QAHOI: Query-Based Anchors for Human-Object Interaction Detection

Junwen Chen and Keiji Yanai

The University of Electro-Communications
Tokyo

Human-Object Interaction (HOI) Detection

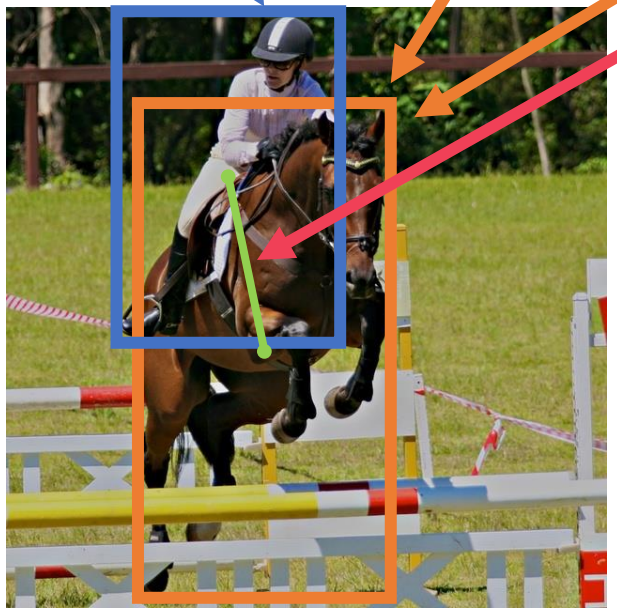
☐ HOI Detection

- Predict a set of <human, object, interaction> triplets within an image

☐ HOI Instance

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

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



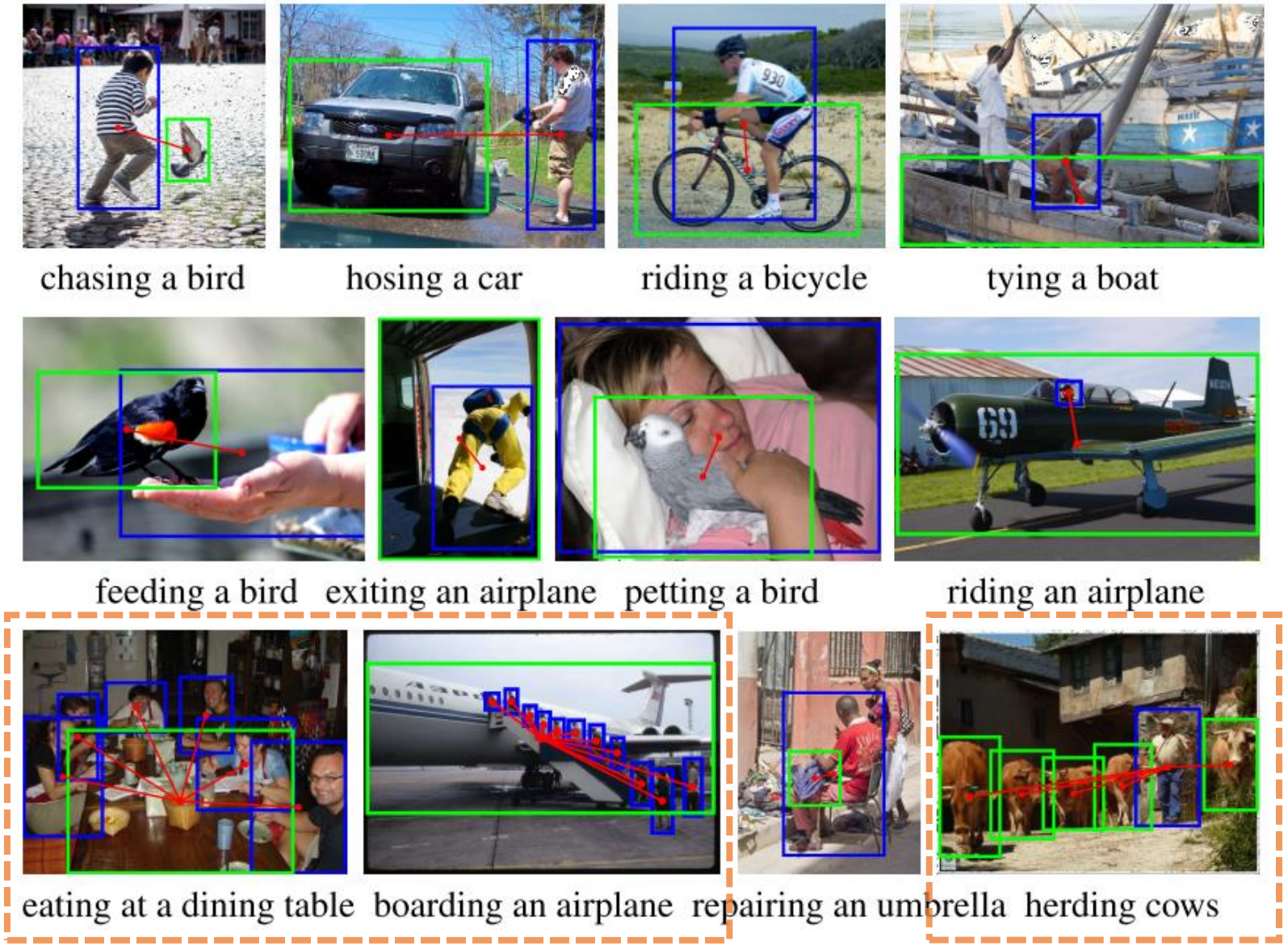
HICO-DET

☐ HOI benchmark

- Training 38,118
- Test: 9,658

☐ Diversity

- 117 action classes
- COCO's 80 object classes
- 600 HOI classes



[1] Chao, Yu-Wei, et al. "Learning to detect human-object interactions." *2018 ieee winter conference on applications of computer vision (wacv)*. IEEE, 2018.

