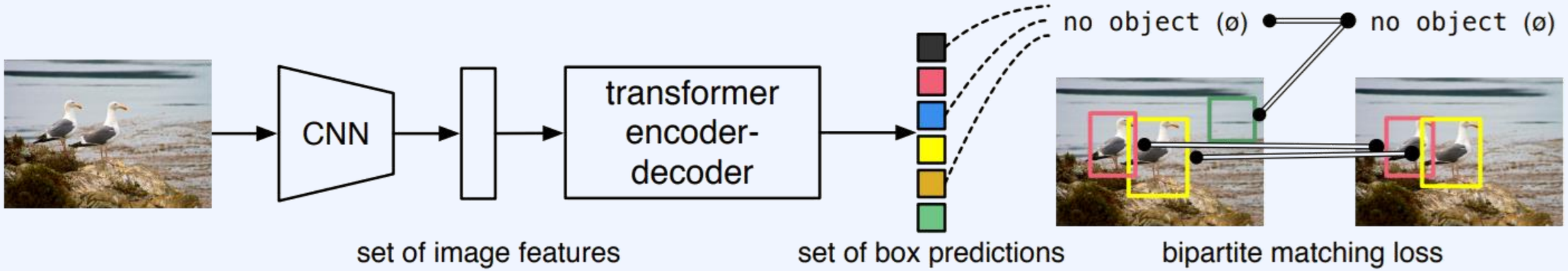
Context Understanding

4 Transformer for OD and HOI

Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

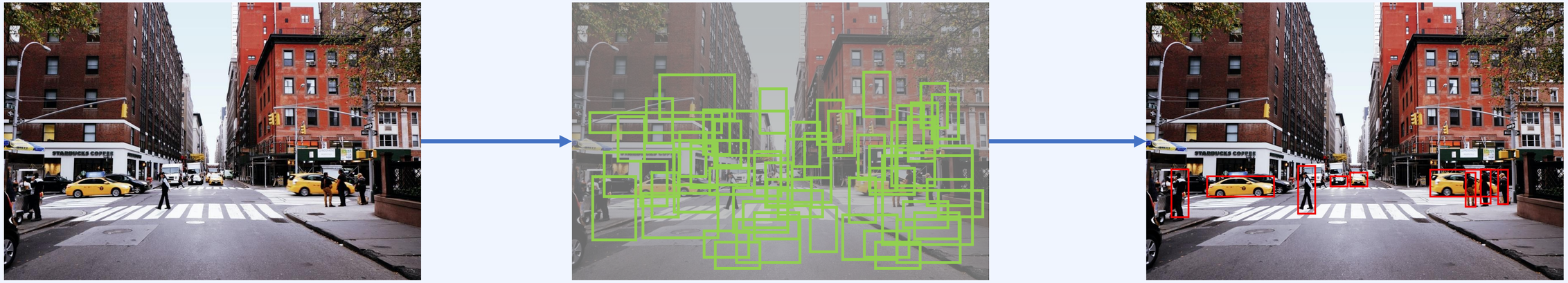
DETR



Object Detection

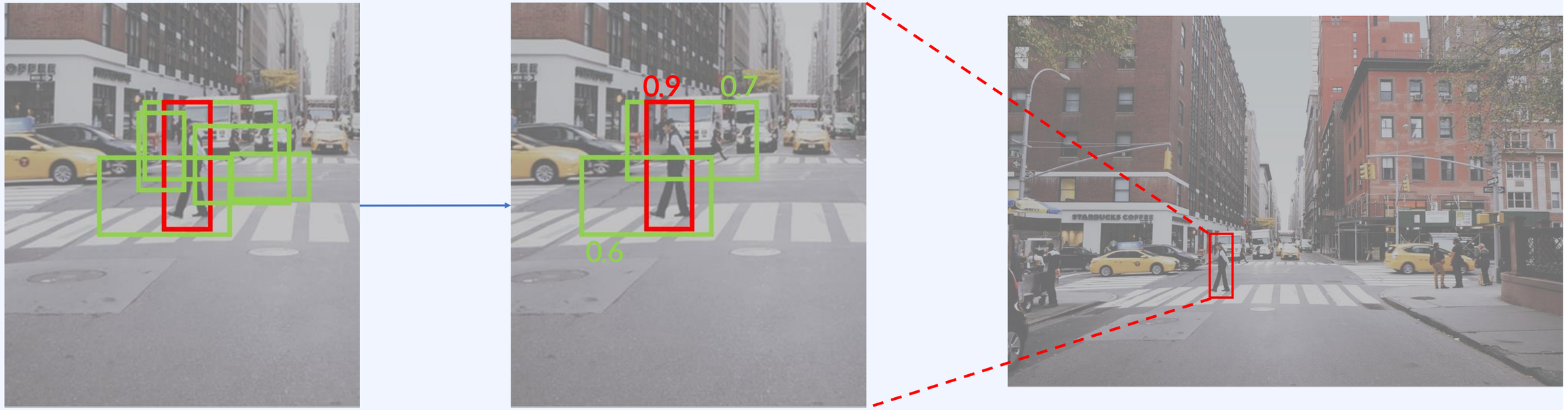
Object Detection

- Non-maximum suppression (NMS)



Object Detection

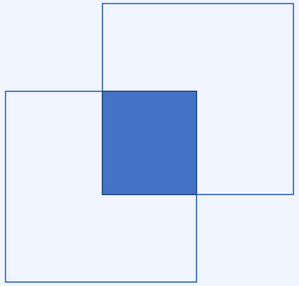
- Non-maximum suppression (NMS)



Object Detection

- Intersection over Union (IoU)

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

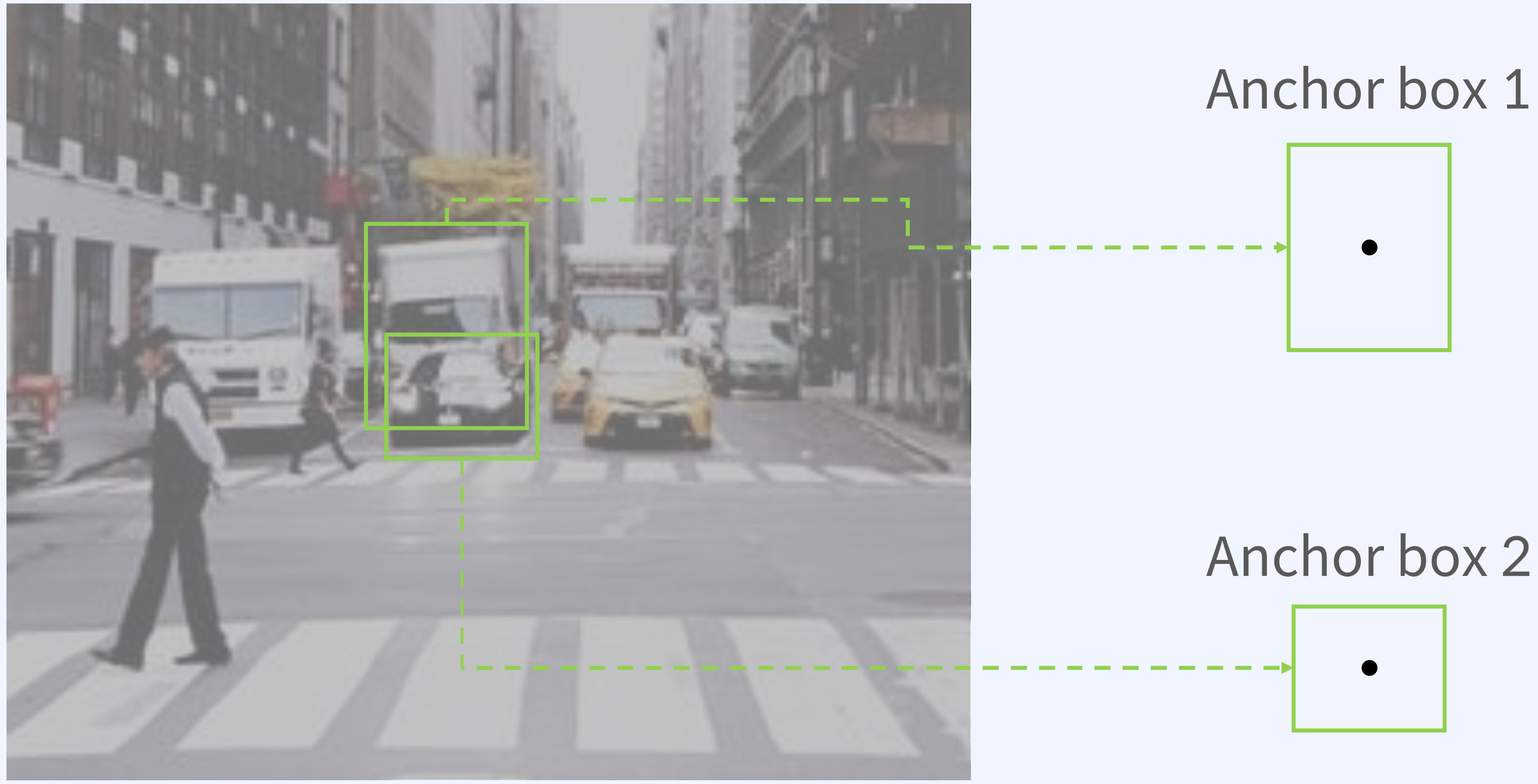


Object Detection

- Anchor box

Object Detection

- Anchor box



Context Understanding
Transformers for OD and HOI

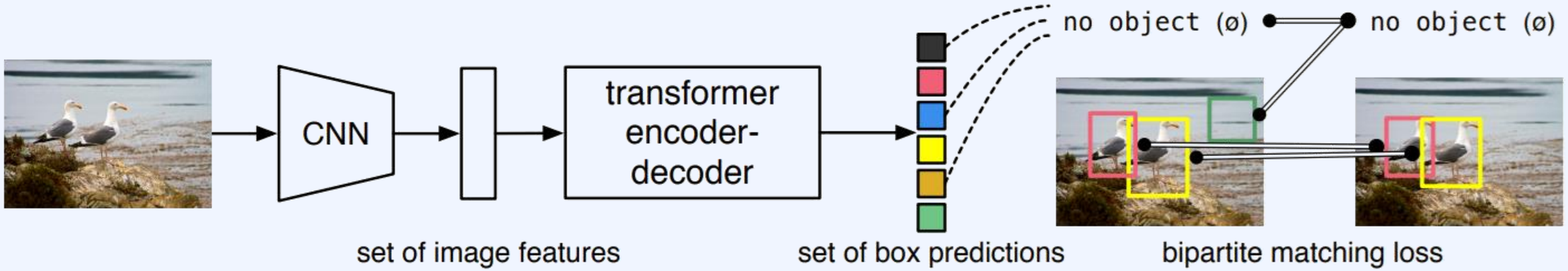
N. Carion et al. End-to-end object detection with transformers. ECCV

DETR

Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

DETR



Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

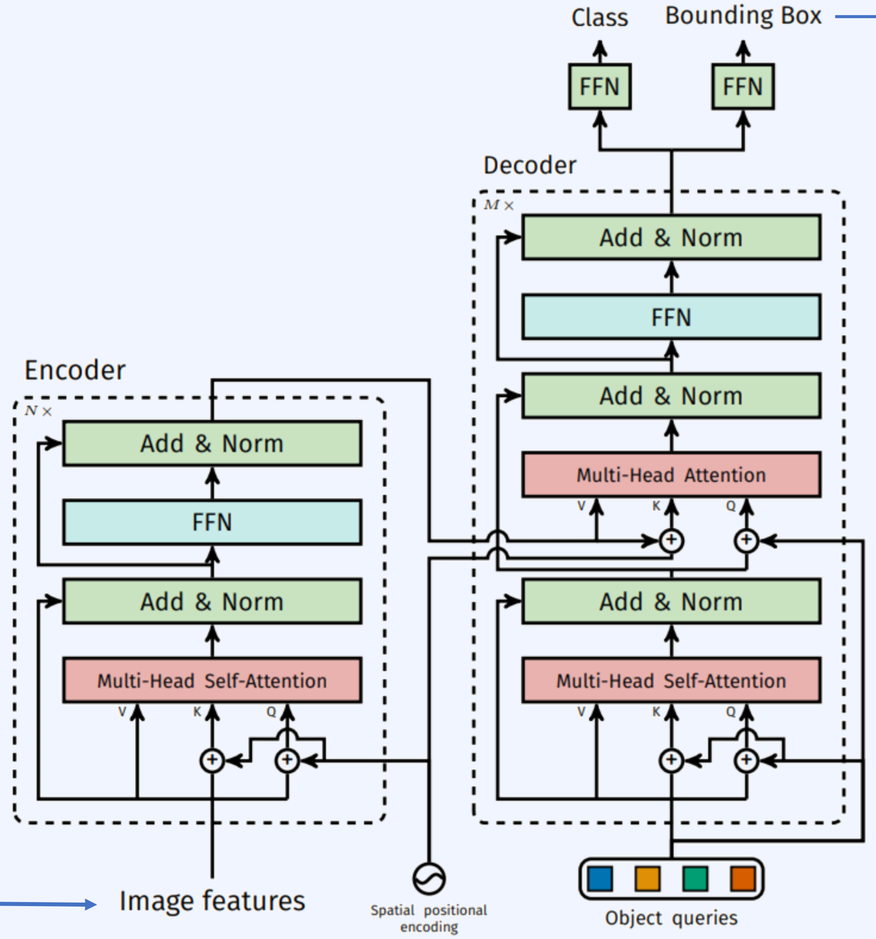
DETR

$$x_{\text{img}} \in \mathbb{R}^{3 \times H_0 \times W_0}$$

$$f \in \mathbb{R}^{C \times H \times W}$$

1x1 conv

$$z_0 \in \mathbb{R}^{d \times H \times W}$$



$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

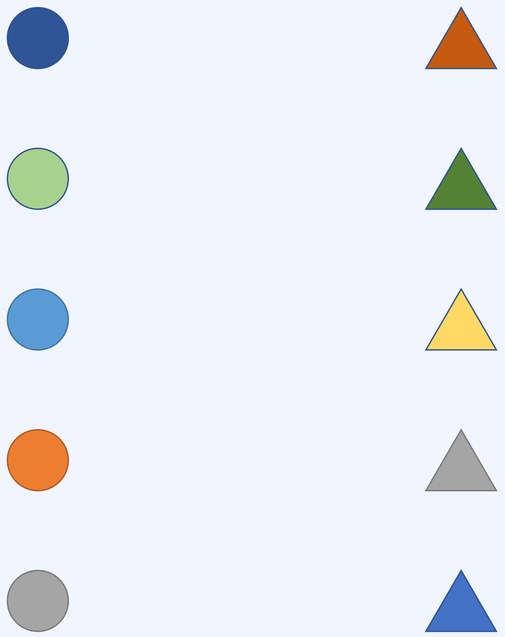
Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

