HOI as Embeddings: Advancements of Model Representation Capability in Human-Object Interaction Detection

Junwen Chen Yingcheng Wang Yanai Keiji

The University of Electro-Communications
Tokyo

Human-Object Interaction (HOI) Detection

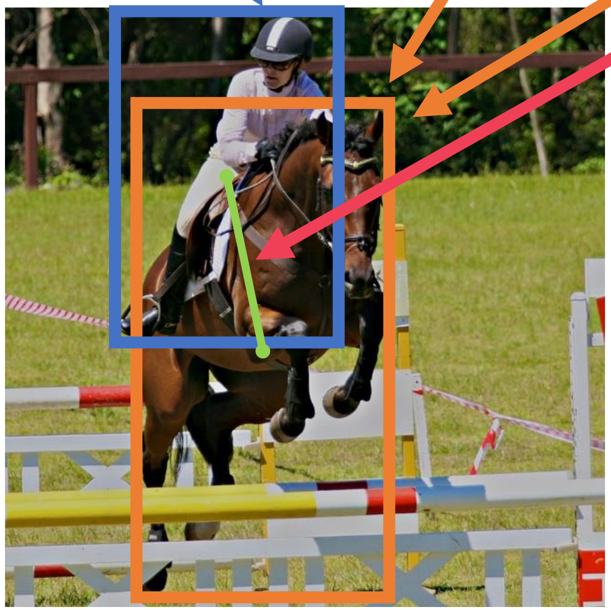
HOI Detection

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

HOI Instance

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

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



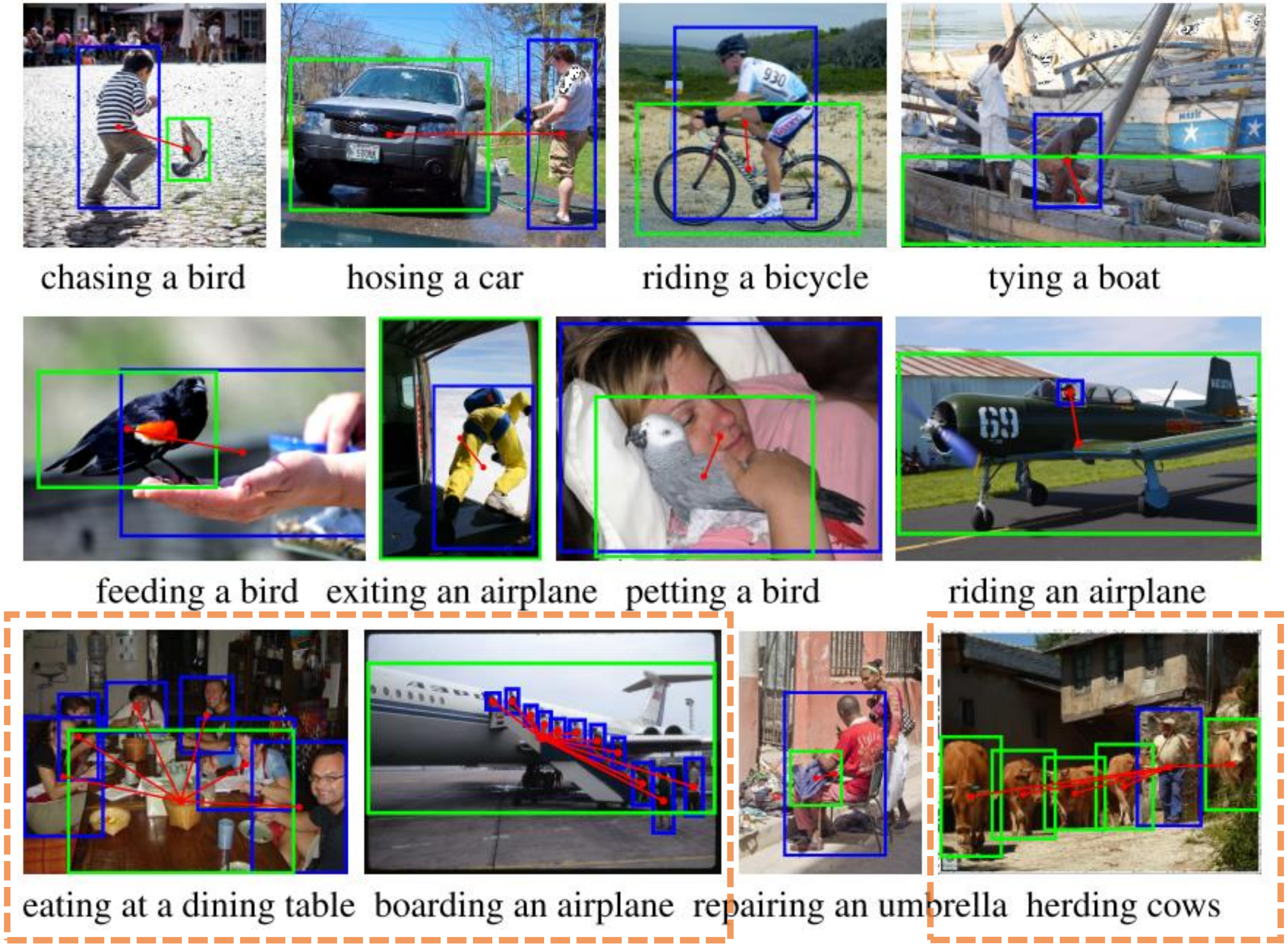
HICO-DET [1]