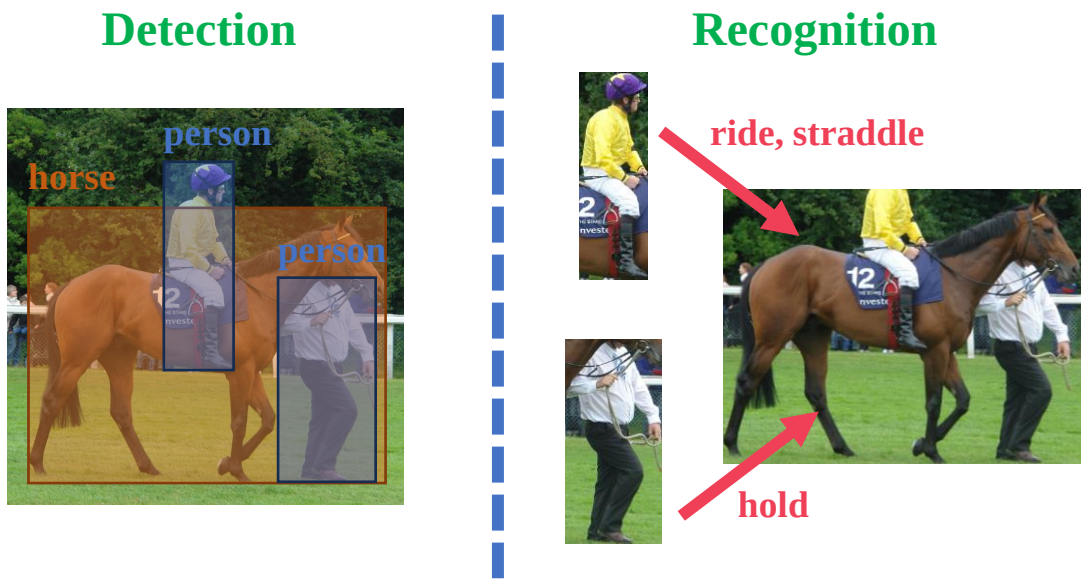
HOI Detection Approaches

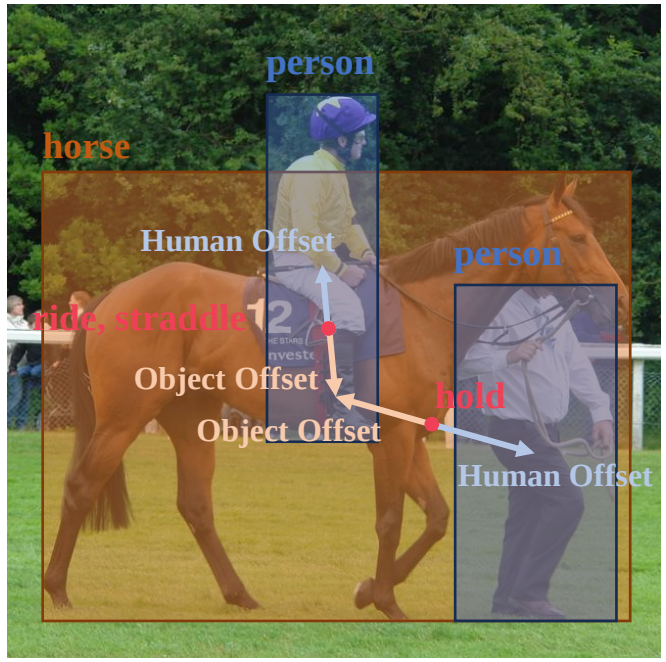
☐ CNN-based Two-stage (Bottom-up)

- Build upon an off-the-shelf object detector
- Object & Human Detection ➡ Interaction Recognition on Pairs

☐ CNN-based One-stage (Top-down)