DETR

$\hat{y} = \{\hat{y}_i\}_{i=1}^{N}$ Prediction

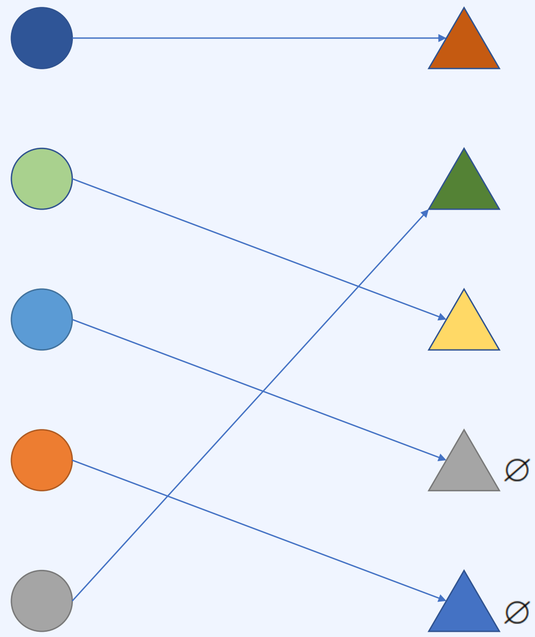
y Ground Truth



$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

Hungarian Algorithm

Prediction Ground Truth



Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

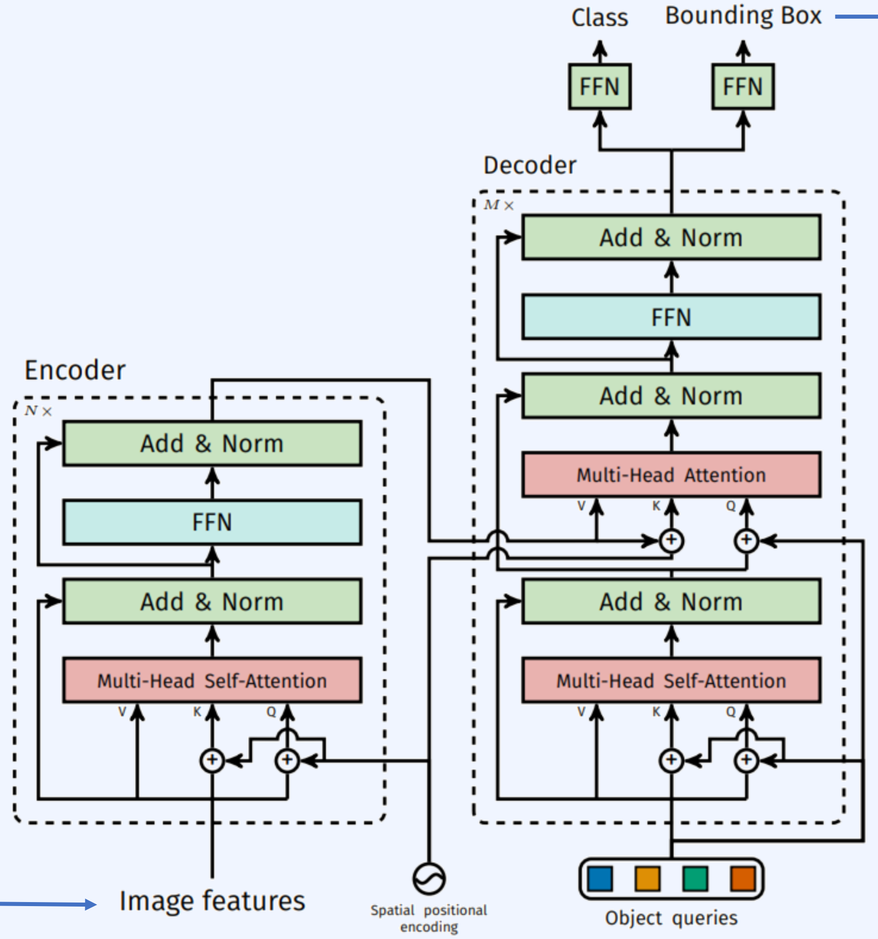
DETR

$$x_{\text{img}} \in \mathbb{R}^{3 \times H_0 \times W_0}$$

$$f \in \mathbb{R}^{C \times H \times W}$$

1x1 conv

$$z_0 \in \mathbb{R}^{d \times H \times W}$$



$$\hat{\sigma} = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$$

$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]$$

$$\lambda_{\text{iou}} \mathcal{L}_{\text{iou}}(b_i, \hat{b}_{\sigma(i)}) + \lambda_{\text{L1}} ||b_i - \hat{b}_{\sigma(i)}||_1$$

Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

DETR

Backbone : ResNet (Torchvision, pretrained over ImageNet, lr = 0.00001, frozen BN weight)

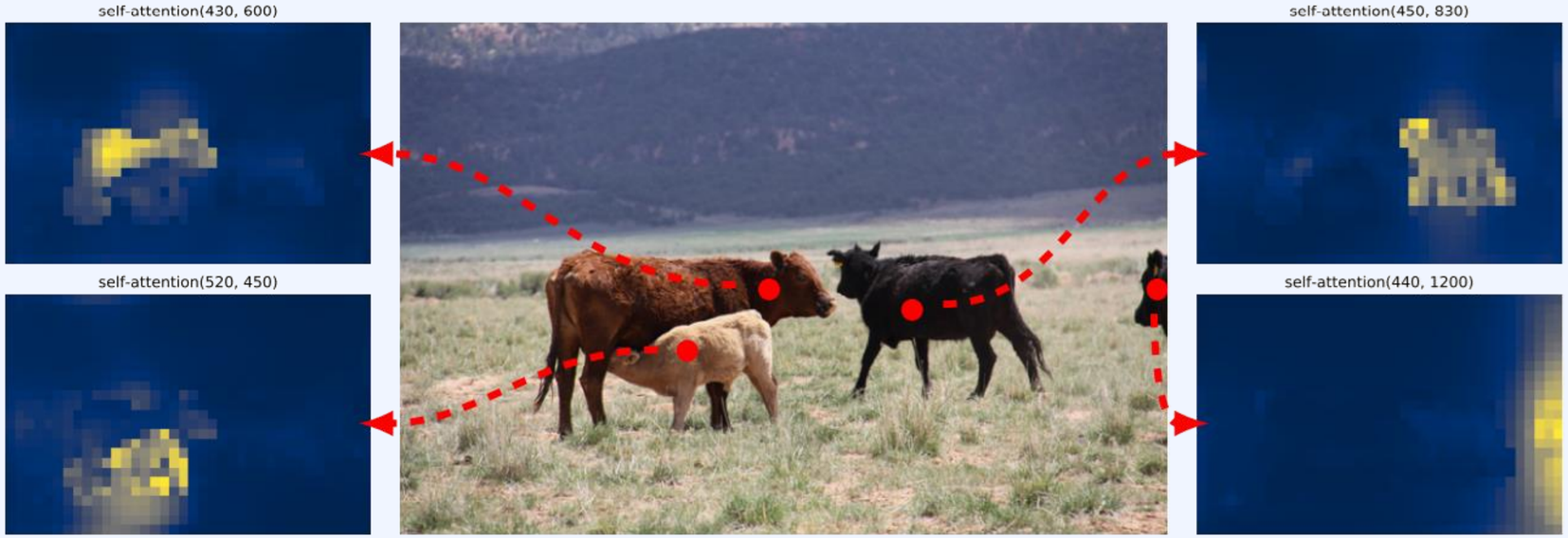
Transformer : lr = 0.0001, dropout = 0.1 after MHA and FFN, Xavier initialization

Optimizer : AdamW with improved weight decay 0.0001 , gradient clipping with a max grad norm 0.1

Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

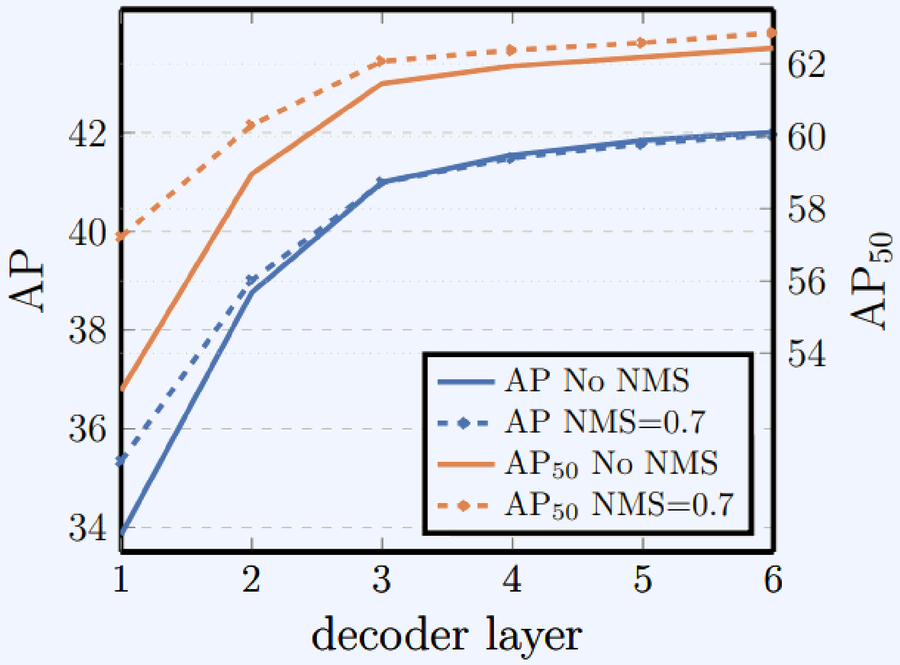
DETR



Context Understanding
Transformers for OD and HOI

N. Carion et al. End-to-end object detection with transformers. ECCV

DETR

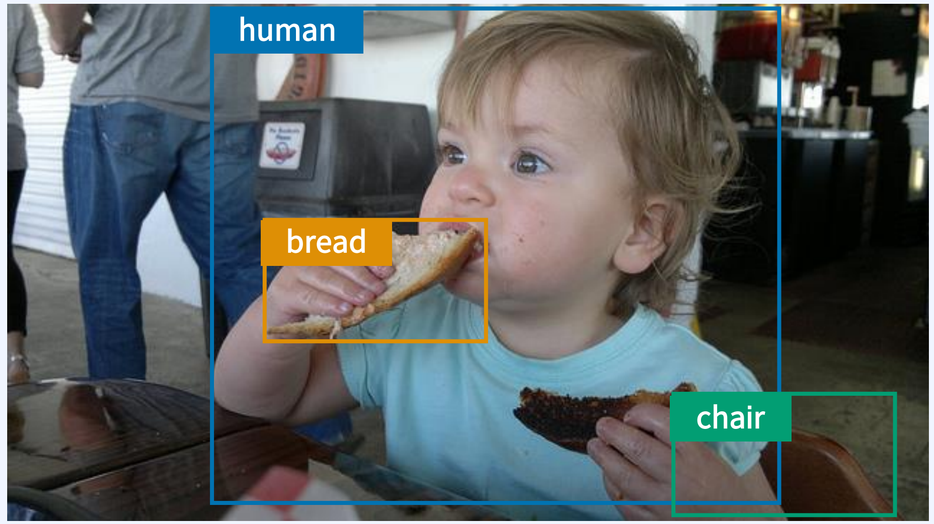


Human-Object Interaction (HOI) Detection

◦ Set $\{(\text{bbox}_1^h, \text{bbox}_1^o, [\text{eat, hold}]), (\text{bbox}_2^h, \text{bbox}_2^o, [\text{sit}])\}$

Human-Object Interaction (HOI) Detection

◦ Set {(bbox_1^h, bbox_1^o,), (bbox_2^h, bbox_2^o, }



Human-Object Interaction (HOI) Detection

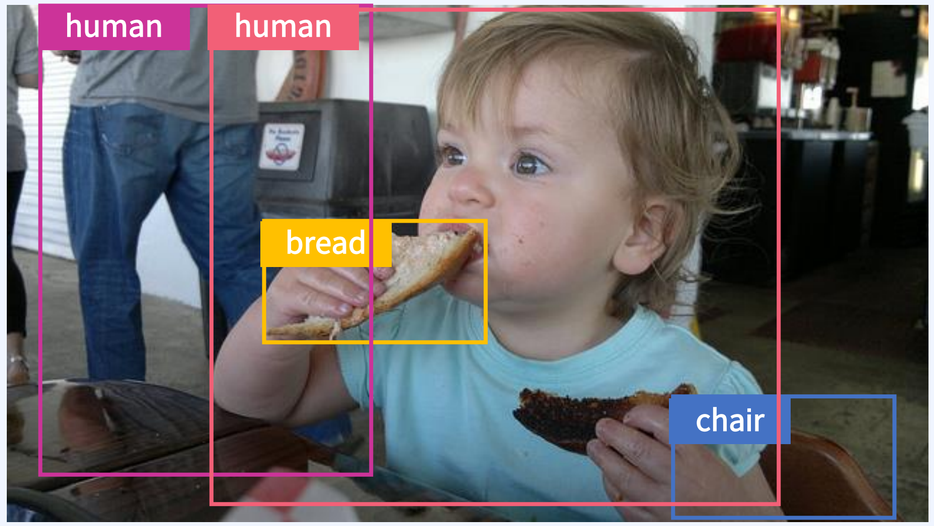
◦ Set {(bbox_1^h, bbox_1^o,[eat, hold]), (bbox_2^h, bbox_2^o, }

Human-Object Interaction (HOI) Detection

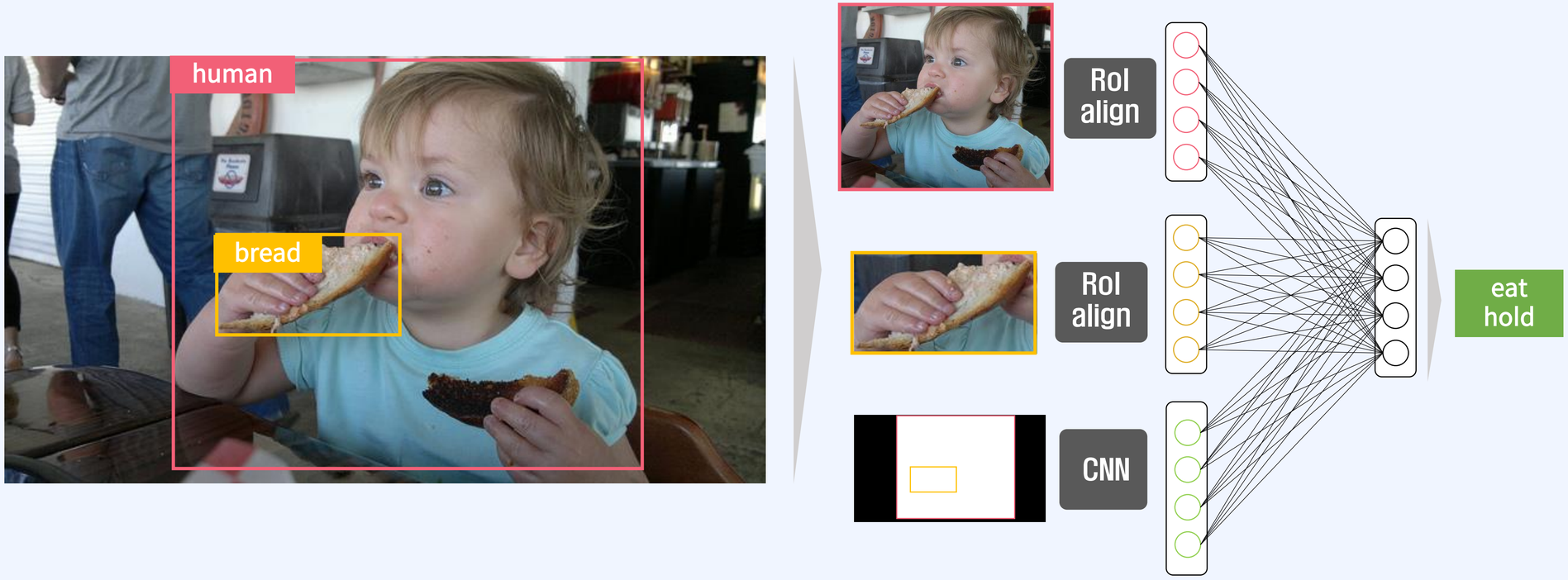
◦ Set $\{(\text{bbox}_1^h, \text{bbox}_1^o, [\text{eat, hold}]), (\text{bbox}_2^h, \text{bbox}_2^o, [\text{sit}])\}$

