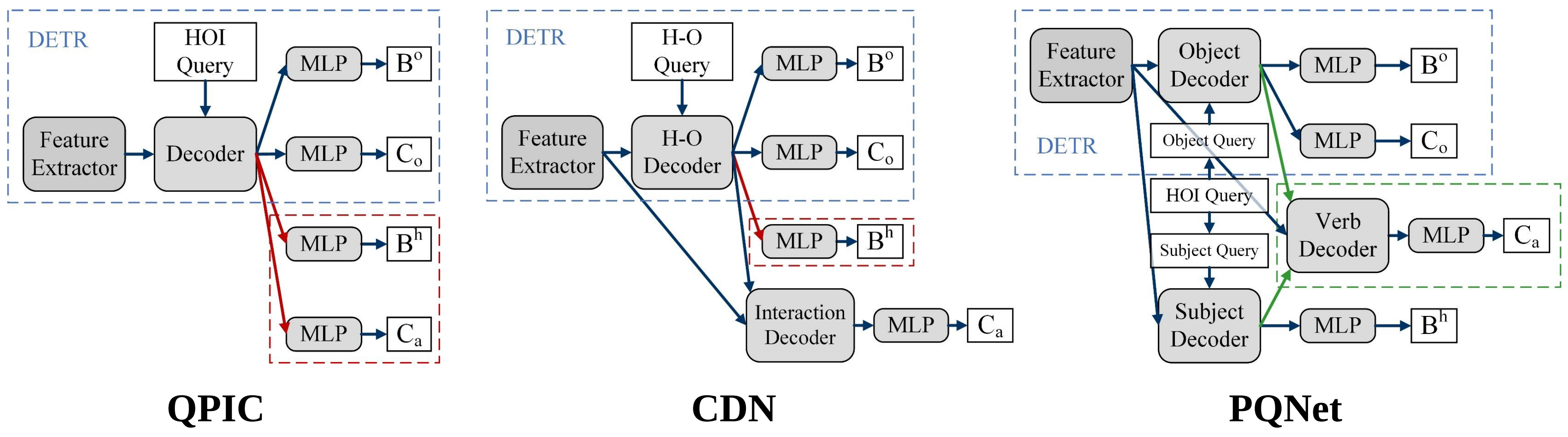
Motivation

□ Previous Methods



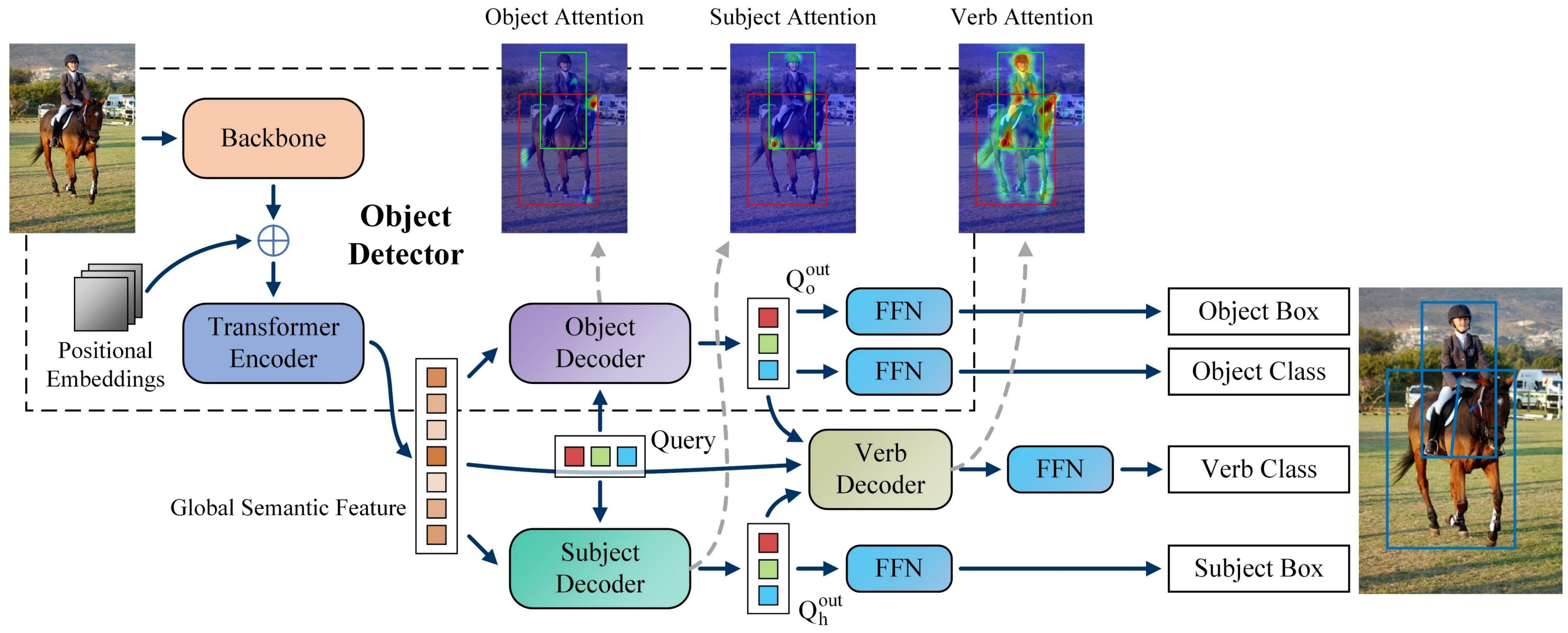


- Transformer-based object detection method, DETR, opens up a new path for object detection by viewing the object detection as a set prediction problem