- Interaction Points & HOI Pair Matching

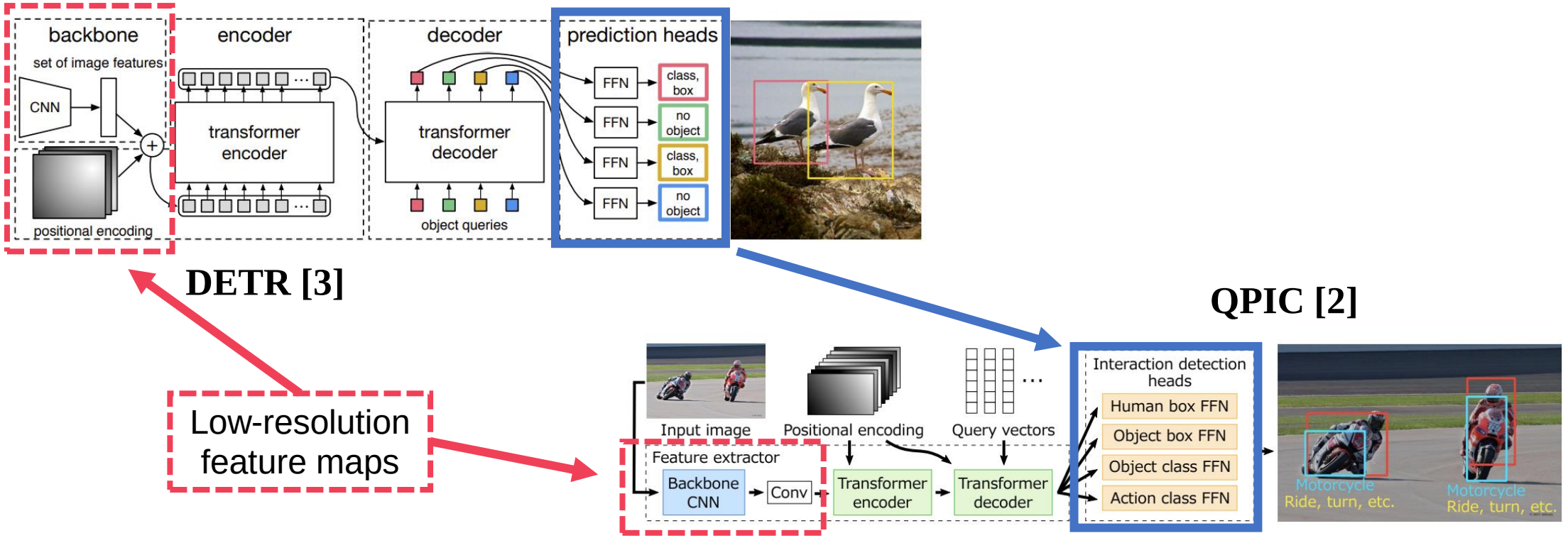




HOI Detection Approaches

☐ Transformer-based One-stage

- Adapted from Transformer-based object detector DETR
- Set-based Prediction

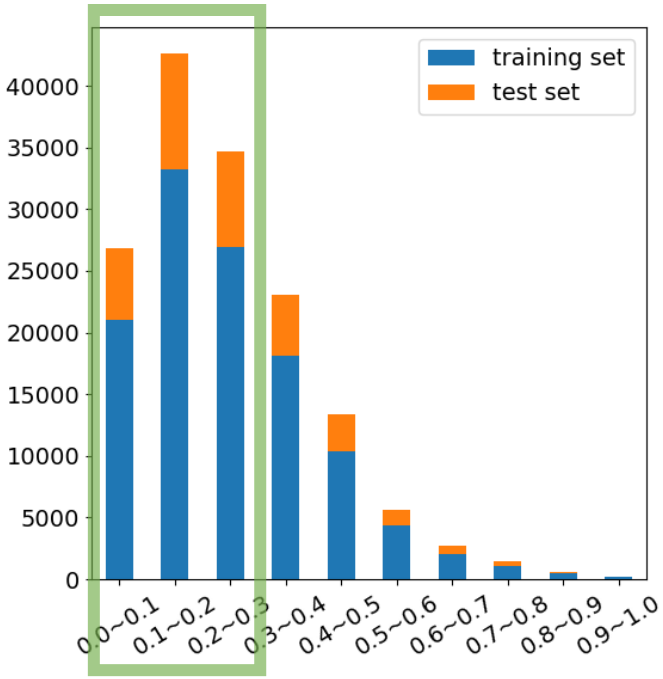


[2] Tamura, Masato, Hiroki Ohashi, and Tomoaki Yoshinaga. "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[3] Carion, Nicolas, et al. "End-to-end object detection with transformers." European conference on computer vision. Springer, Cham, 2020.

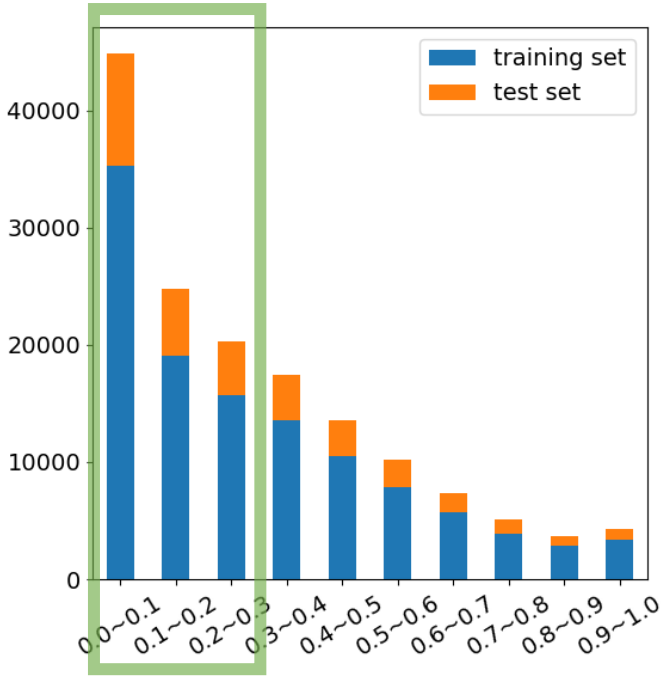
☐ The spatial distribution of the HOI instances in HICO-DET

- Small objects & Close human-object pairs
- High-resolution feature maps are better to restore detailed features