◻ 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

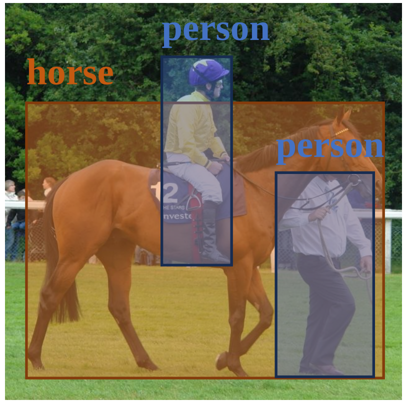
☐ Two-stage (Bottom-up)

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

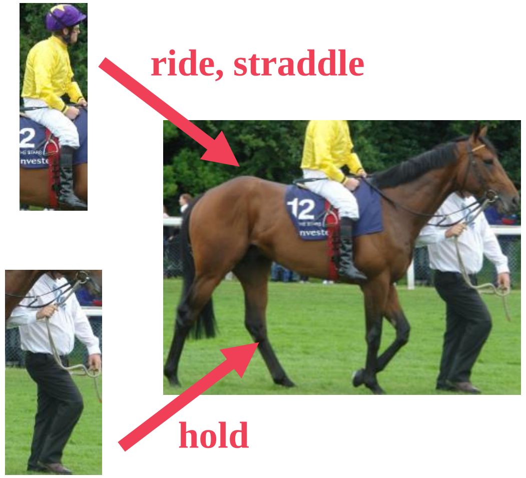
☐ One-stage (Top-down)