- QPIC replaces the detection head of DETR with an interaction head



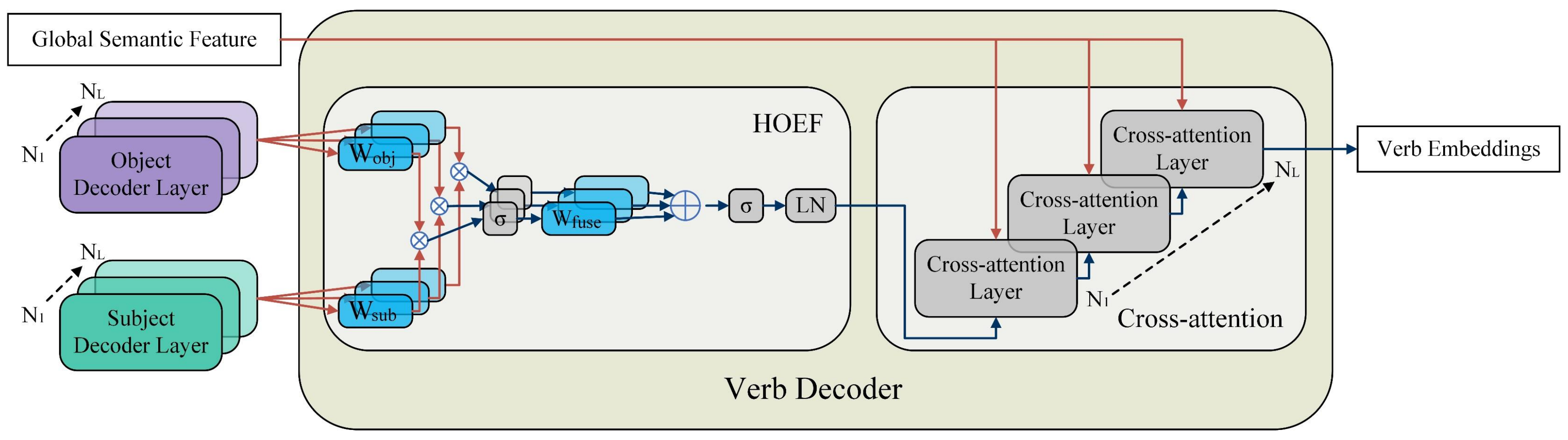
- Previous methods modify the decoding part or change the decoding target of DETR
- The training of these models needs to adapt the object detection part to a new target which leads to slower convergence