☐ Transformer-based methods lack a multi-scale architecture



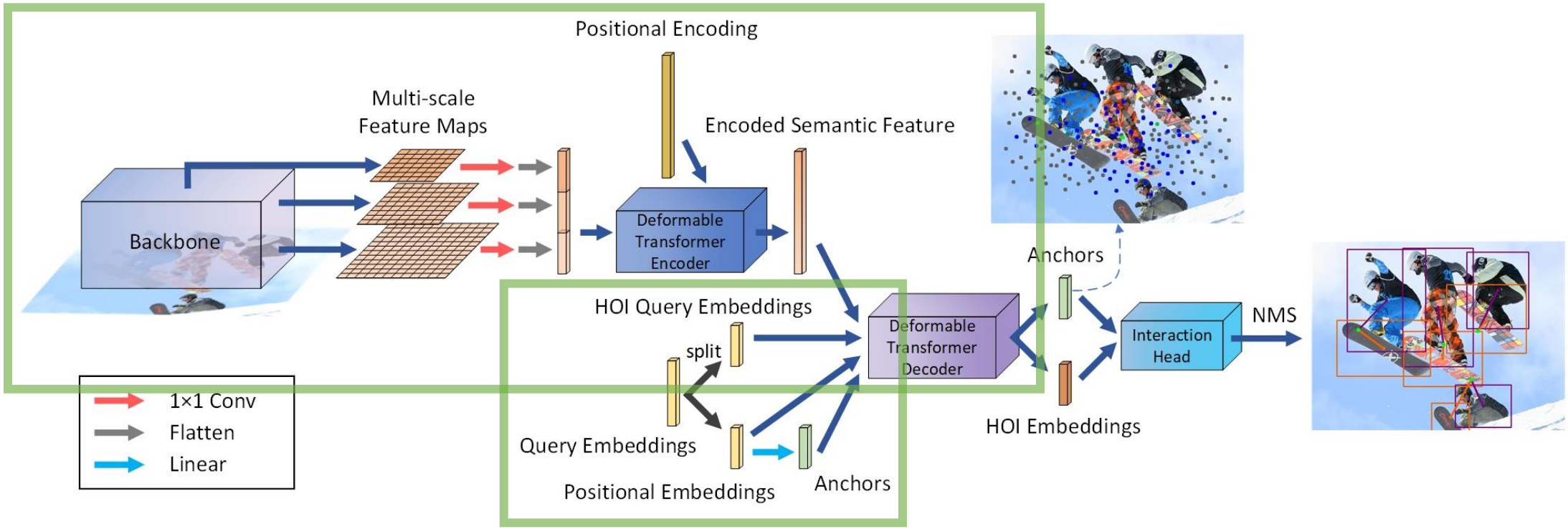
(a) Larger Area



(b) Center Distance

Overview of QAHOI

- **Multi-scale feature maps** from a hierarchical backbone
- A new representation of HOI instances: **Query-based Anchors**
- **Deformable Transformer** Encoder-Decoder Architecture [4]
- Training from scratch



[4] Zhu, Xizhou, et al. "Deformable DETR: Deformable Transformers for End-to-End Object Detection." International Conference on Learning Representations. 2020.