- Interaction Points & HOI Pair Matching

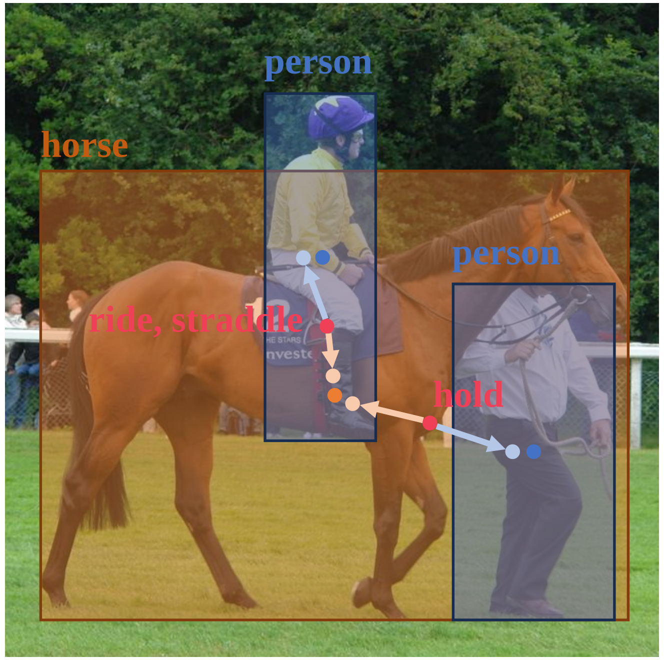
Detection



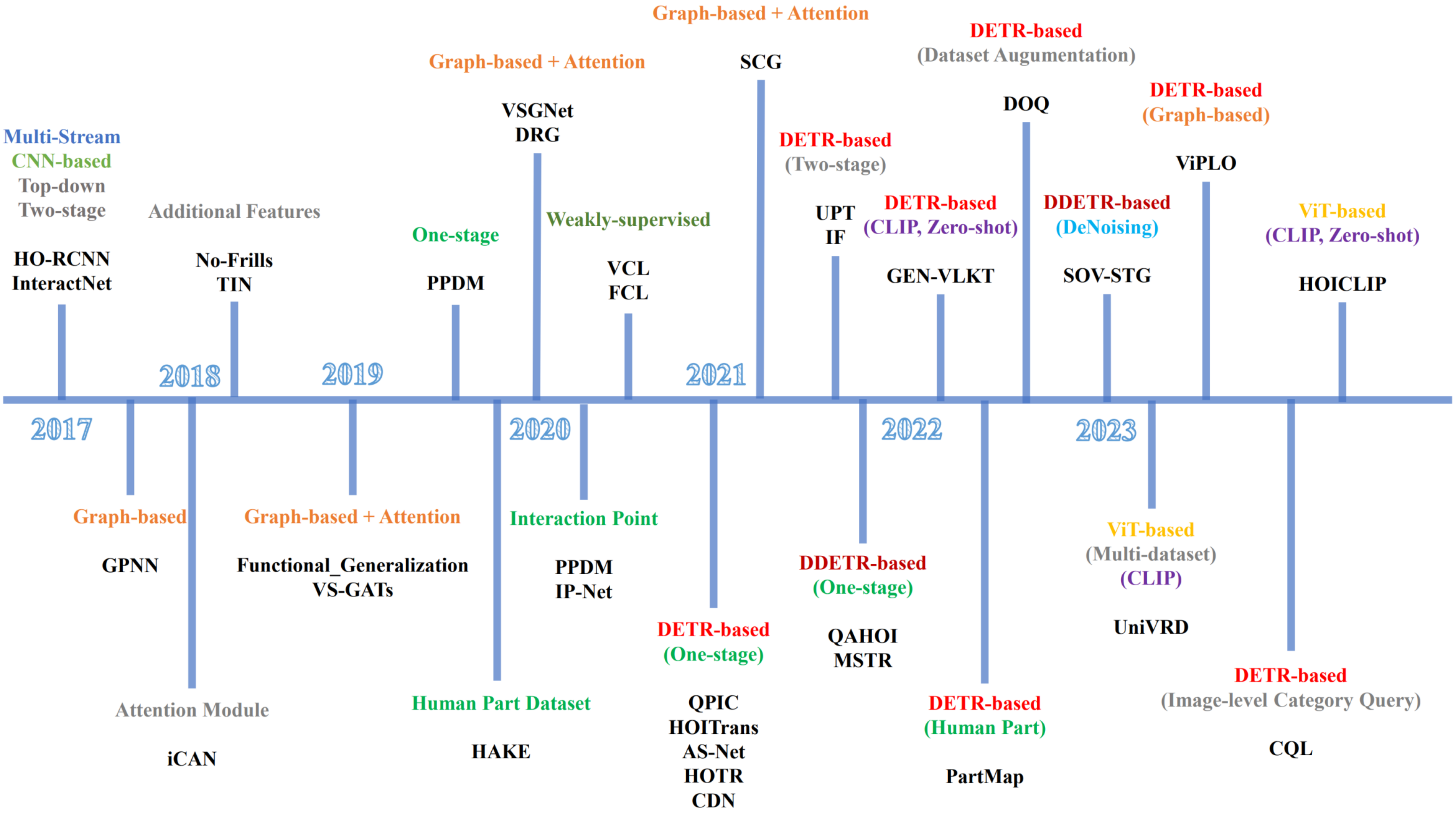
Recognition



Detection & Recognition