◦ Sequential HOI Detectors

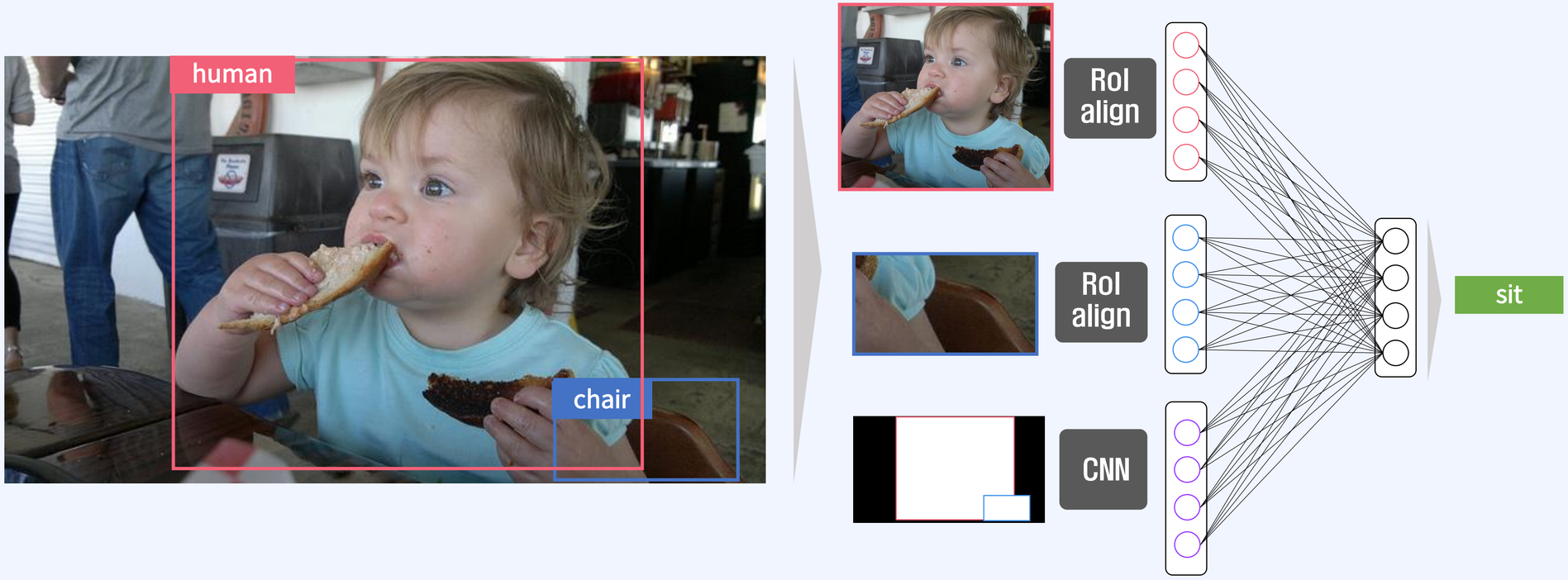
◦ Sequential HOI Detectors



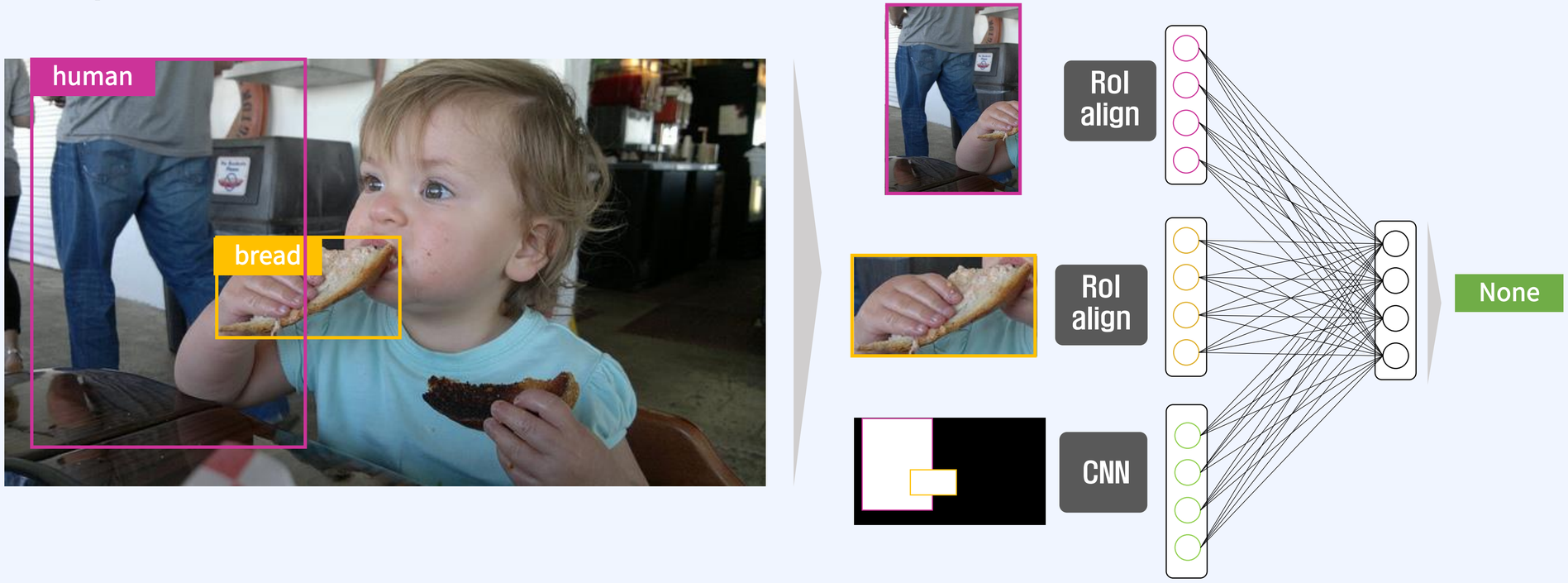
◦ Sequential HOI Detectors



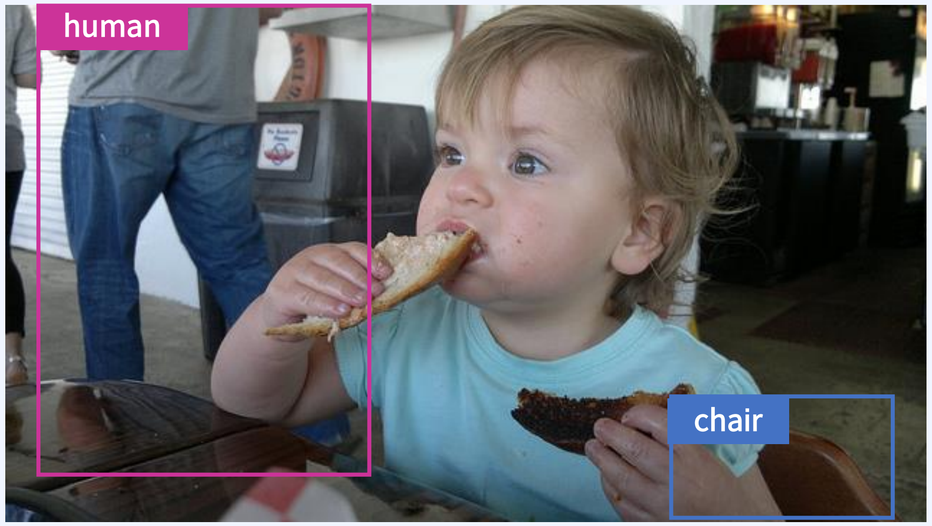
◦ Sequential HOI Detectors

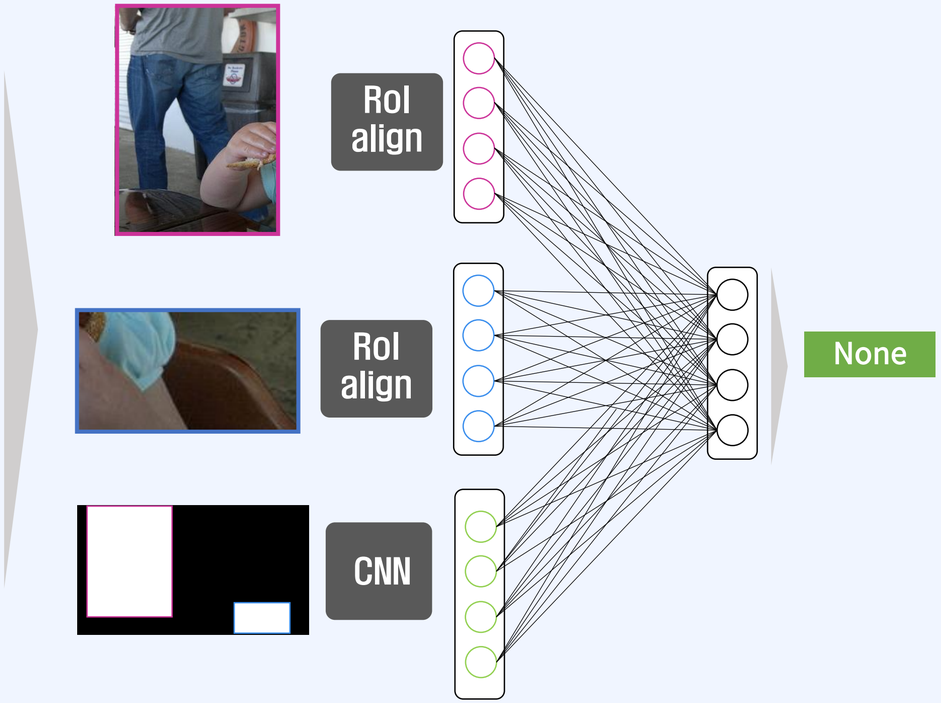


◦ Sequential HOI Detectors



◦ Sequential HOI Detectors

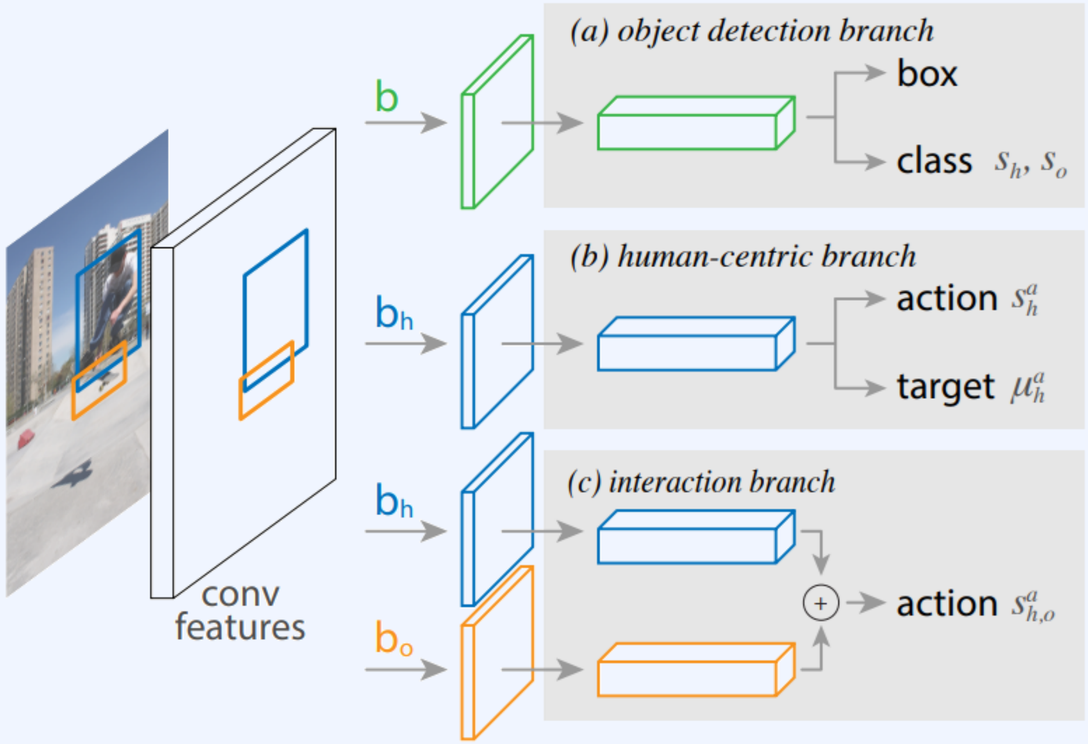




◦ Sequential HOI Detectors

G. Gkioxari et al. Detecting and Recognizing Human-Object Interactions. CVPR

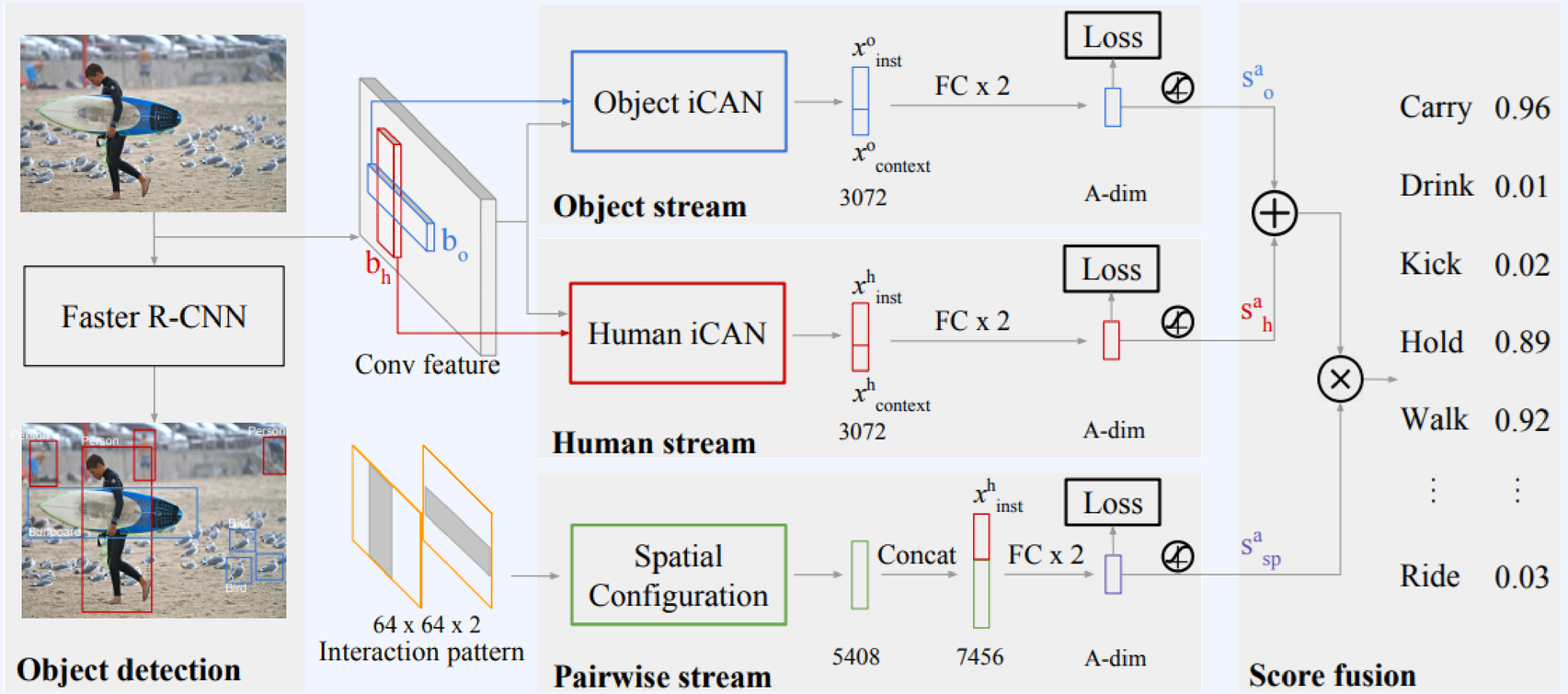
InteractNet



◦ Sequential HOI Detectors

C. Gao et al. iCAN: Instance-Centric Attention Network for Human-Object Interaction Detection. BMVC

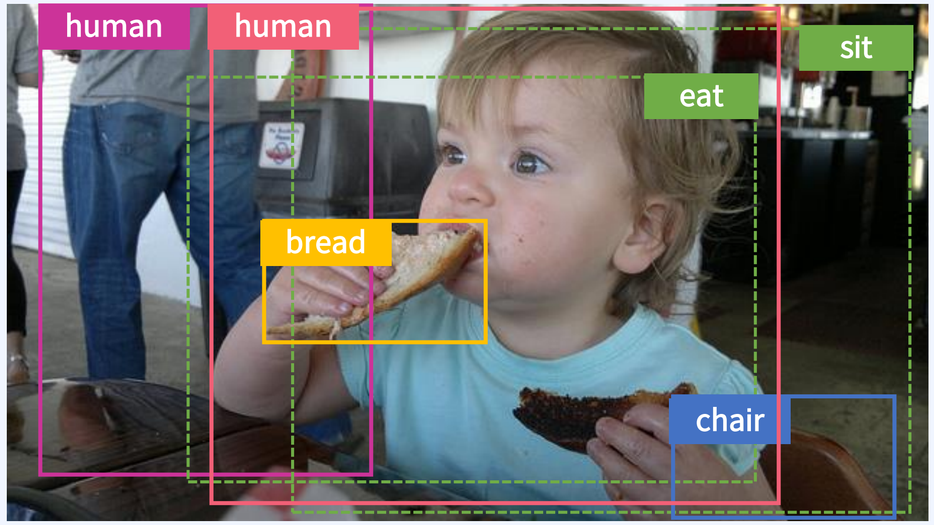
iCAN



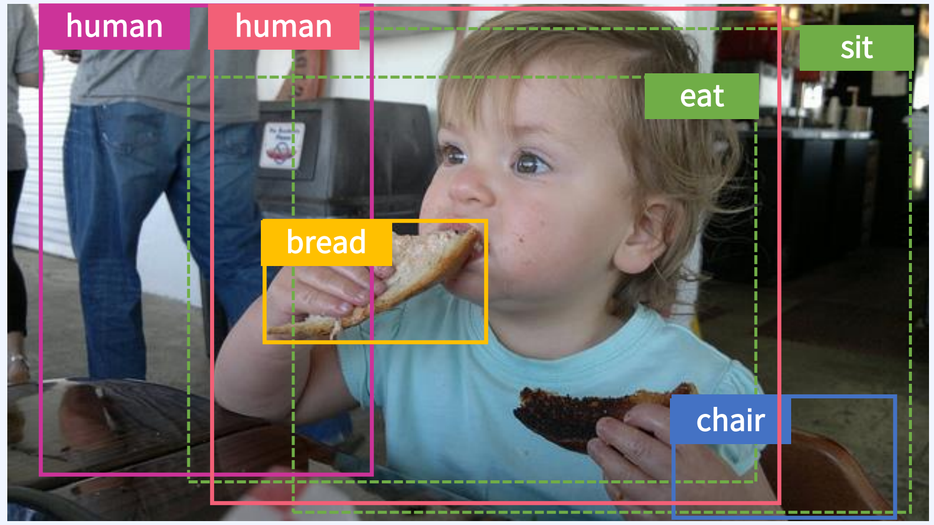
◦ Sequential HOI Detectors

- Intuitive Pipeline
- Pairwise Neural Network Inference : Slow

◦ Parallel HOI Detectors

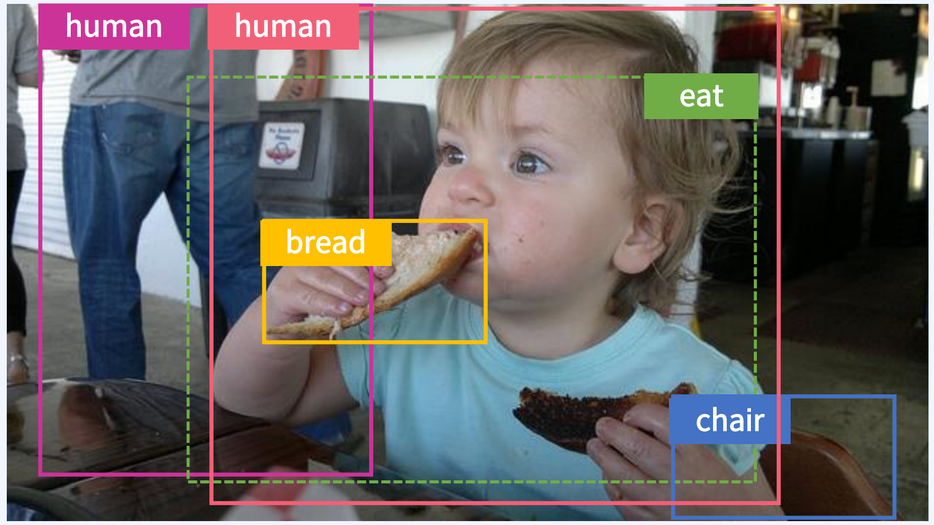