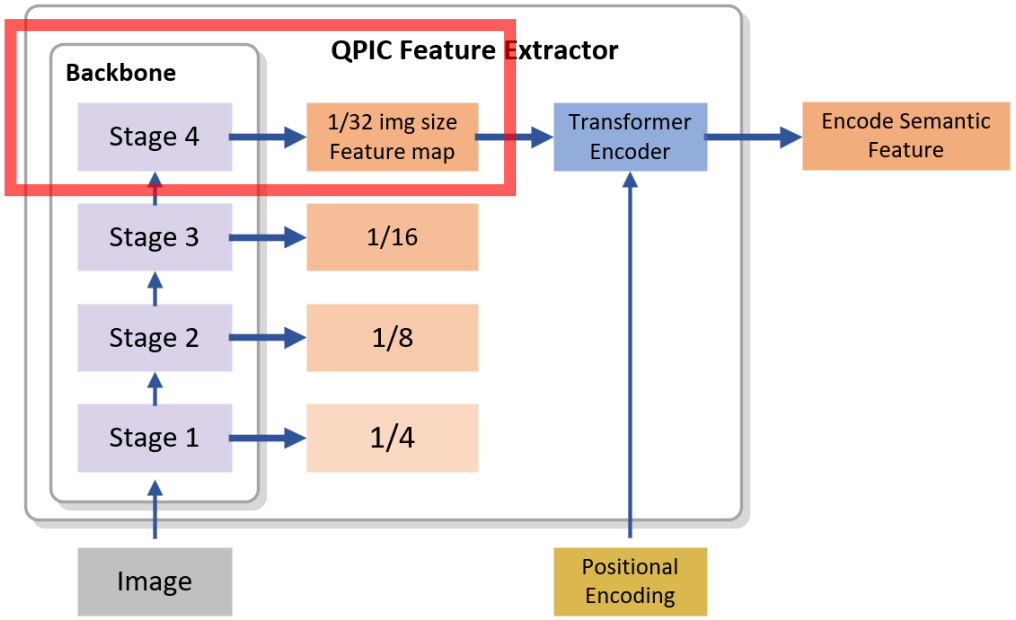
Multi-scale Feature Extractor

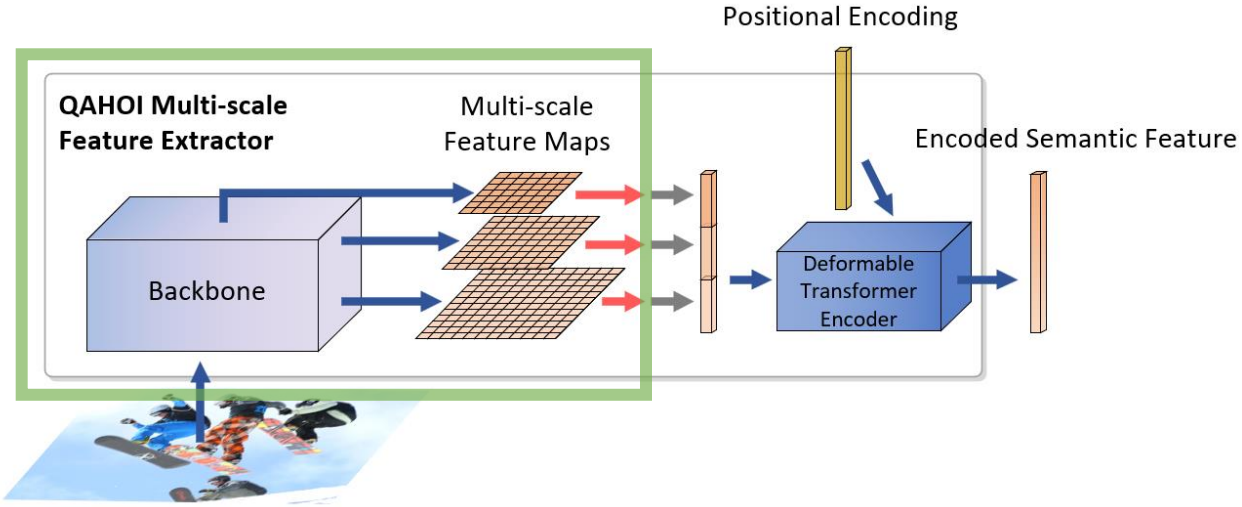
☐ Feature Extractor of QPIC

- CNN Backbone + Transformer Encoder [5]
- Low-resolution feature maps from last Stage

☐ Multi-scale Feature Extractor of QAHOI

- Hierarchical Backbone (CNN-based or Transformer-based) + Deformable Transformer Encoder
- Multi-scale feature maps from multiple stages

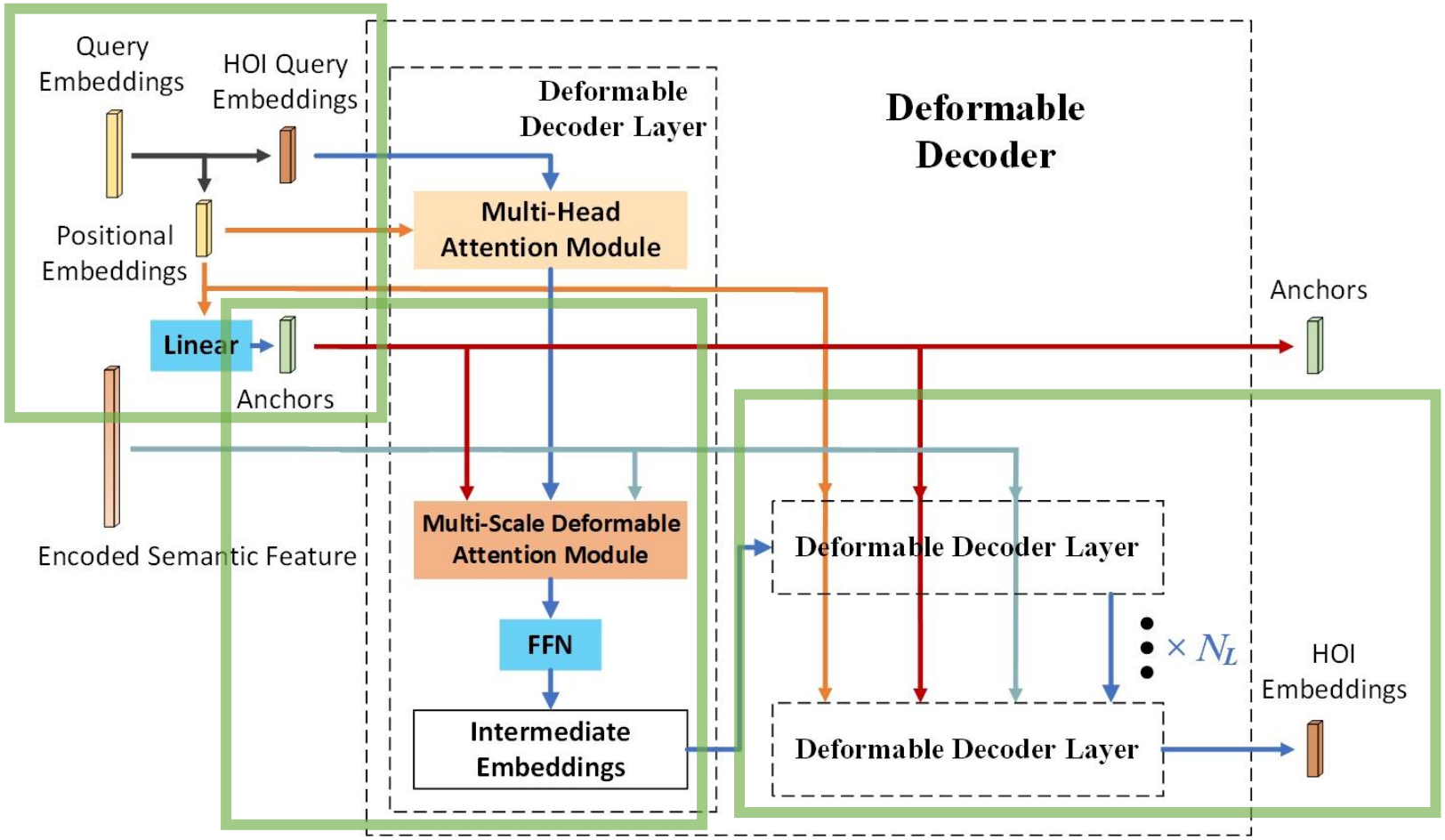




[5] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.