PQNet



- PQNet comprises four parts, the object detector, subject decoder, verb decoder, and feed-forward network (FFN) heads
- The object decoder and subject decoder uses the global semantic feature to predict the object boxes and human boxes with the FFN head, respectively
- The verb decoder fuses the object and subject embeddings and extracts the verb representations from the global semantic feature

Verb Decoder

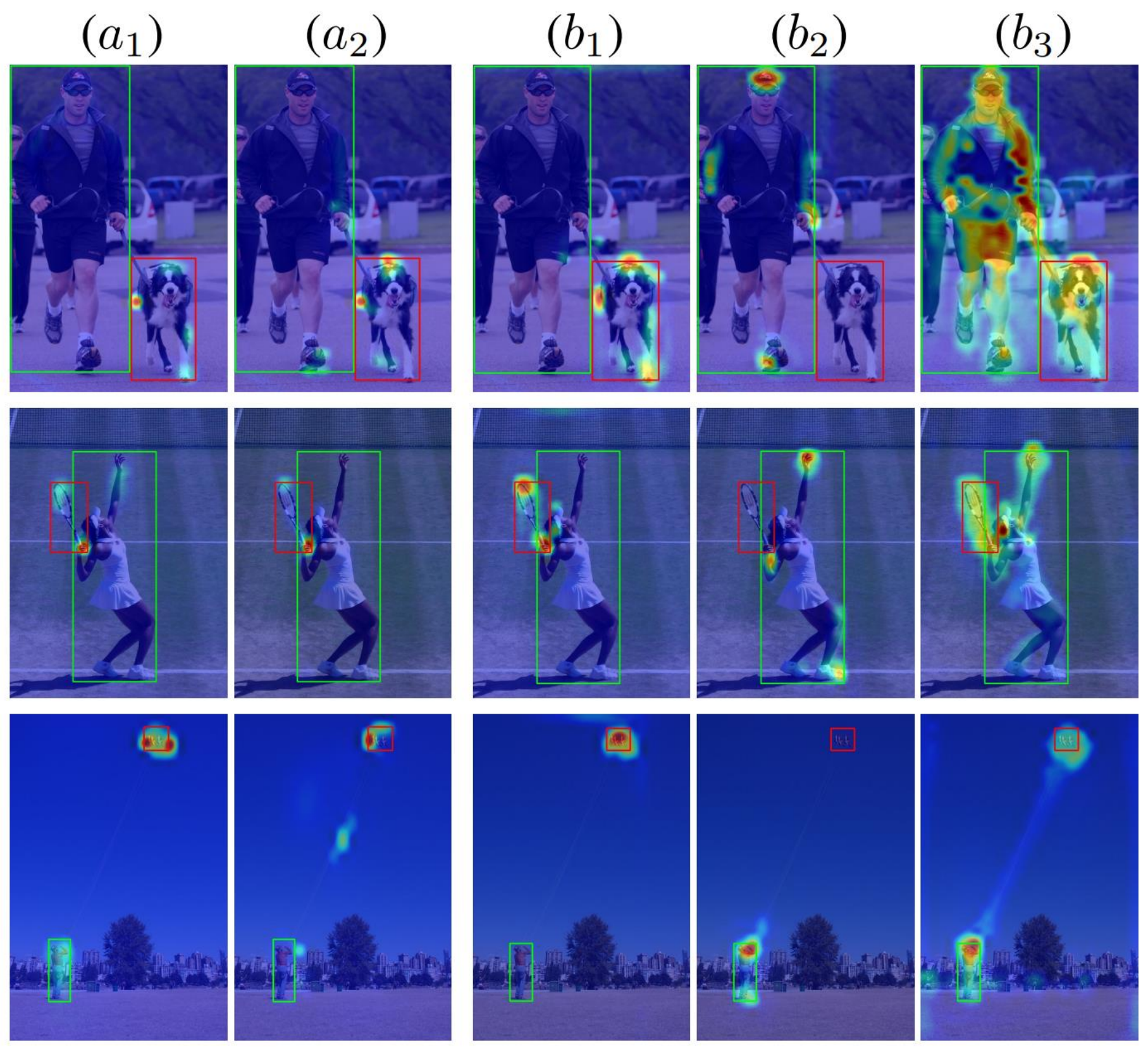


- The HOEF module fuses the last layer's embeddings from the object decoder and subject decoder
- The cross-attention module calculates the attention between the global semantic feature and the fused embeddings layer by layer
- The verb embeddings from the last layer of cross-attention are used to predict the verb classes of HOI instances