HOI Detection Timeline



Advancements of HOID

① QAHOI

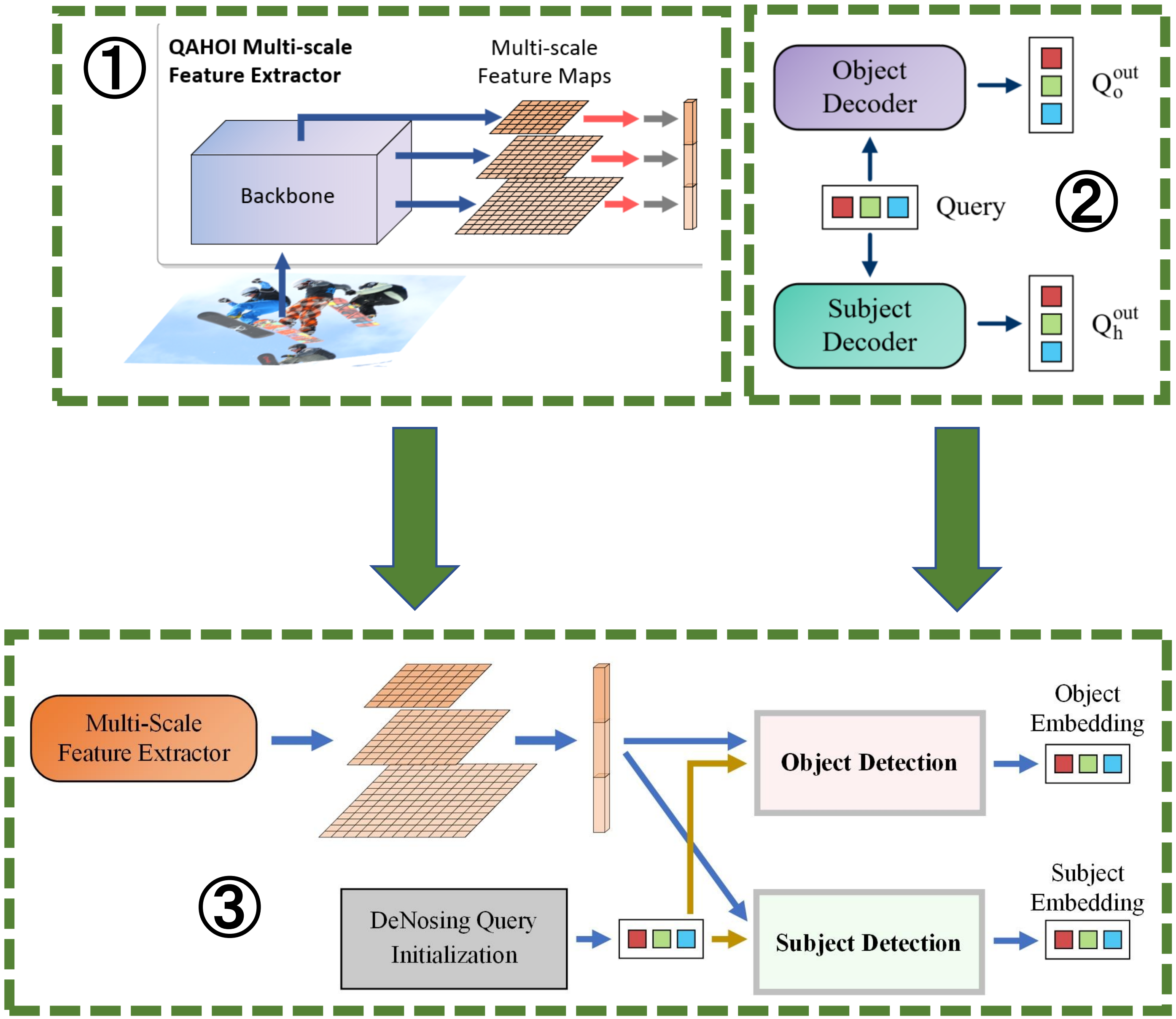
- Use multi-scale feature maps to utilize features at different scales