Predicting HOI with Query-Based Anchors

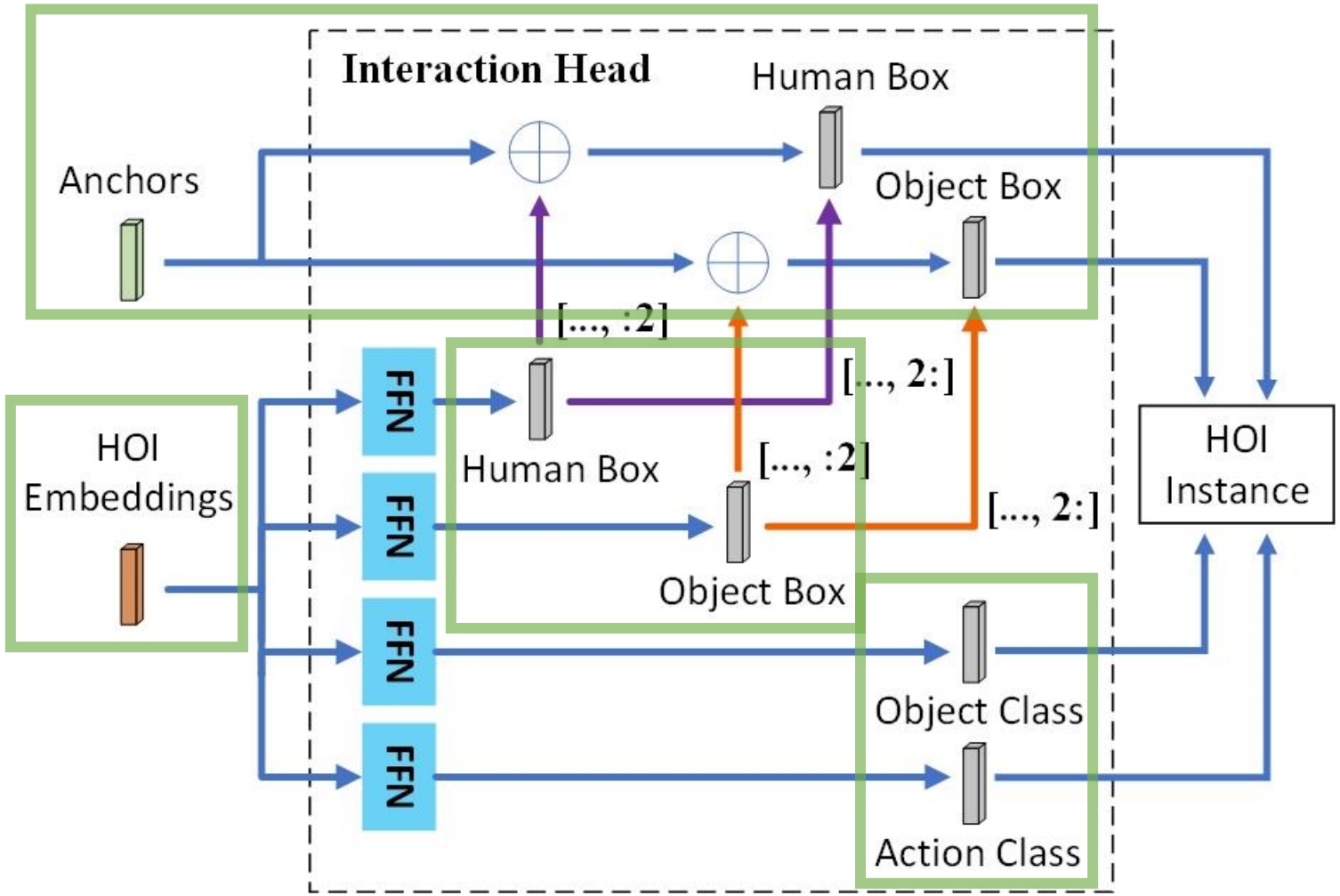
- Anchor guided Decoding
- Multi-scale cross-attention for multi-scale context features



Deformable Transformer Decoder of QAHOI

Anchor-based Interaction Detection Heads

- Predict boxes according to anchors
- Human-object pairs are combined with corresponding anchors



Top K Scores and HOI NMS

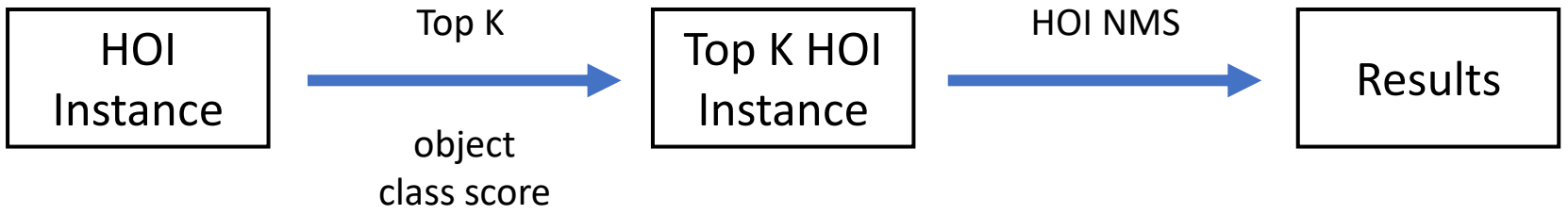
☐ Top K

- Selected by object class scores

☐ HOI NMS

- Based on the IoU and HOI score $c_{\text{HOI}} = c_o \cdot c_a$

$$\text{IoU}(i, j) = \text{IoU}(B_i^{(h)}, B_j^{(h)}) \cdot \text{IoU}(B_j^{(o)}, B_j^{(o)})$$