Results

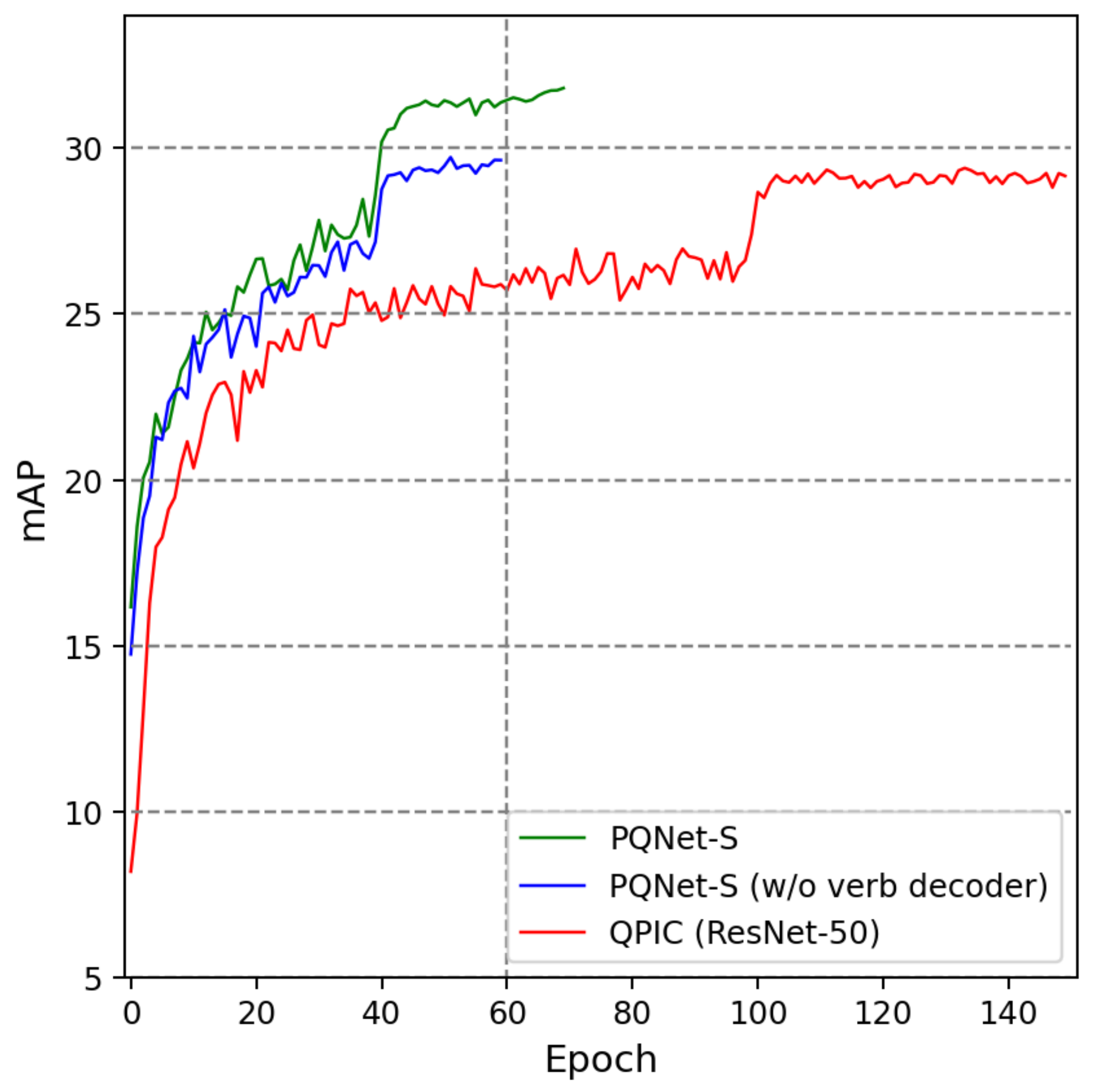
Method	Fine-tuned Detector	Backbone	Feature	Default			Known Object		
				Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
Two-stage									
No-Frills [8]	✗	ResNet-152	A+S+P	17.18	12.17	18.68	-	-	-
RPNN [32]	✗	ResNet-50	A+P	17.35	12.78	18.71	-	-	-
PMFNet [26]	✗	ResNet-50-FPN	A+S+P	17.46	15.65	18.00	20.34	17.47	21.20
VSGNet [25]	✗	ResNet-152	A+S	19.80	16.05	20.91	-	-	-
FCMNet [18]	✗	ResNet-50	A+S+L	20.41	17.34	21.56	22.04	18.97	23.12
ACP [13]	✗	ResNet-152	A+P+L	20.59	15.92	21.98	-	-	-
DJ-RN [15]	✗	ResNet-50	A+S+P	21.34	18.53	22.18	23.69	20.64	24.60
PD-Net [30]	✗	ResNet-152	A+S+P+L	22.37	17.61	23.79	26.86	21.70	28.44
DRG [5]	✓	ResNet-50-FPN	A+S+L	24.53	19.47	26.04	27.98	23.11	29.43
SCG [29]	✓	ResNet-50-FPN	A+S	31.33	24.72	33.31	34.37	27.18	36.52
One-stage									
PPDM [16]	✓	Hourglass-104	A	21.73	13.78	24.10	24.58	16.65	26.84
GGNet [31]	✓	Hourglass-104	A	23.47	16.48	25.60	27.36	20.23	29.48
HOITrans [34]	✓	ResNet-101	A	26.61	19.15	28.84	29.13	20.98	31.57
HOTR [12]	✓	ResNet-50	A	25.10	17.34	27.42	-	-	-
AS-Net [4]	✓	ResNet-50	A	28.87	24.25	30.25	31.74	27.07	33.14
QPIC [24]	✓	ResNet-50	A	29.07	21.85	31.23	31.68	24.14	33.93
QPIC [24]	✓	ResNet-101	A	29.90	23.92	31.69	32.38	26.06	34.27