② PQNet

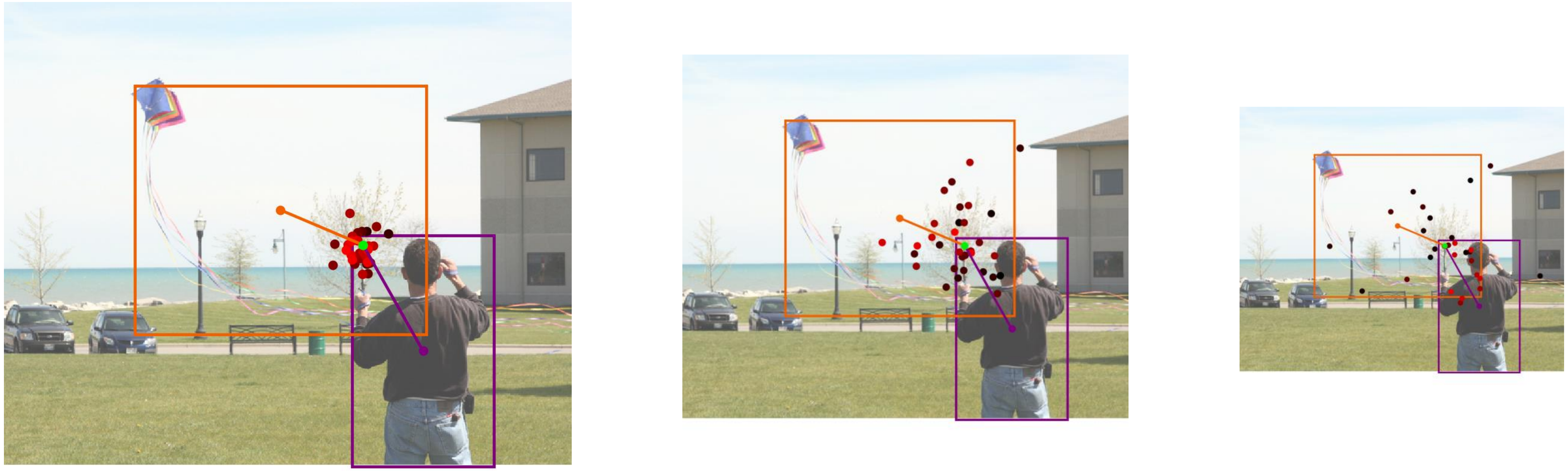
- Parallelize queries to speed up convergence

③ SOV-STG

- Combine the advantages of QAHOI and PQNet, and introduce denoising learning



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

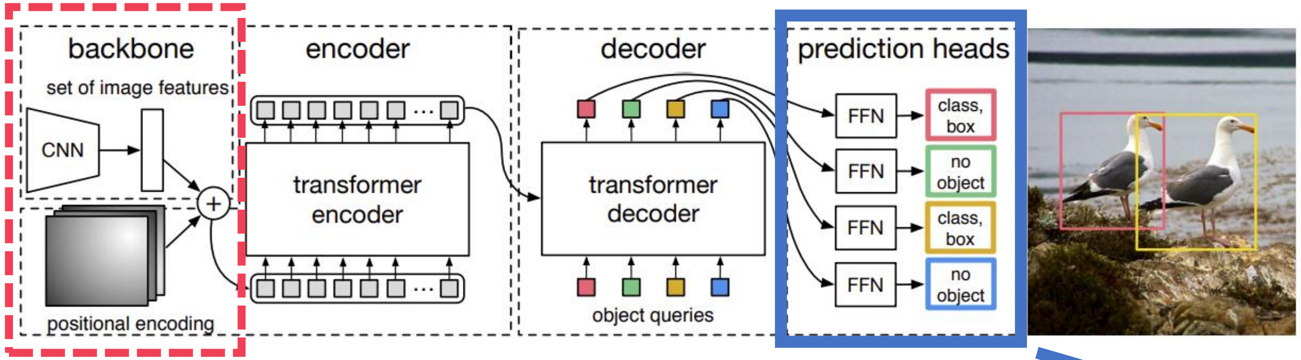


Chen, Junwen, and Keiji Yanai. "QAHOI: Query-Based Anchors for Human-Object Interaction Detection." *arXiv e-prints* (2021): arXiv-2112.