◦ Parallel HOI Detectors

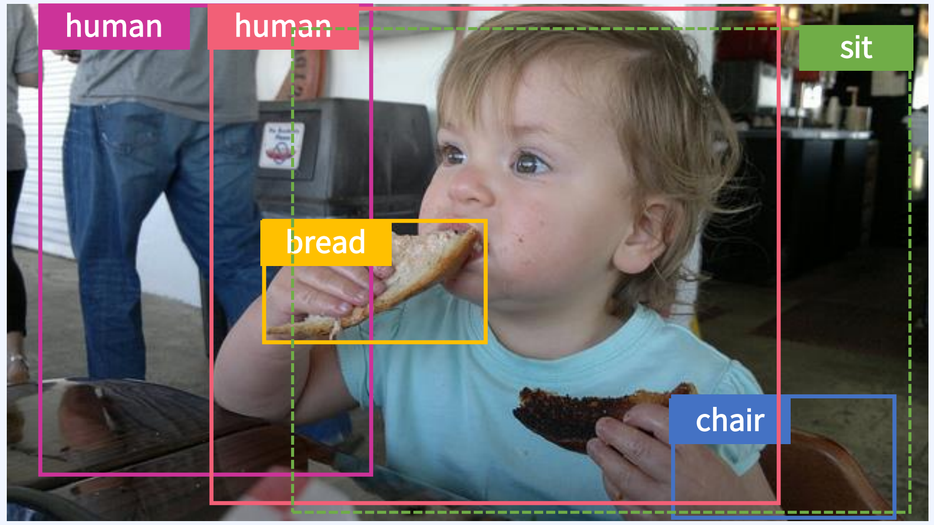


Region of Interaction

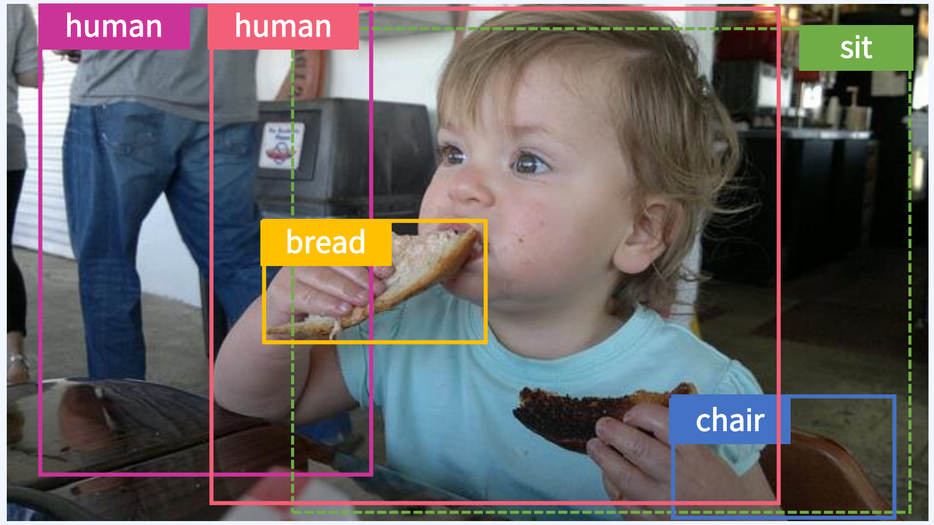
◦ Parallel HOI Detectors



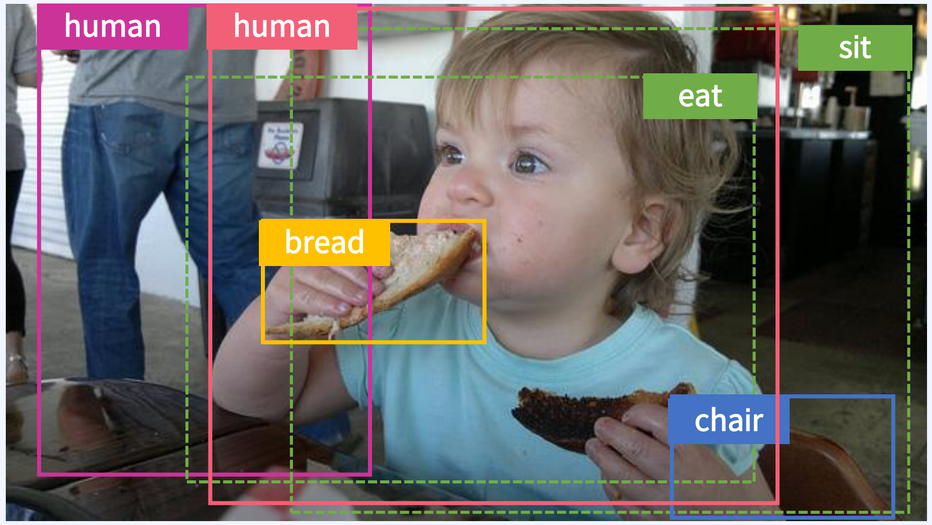
◦ Parallel HOI Detectors

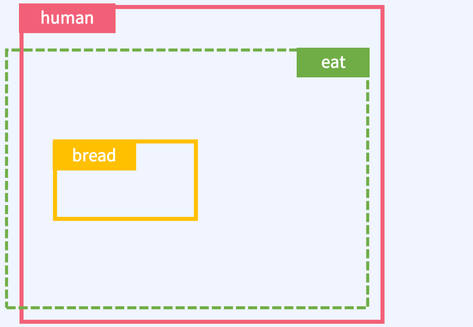


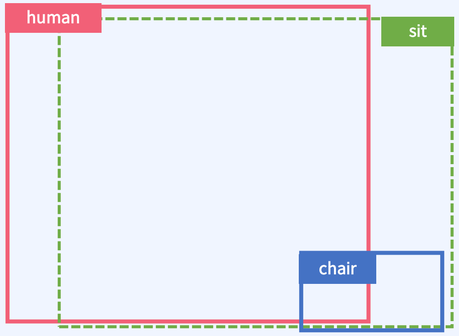
◦ Parallel HOI Detectors



◦ Parallel HOI Detectors



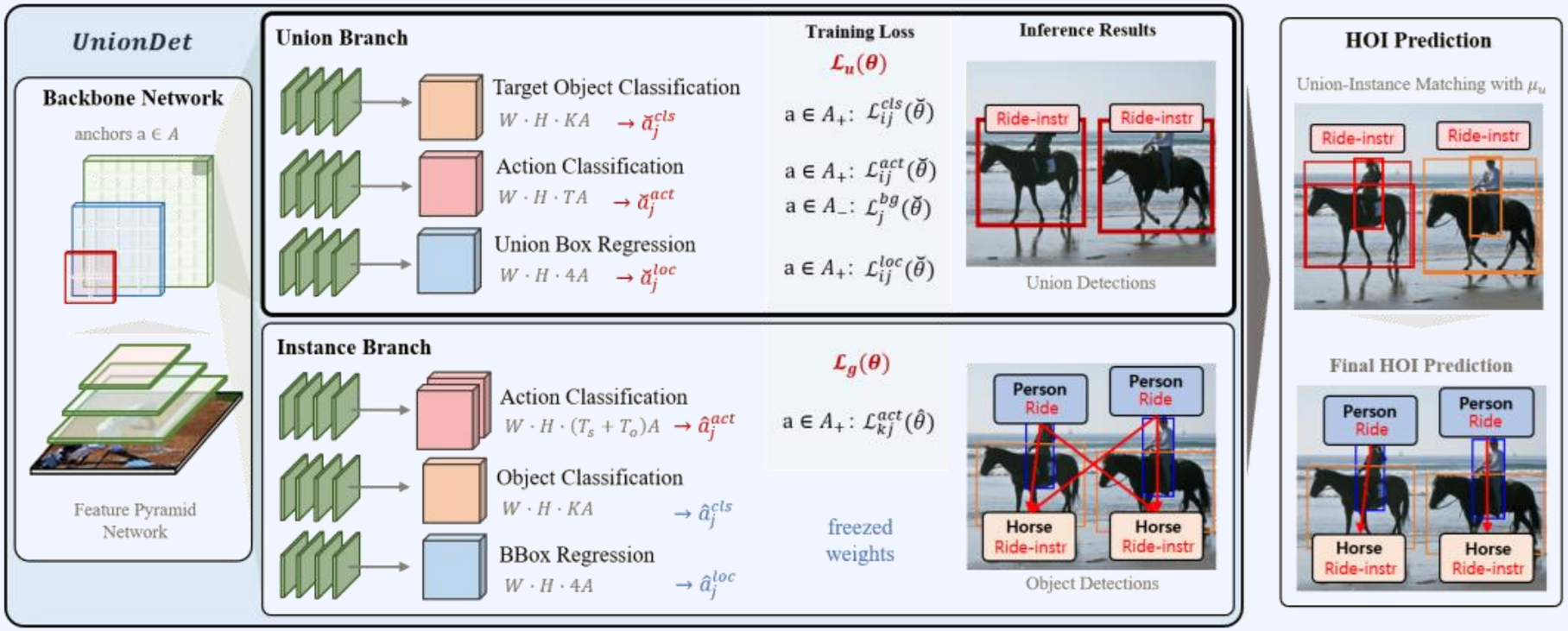




◦ Parallel HOI Detectors

B. Kim et al. UnionDet: Union-Level Detector Towards Real-Time Human-Object Interaction Detection. ECCV

UnionDet

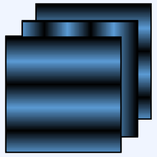


Context Understanding
Transformers for OD and HOI

- Sequential HOI Detectors
  - Intuitive Pipeline
  - Pairwise Neural Network Inference : Slow
- Parallel HOI Detectors
  - Define “region of interaction” : Union / Interaction
  - Speed-up in HOI inference time
  - However, the triplet search is still a bottleneck and has room for improvement