Comparison with State-of-the-Arts

- Best Model: QAHOI with Swin-Transformer [6] Backbone
- 150 epochs of training

Arch.	Method	Backbone	Fine-tuned Detection	Default			Known Object		
				Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
Points	IP-Net [16]	ResNet-50-FPN	✗	19.56	12.79	21.58	22.05	15.77	23.92
	PPDM [9]	Hourglass-104	✓	21.73	13.78	24.10	24.58	16.65	26.84
	GGNet [18]	Hourglass-104	✓	23.47	16.48	25.60	27.36	20.23	29.48
Query	HOITrans [20]	ResNet-101	✓	26.61	19.15	28.84	29.13	20.98	31.57
	HOTR [7]	ResNet-50	✗	23.46	16.21	25.65	-	-	-
	HOTR [7]	ResNet-50	✓	25.10	17.34	27.42	-	-	-
	AS-Net [3]	ResNet-50	✗	24.40	22.39	25.01	27.41	25.44	28.00
	AS-Net [3]	ResNet-50	✓	28.87	24.25	30.25	31.74	27.07	33.14
	QPIC [15]	ResNet-101	✓	29.90	23.92	31.69	32.38	26.06	34.27
	QAHOI	**Swin-Tiny**	✗	28.47	22.44	30.27	30.99	24.83	32.84
	QAHOI	**Swin-Base**	✗	29.47	22.24	31.63	31.45	24.00	33.68
	QAHOI	**Swin-Base**$^{*+}$	✗	33.58	25.86	35.88	35.34	27.24	37.76
	QAHOI	**Swin-Large**$^{*+}$	✗	**35.78**	**29.80**	**37.56**	**37.59**	**31.66**	**39.36**

+5.88 (19.7%)

+4.1 (13.9%)

[5] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

Ablation Study

☐ Training Strategy and Multi-scale Feature Maps

Arch.	Model	Backbone	Fine-tuned Detection	Multi-scale	Default *Full*	Default *Rare*	Default *Non-Rare*
QPIC	(1)	ResNet-50	✗	x_3	24.21	17.51	26.21
	(2)	ResNet-50	✓	x_3	29.07	21.85	31.23
	(3)	Swin-Tiny	✗	x_3	27.19	21.32	28.95
QAHOI	(4)	ResNet-50	✗	x_1, x_2, x_3, x_4	24.35	16.18	26.80
	(5)	ResNet-50	✓	x_1, x_2, x_3, x_4	26.18	18.06	28.61
	(6)	Swin-Tiny	✓	x_1, x_2, x_3	30.15	22.83	32.34
	(7)	Swin-Tiny	✗	x_1, x_2, x_3, x_4	28.09	21.65	30.01
	(8)	Swin-Tiny	✗	x_1, x_2, x_3	28.47	22.44	30.27
	(9)	Swin-Tiny	✗	x_2, x_3	28.12	20.43	30.41
	(10)	Swin-Tiny	✗	x_3	26.65	19.13	28.89

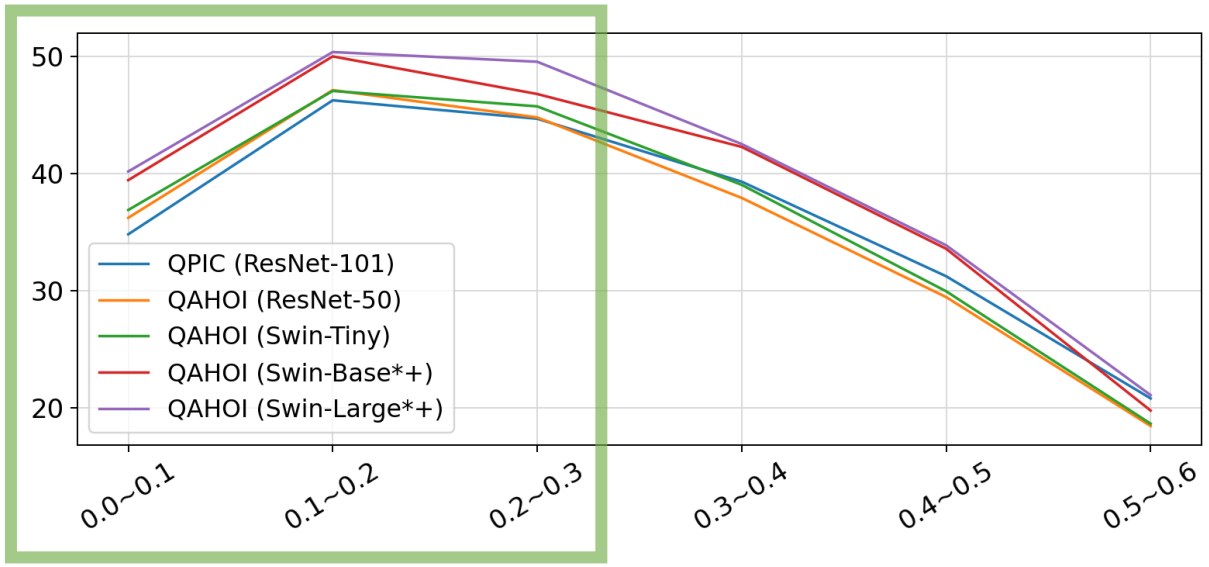
Without Pre-training, QAHOI is better than QPIC