HOI Detection Approaches

◻ Transformer-based One-stage

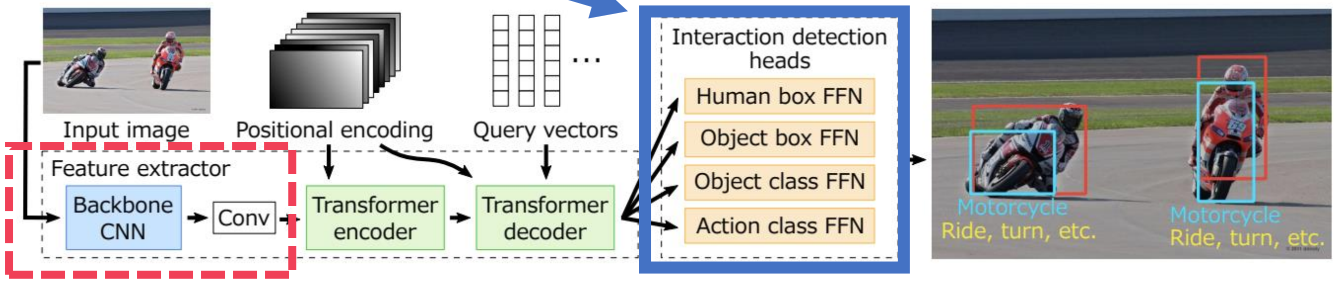
- Adapted from Transformer-based object detector DETR
- Set-based Prediction



DETR [3]

Low-resolution feature maps

QPIC [2]



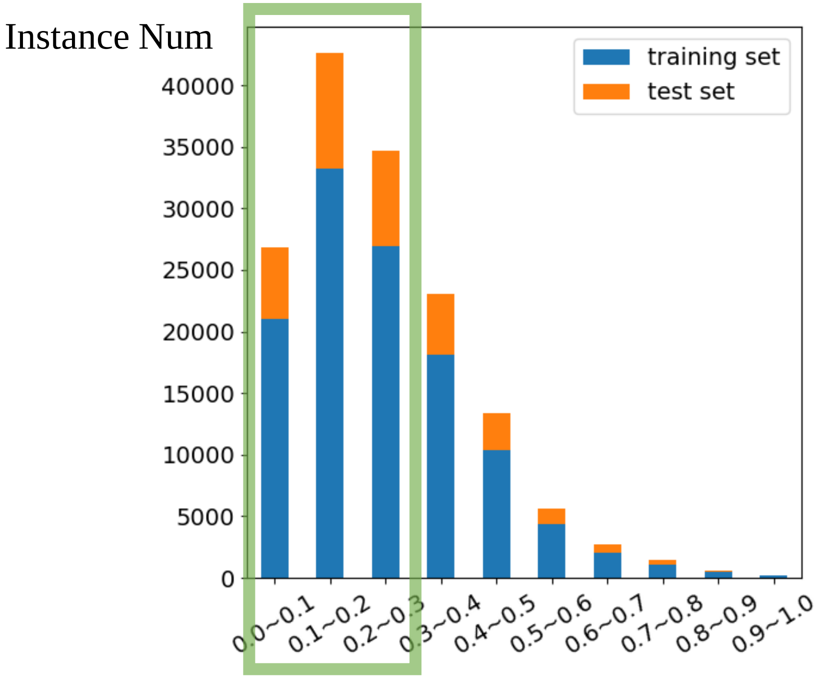
[2] Tamura, Masato, Hiroki Ohashi, and Tomoaki Yoshinaga. "QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[3] Carion, Nicolas, et al. "End-to-end object detection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