Context Understanding
Transformers for OD and HOI
HOTR

Convolutional Neural Network

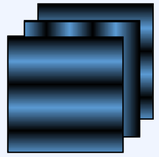


Positional Encoding

B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR

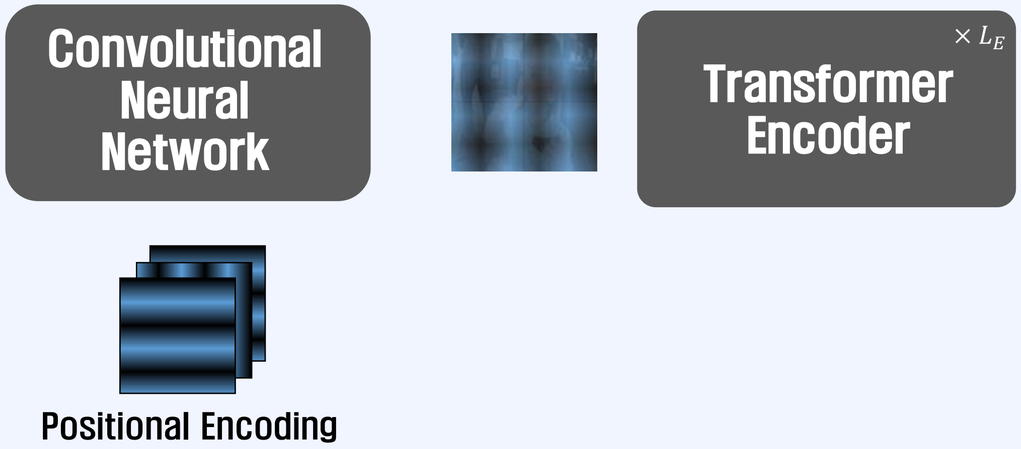
Convolutional Neural Network



Positional Encoding

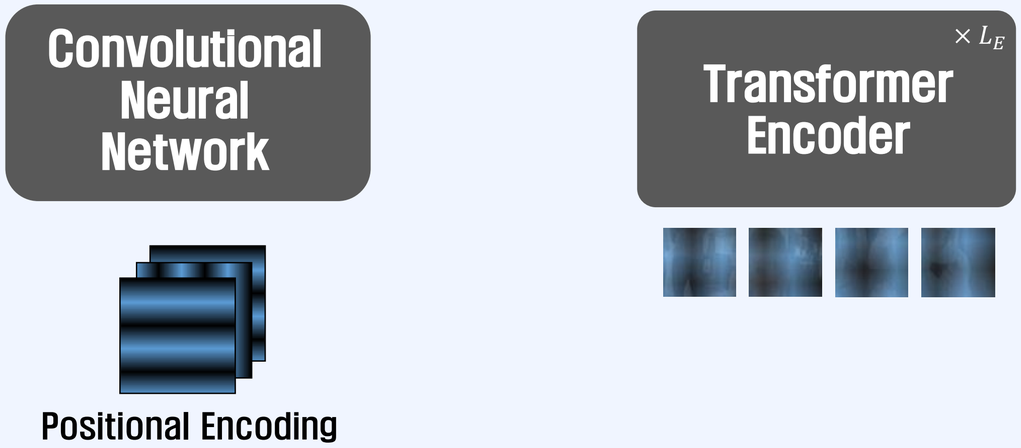
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

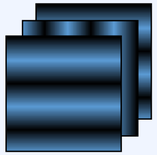
Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR

Convolutional Neural Network



Positional Encoding

Transformer Encoder $\times L_E$

Instance Queries

Instance Decoder $\times L_O$

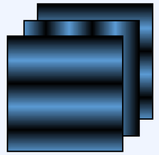
Interaction Queries

Interaction Decoder $\times L_I$

B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR

Convolutional Neural Network



Positional Encoding

Transformer Encoder $\times L_E$

Instance Queries

Instance Decoder $\times L_O$

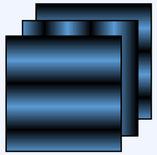
Interaction Queries

Interaction Decoder $\times L_I$

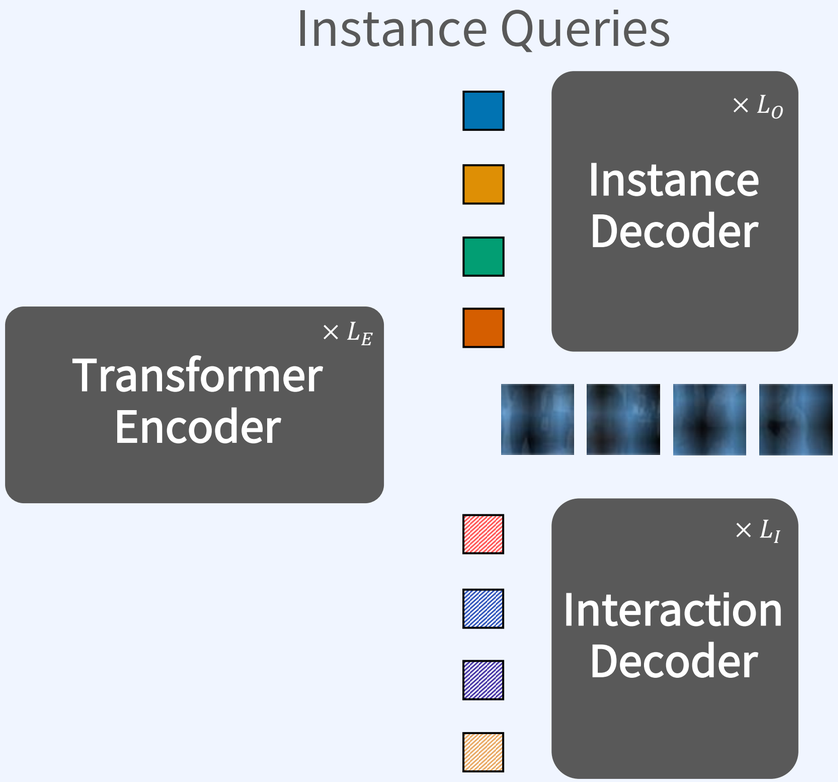
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR

Convolutional Neural Network



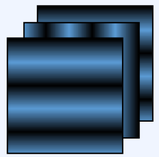
Positional Encoding



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR

Convolutional Neural Network



Positional Encoding

Transformer Encoder $\times L_E$

Instance Decoder $\times L_O$

FFN

1 class, box
2 class, box
3 class, box
4 class, box

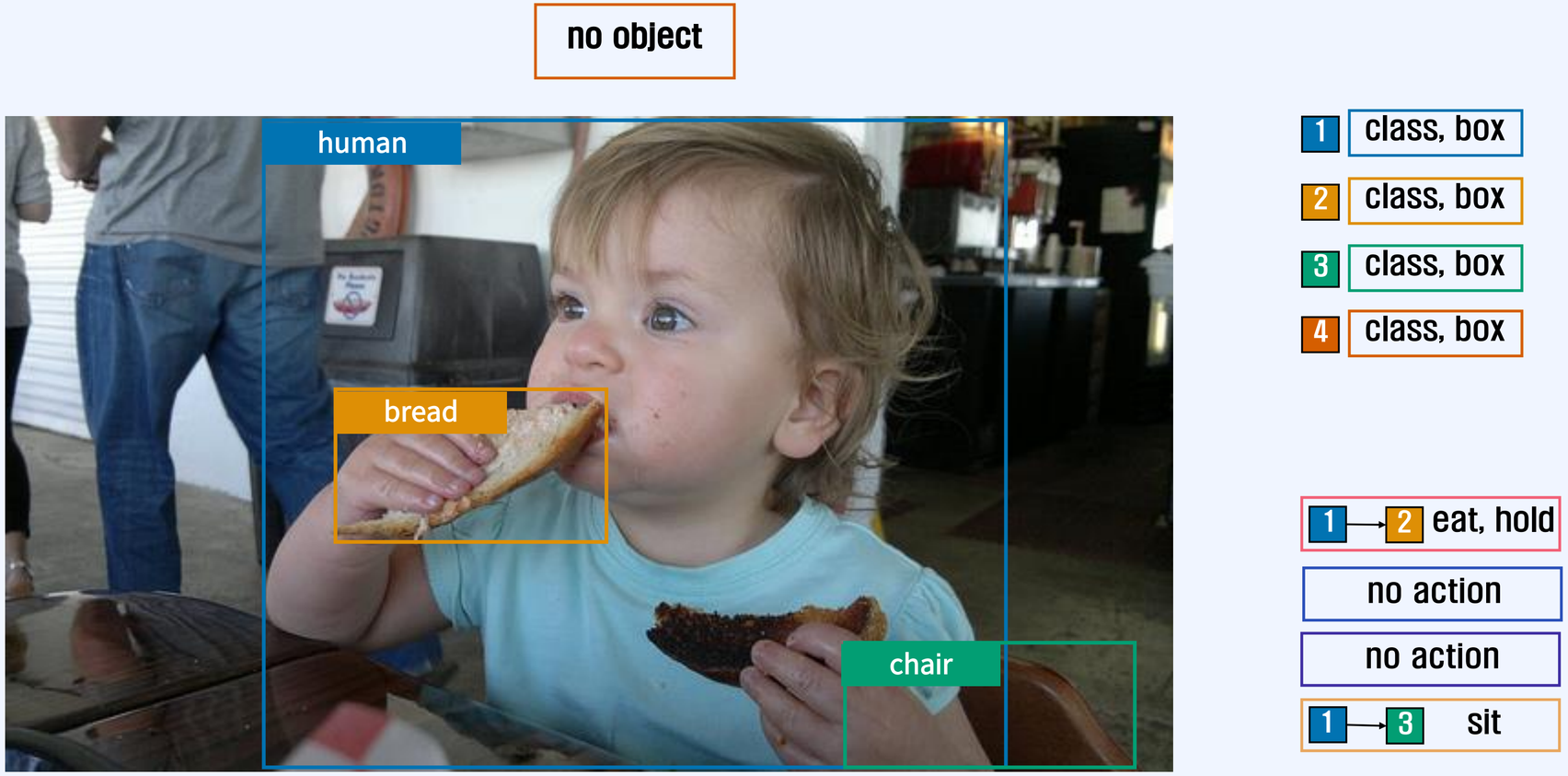
Interaction Decoder $\times L_I$

FFN

1 → 2 eat, hold
no action
no action
1 → 3 sit

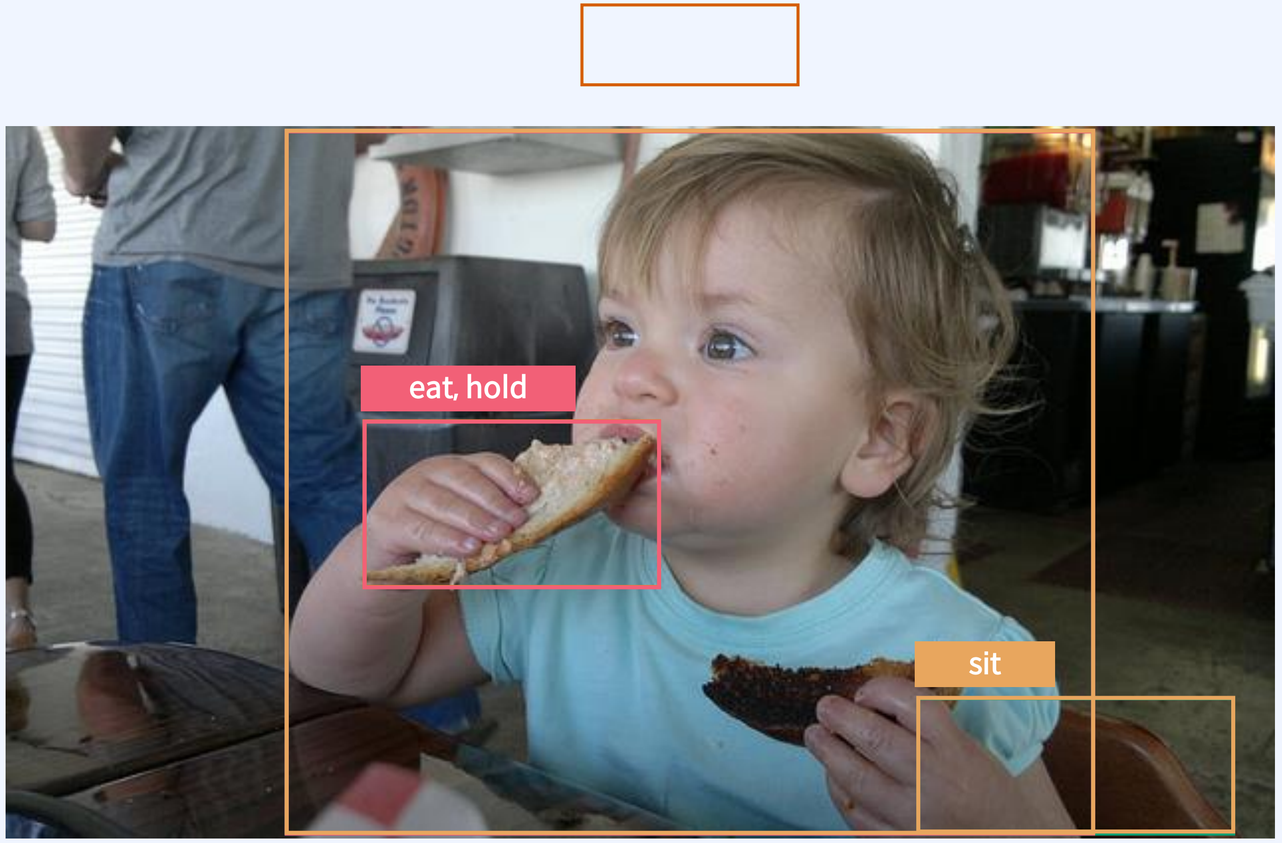
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR

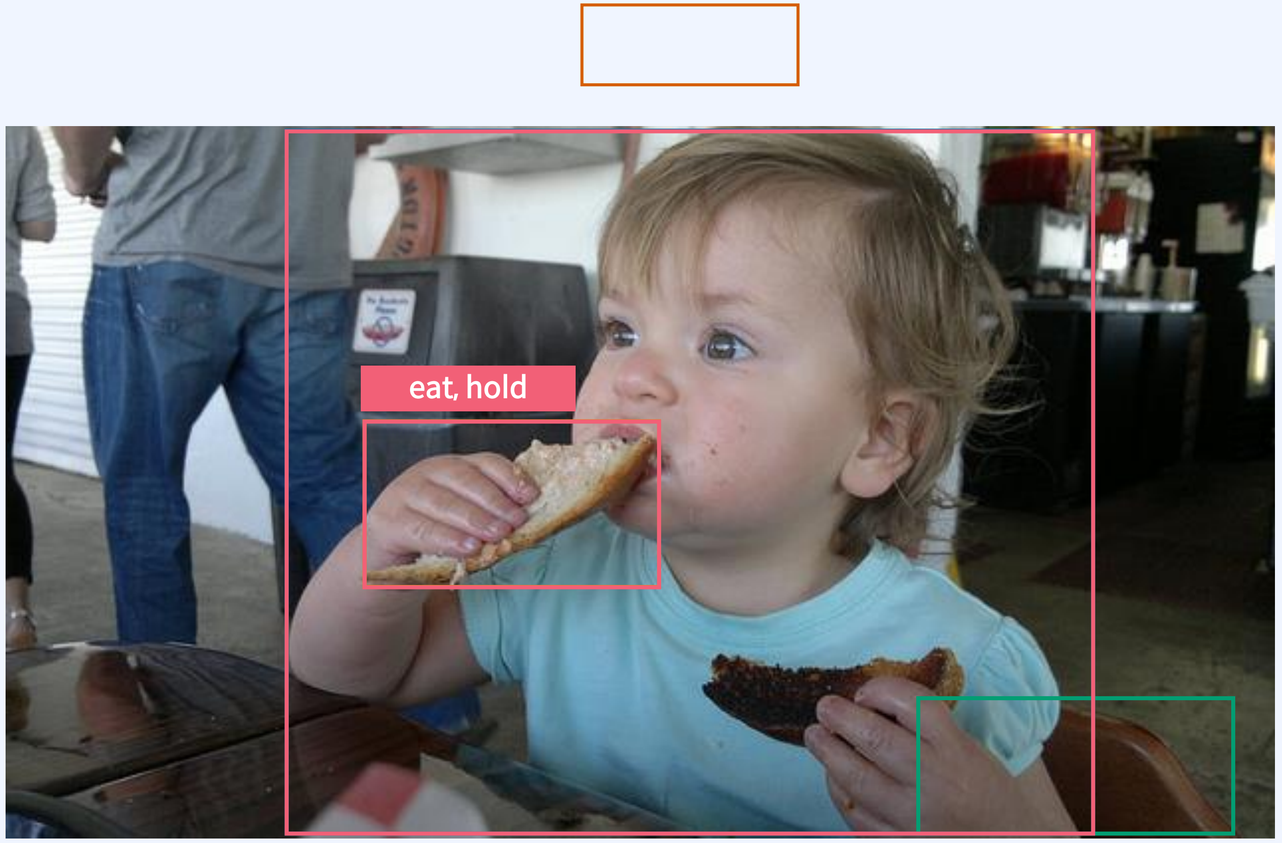


1 class, box
2 class, box
3 class, box
4 class, box

1 → 2 eat, hold
no action
no action
1 → 3 sit

B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



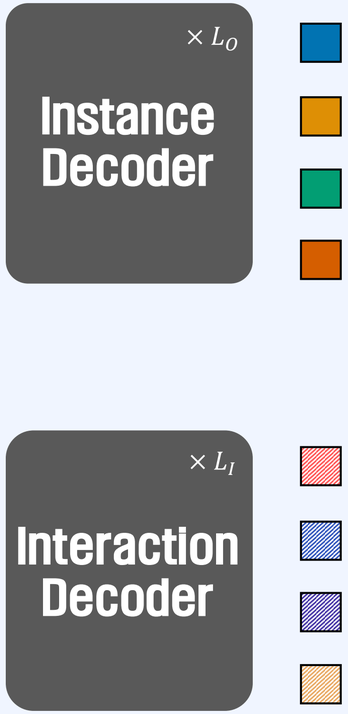
1 class, box
2 class, box
3 class, box
4 class, box

1 → 2 eat, hold
no action
no action
1 → 3 sit

HO Pointers

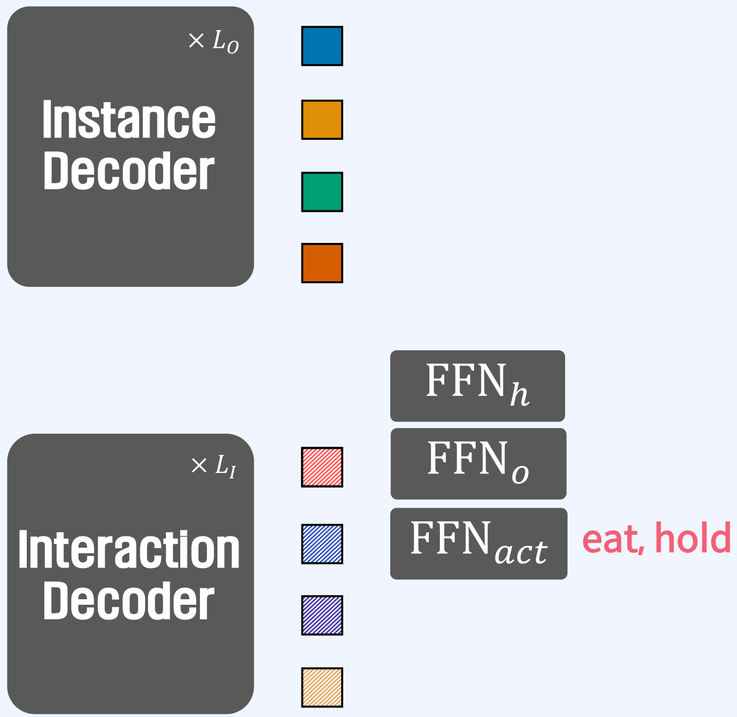
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



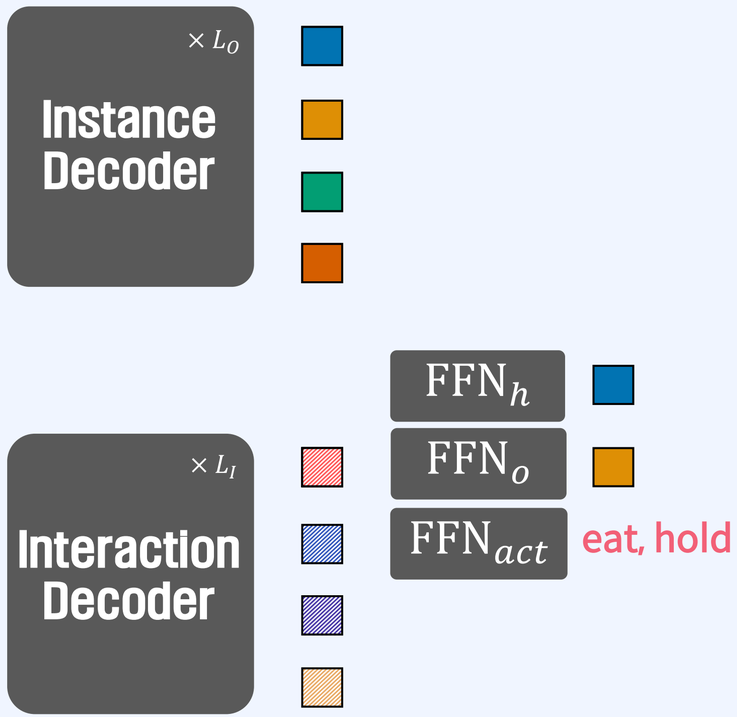
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



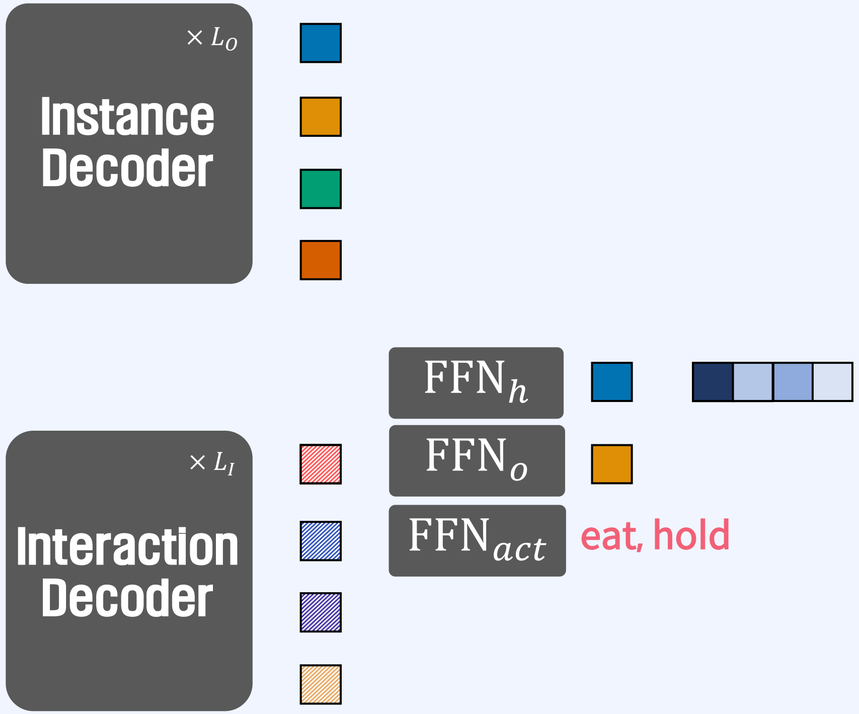
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



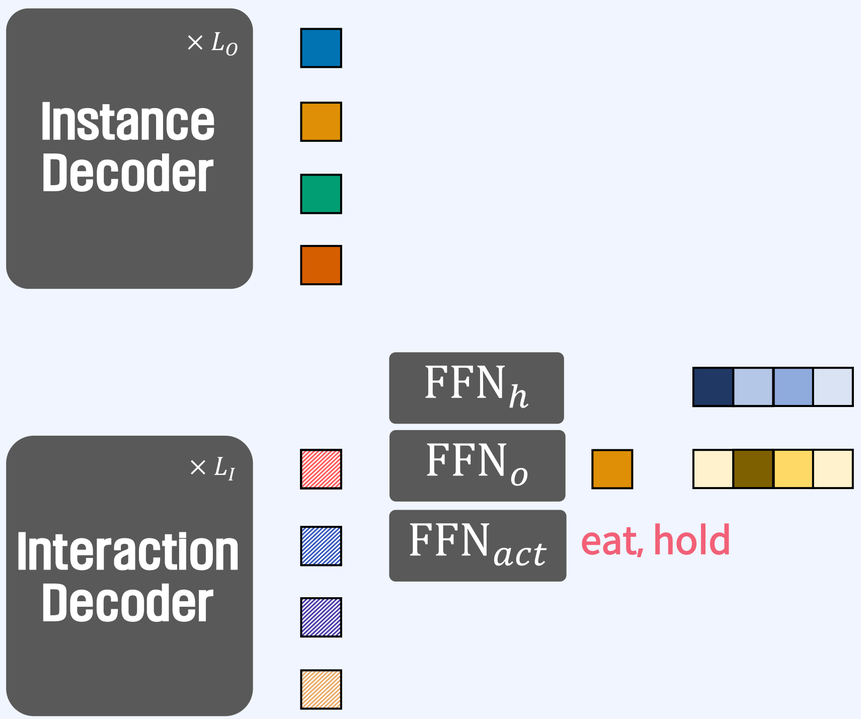
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



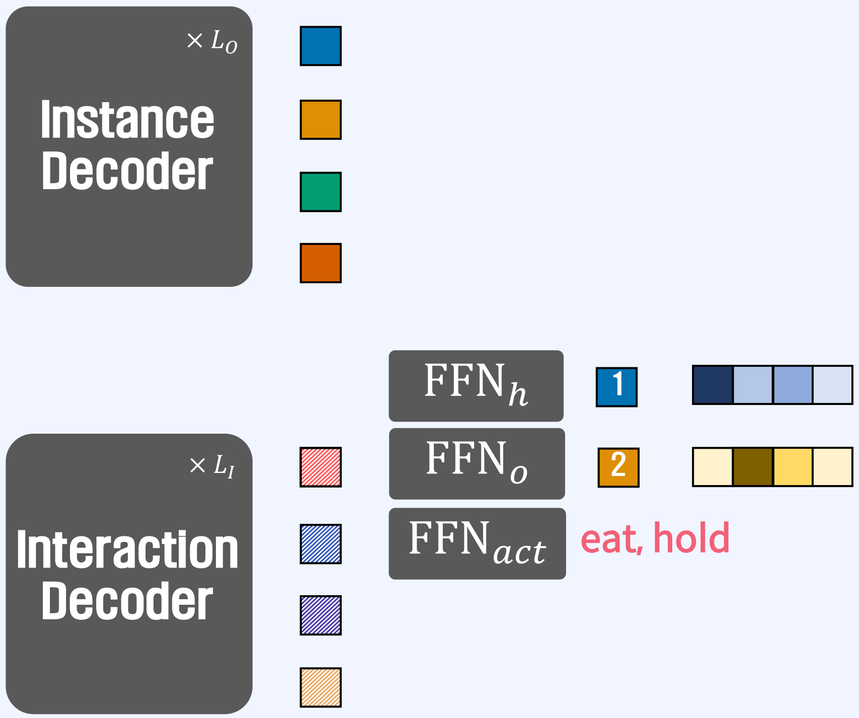
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