Contribution of Multi-scale Feature Maps

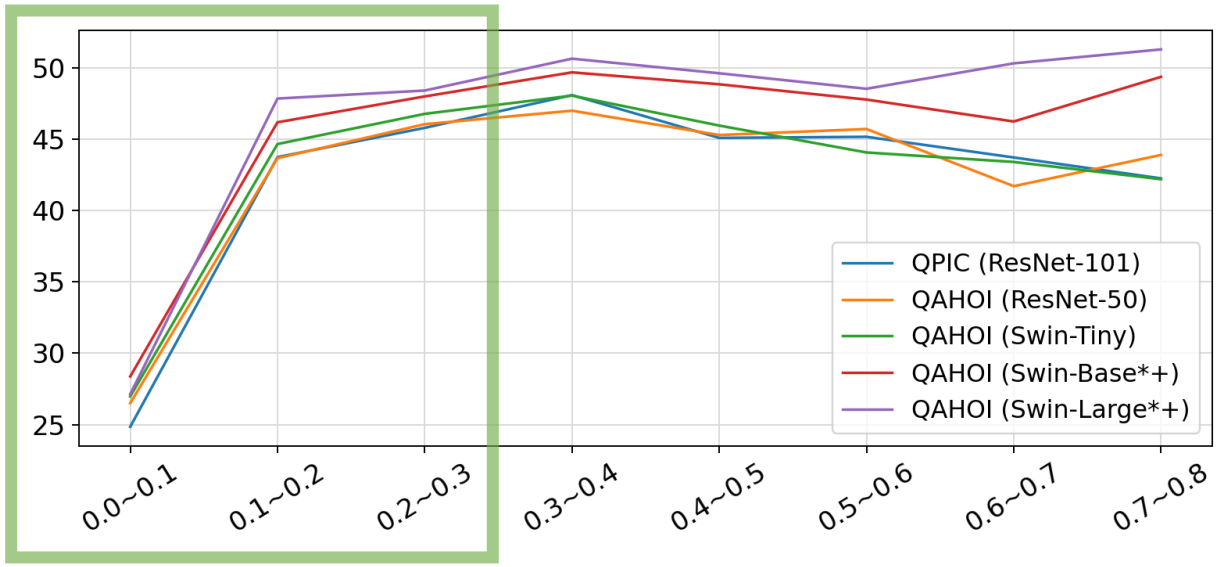
$x_4 \in \mathbb{R}^{C_d \times \frac{H}{64} \times \frac{W}{64}}$ is generated by using a 3×3 convolution on the last stage feature map x_3

Contribution at Different Spatial Scales

- The ground-truth HOI instances in the test set of HICO-DET is divided into 10 bins
- The bins with more than 1,000 instances are selected to display the AP results



(a) AP results on different large areas.

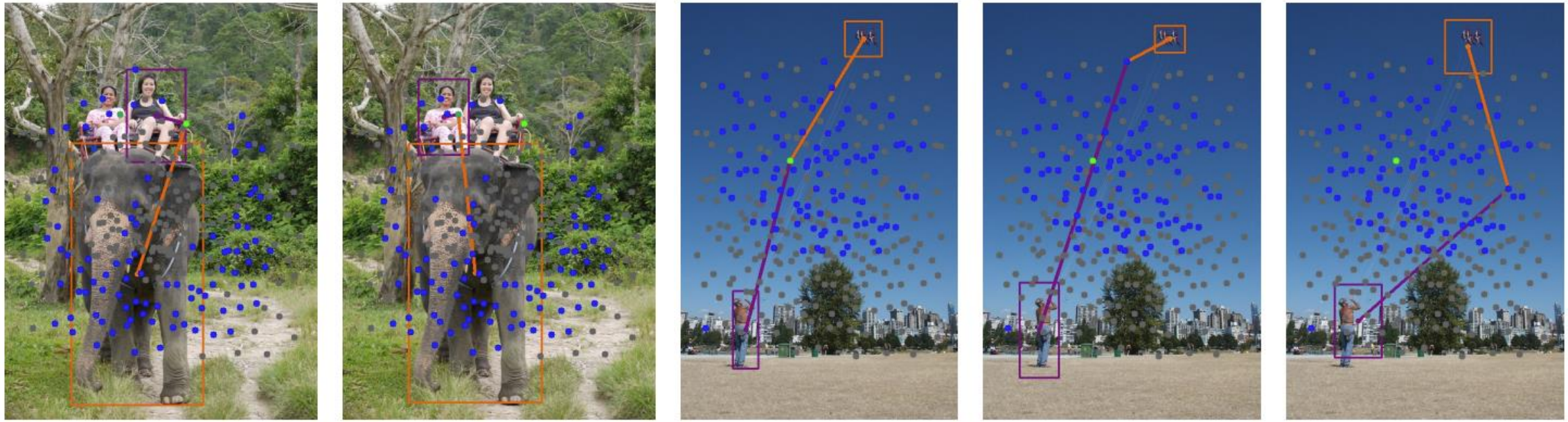


(b) AP results on different center distances.

Qualitative Analysis

The flexibility of Query-Based anchors

- Far from center
- Close to person or object



(a) ride, elephant (b) ride, elephant (c) fly, kite (d) fly, kite (e) fly, kite

The flexibility of the anchors.

Conclusion & Future Work

☐ Conclusion

- A multi-scale transformer-based method, QAHOI for HOI
- A new way to represent HOI instance based on query-based anchors
- Explore the benefits of transformer-based backbone

☐ Future Work

- Better detection framework
- Further reduce the training cost


UEC
TOKYO