CDN-S [28]	✓	ResNet-50	A	31.44	27.39	32.64	34.09	29.63	35.42
CDN-B [28]	✓	ResNet-50	A	31.78	27.55	33.05	34.53	29.73	35.96
CDN-L [28]	✓	ResNet-101	A	32.07	27.19	33.53	34.79	29.48	36.38
PQNet-S	✓	ResNet-50	A	31.92	28.06	33.08	34.58	30.71	35.74
PQNet-B	✓	ResNet-50	A	32.13	**29.43**	32.93	34.68	**32.06**	35.47
PQNet-L	✓	ResNet-101	A	**32.45**	27.80	**33.84**	**35.28**	30.72	**36.64**

- Comparing to recent one-stage methods, PQNet-B exceeds QPIC (ResNet-50) and CDN-B by 3.06 mAP (relatively 10.5%) and 0.35 mAP (relatively 1.1%), respectively
- PQNet-L achieves 32.45 mAP, 0.38 mAP (relatively 1.2%) higher than CDN-L



- Attention visualization of the decoders' last layer
- (a1) and (a2) are from CDN's instance decoder and interaction decoder
- (b1), (b2), and (b3) are from PQNet's object decoder, subject decoder, and verb decoder

- PQNet learned to focus on the extreme points of the target
- The verb decoder focuses on the whole part of the human and object but pays more attention to the interaction regions



- The training process of PQNet and QPIC
- PQNet-S achieves more than twice mAP at the first epoch and shows a fast convergence in the first 40 epochs before the learning rate drops