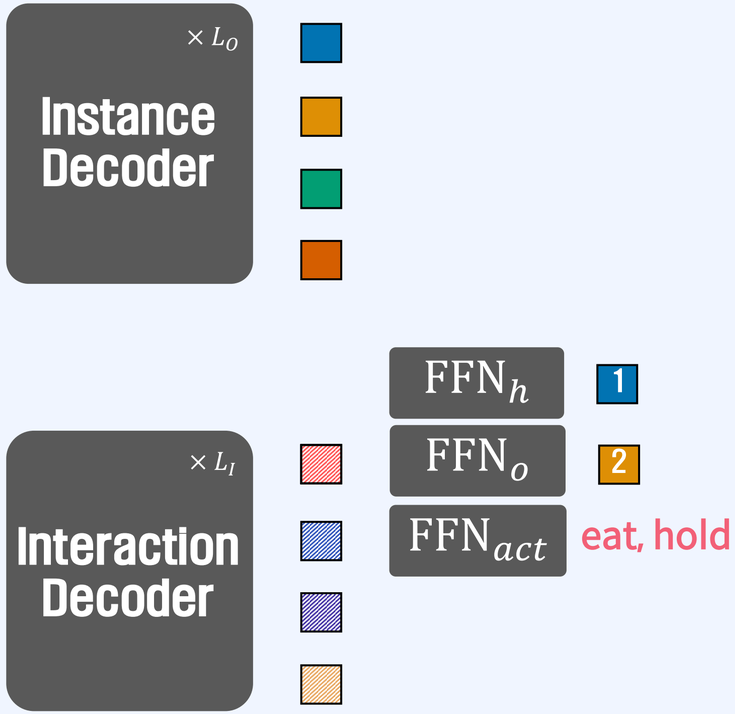
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

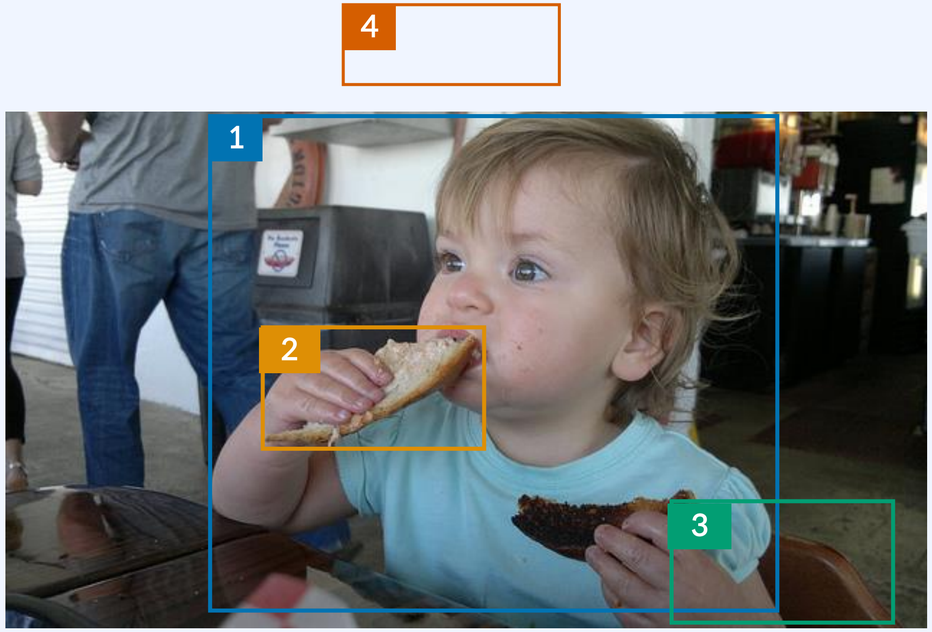
Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

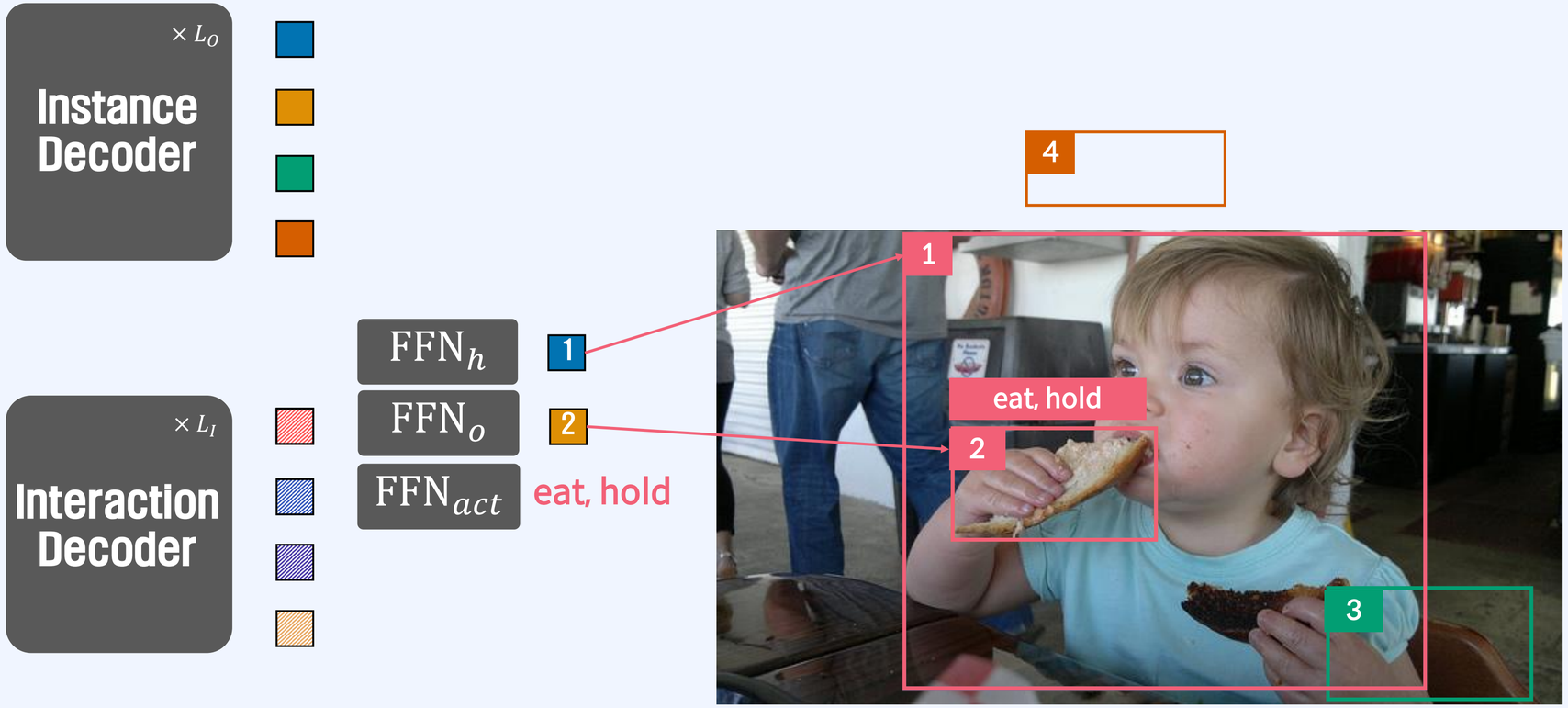
Context Understanding
Transformers for OD and HOI
HOTR





B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



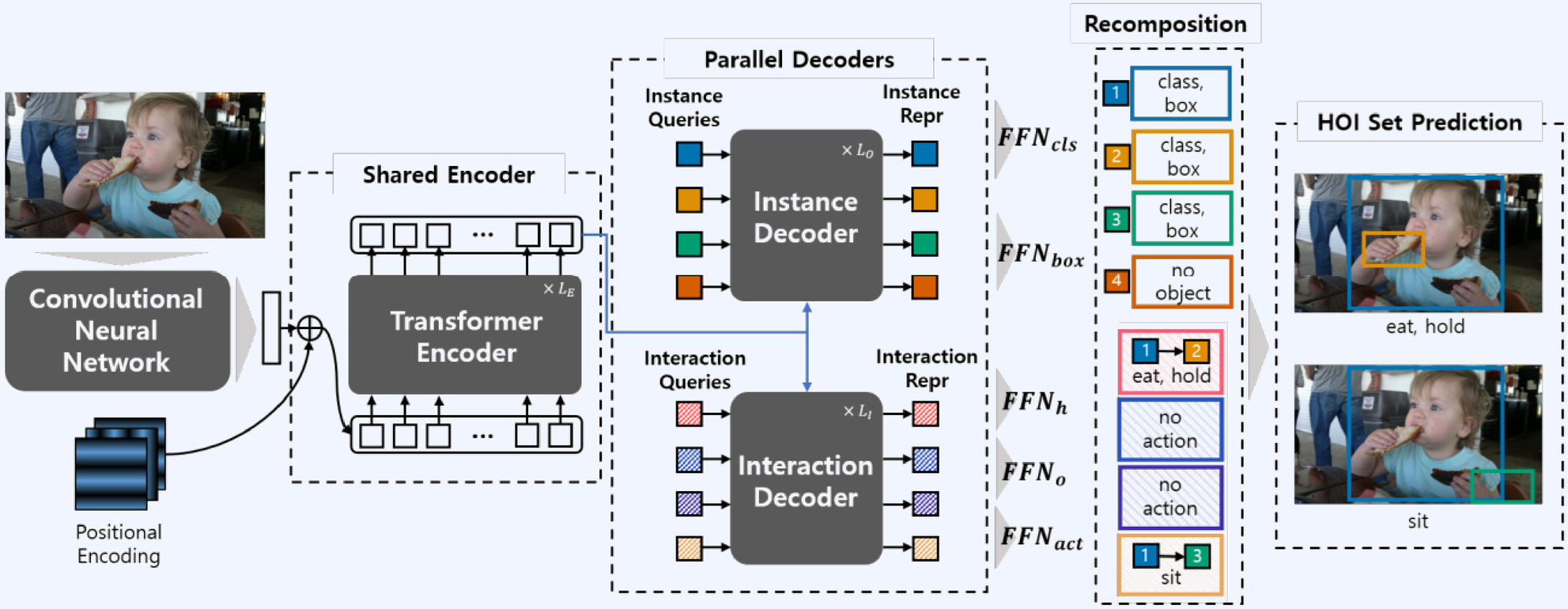
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding Transformers for OD and HOI

HOTR

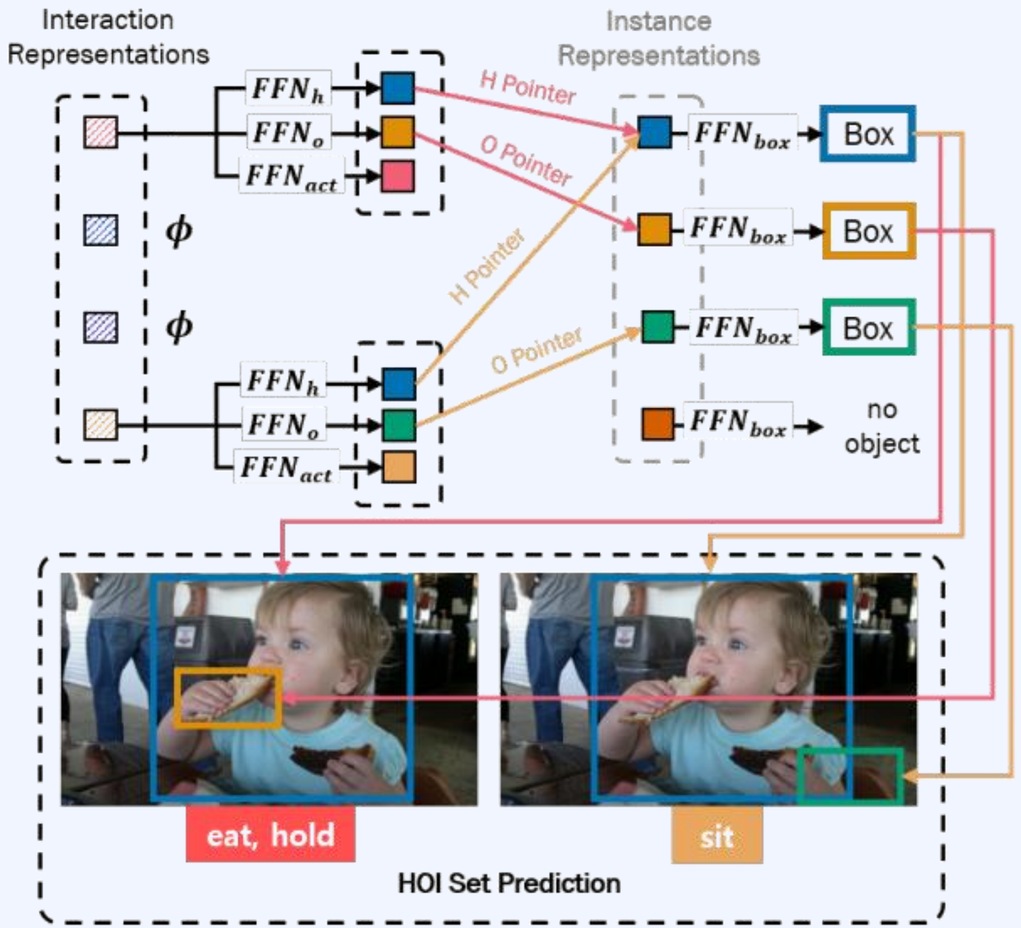
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

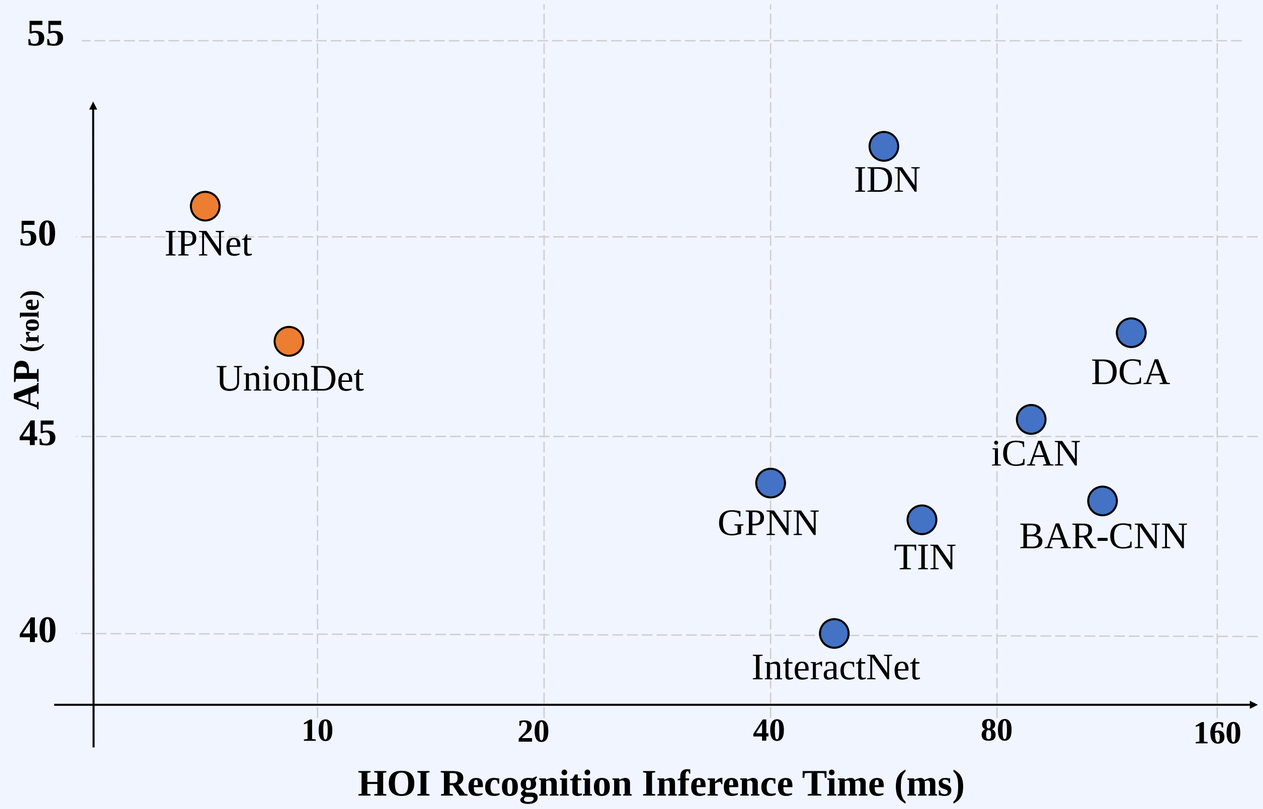
Context Understanding
Transformers for OD and HOI
HOTR

Method	Backbone	$AP^{\#1}_{\text{role}}$	$AP^{\#2}_{\text{role}}$
*Models with **external features***			
TIN ($\text{RP}_\text{D}\text{C}_\text{D}$) [18]	R50	47.8	
Verb Embedding [31]	R50	45.9	
RPNN [33]	R50	-	47.5
PMFNet [27]	R50-FPN	52.0	
PastaNet [17]	R50-FPN	51.0	57.5
PD-Net [32]	R50	52.0	-
ACP [13]	R152	53.0	
FCMNet [20]	R50	53.1	-
ConsNet [21]	R50-FPN	53.2	-
Sequential HOI Detectors			
VSRL [8]	R50-FPN	31.8	-
InteractNet [6]	R50-FPN	40.0	48.0
BAR-CNN [14]	R50-FPN	43.6	-
GPNN [24]	R152	44.0	-
iCAN [5]	R50	45.3	52.4
TIN (RC_D) [18]	R50	43.2	-
DCA [29]	R50	47.3	-
VSGNet [26]	R152	51.8	57.0
VCL [10]	R50-FPN	48.3	
DRG [4]	R50-FPN	51.0	
IDN [16]	R50	53.3	60.3
Parallel HOI Detectors			
IPNet [30]	HG104	51.0	-
UnionDet [12]	R50-FPN	47.5	56.2
Ours	R50	**55.2**	**64.4**

				Default		
Method	Detector	Backbone	Feature	Full	Rare	Non Rare
Sequential HOI Detectors						
InteractNet [6]	COCO	R50-FPN	A	9.94	7.16	10.77
GPNN [24]	COCO	R101	A	13.11	9.41	14.23
iCAN [5]	COCO	R50	A+S	14.84	10.45	16.15
DCA [29]	COCO	R50	A+S	16.24	11.16	17.75
TIN [18]	COCO	R50	A+S+P	17.03	13.42	18.11
RPNN [33]	COCO	R50	A+P	17.35	12.78	18.71
PMFNet [27]	COCO	R50-FPN	A+S+P	17.46	15.65	18.00
No-Frills HOI [9]	COCO	R152	A+S+P	17.18	12.17	18.68
DRG [4]	COCO	R50-FPN	A+S+L	19.26	17.74	19.71
VCL [10]	COCO	R50	A+S	19.43	16.55	20.29
VSGNet [26]	COCO	R152	A+S	19.80	16.05	20.91
FCMNet [20]	COCO	R50	A+S+P	20.41	17.34	21.56
ACP [13]	COCO	R152	A+S+P	20.59	15.92	21.98
PD-Net [32]	COCO	R50	A+S+P+L	20.81	15.90	22.28
DJ-RN [15]	COCO	R50	A+S+V	21.34	18.53	22.18
ConsNet [21]	COCO	R50-FPN	A+S+L	22.15	17.12	23.65
PastaNet [17]	COCO	R50	A+S+P+L	22.65	21.17	23.09
IDN [16]	COCO	R50	A+S	23.36	**22.47**	23.63
Functional Gen. [1]	HICO-DET	R101	A+S+L	21.96	16.43	23.62
TIN [18]	HICO-DET	R50	A+S+P	22.90	14.97	25.26
VCL [10]	HICO-DET	R50	A+S	23.63	17.21	25.55
ConsNet [21]	HICO-DET	R50-FPN	A+S+L	24.39	17.10	26.56
DRG [4]	HICO-DET	R50-FPN	A+S	24.53	19.47	26.04
IDN [16]	HICO-DET	R50	A+S	24.58	**20.33**	25.86
Parallel HOI Detectors						
UnionDet [12]	COCO	R50-FPN	A	14.25	10.23	15.46
IPNet [30]	COCO	R50-FPN	A	19.56	12.79	21.58
Ours	COCO	R50	A	**23.46**	16.21	**25.62**
UnionDet [12]	HICO-DET	R50-FPN	A	17.58	11.72	19.33
PPDM [19]	HICO-DET	HG104	A	21.10	14.46	23.09
Ours	HICO-DET	R50	A	**25.10**	17.34	**27.42**

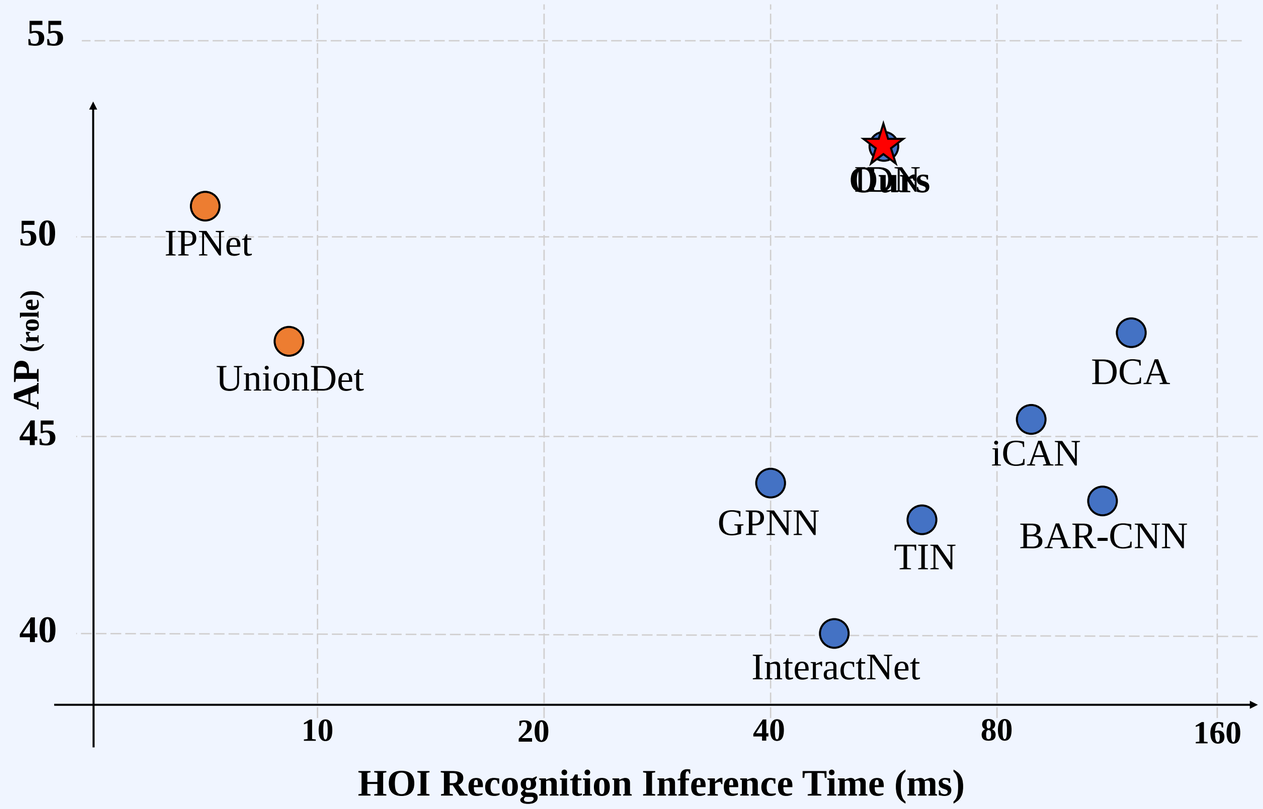
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



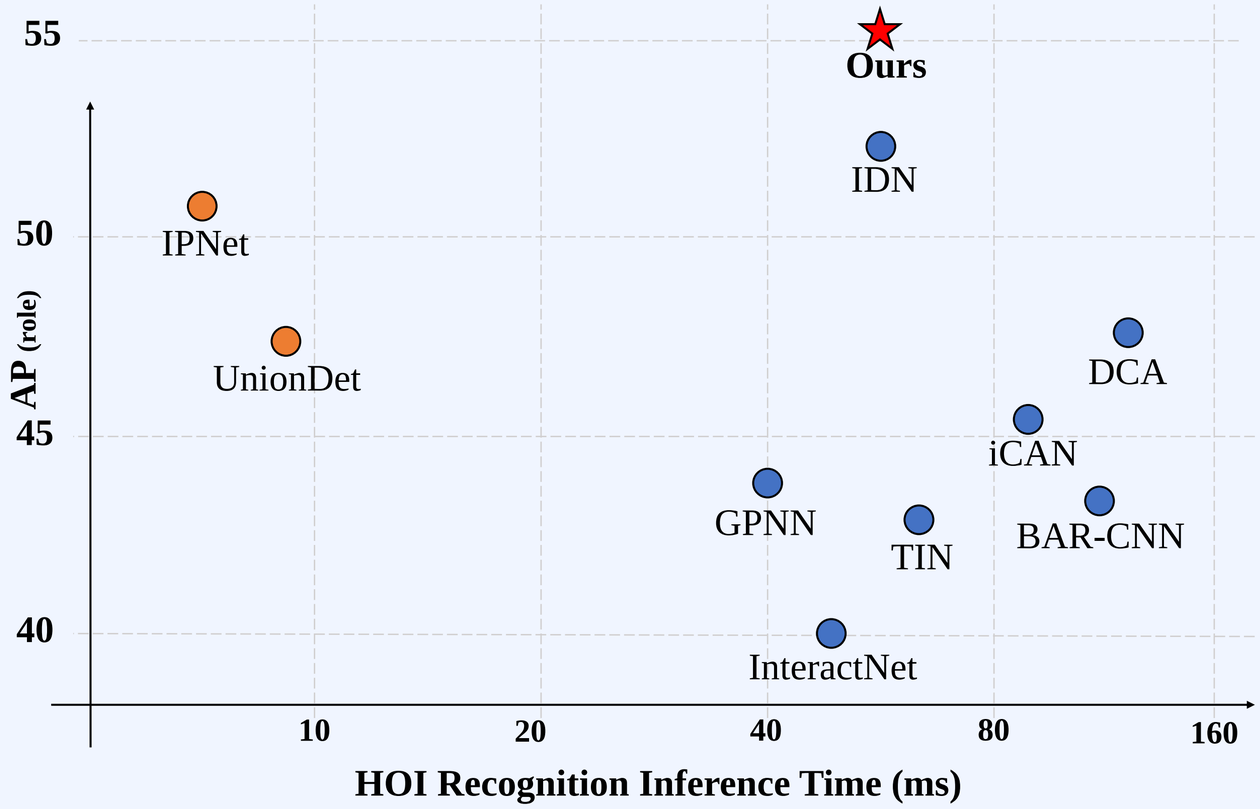
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



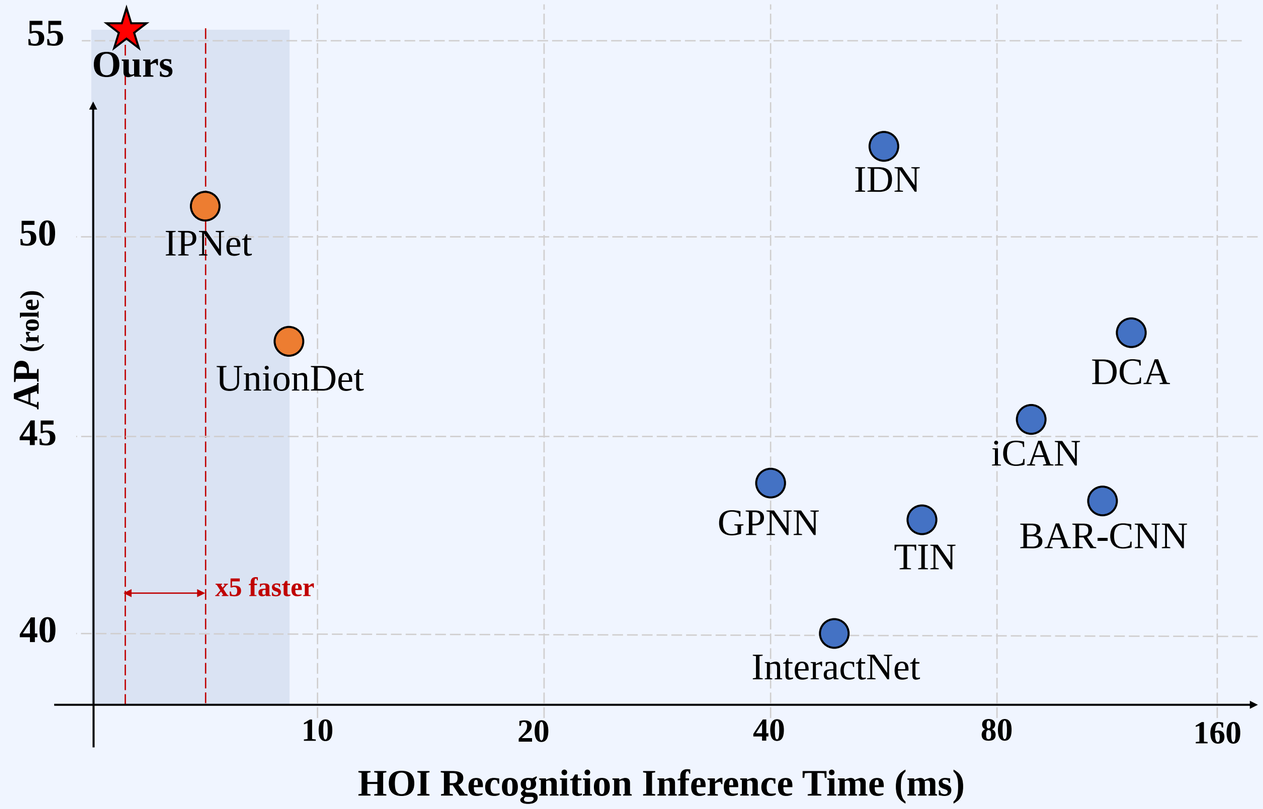
B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR

Context Understanding
Transformers for OD and HOI
HOTR



B. Kim et al. HOTR: End-to-End Human-Object Interaction Detection with Transformers. CVPR