Motivation

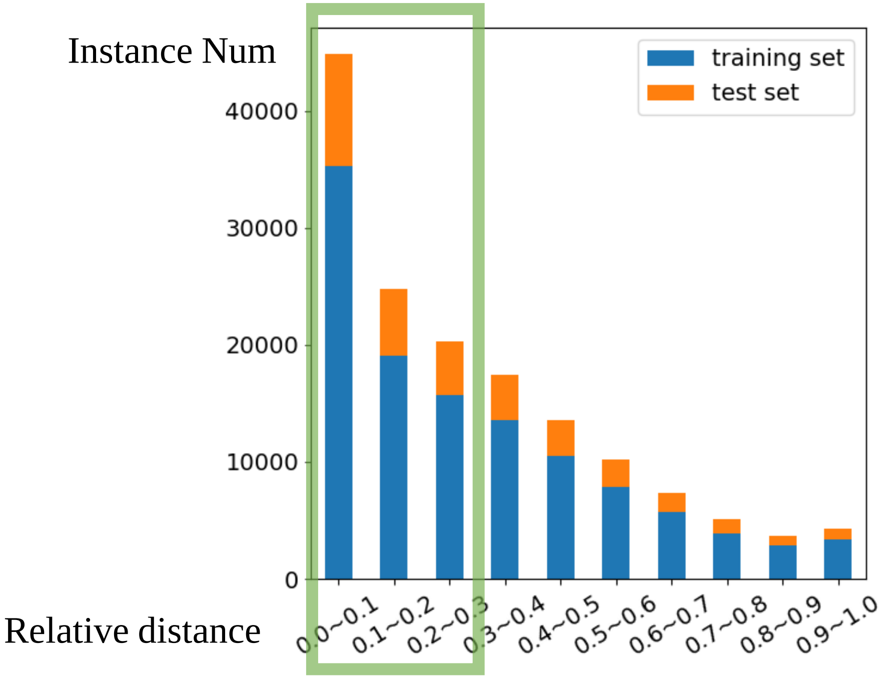
◻ The spatial distribution of the HOI instances in HICO-DET

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

◻ Transformer-based methods lack a multi-scale architecture



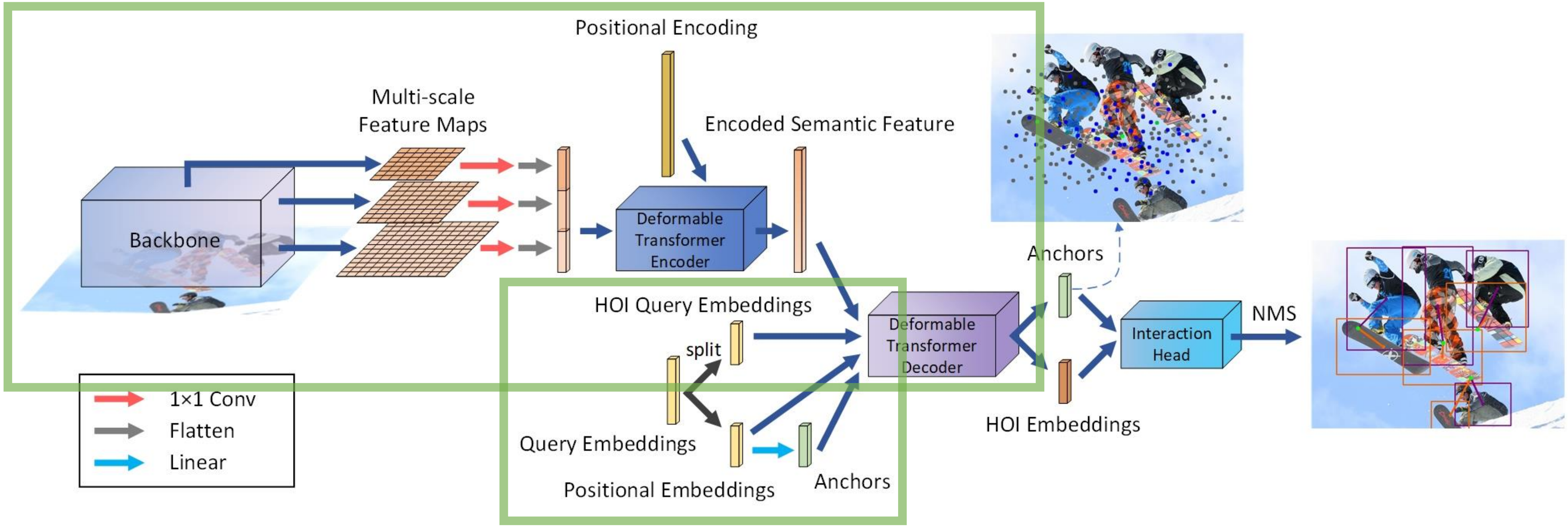
(a) Larger Area



(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.

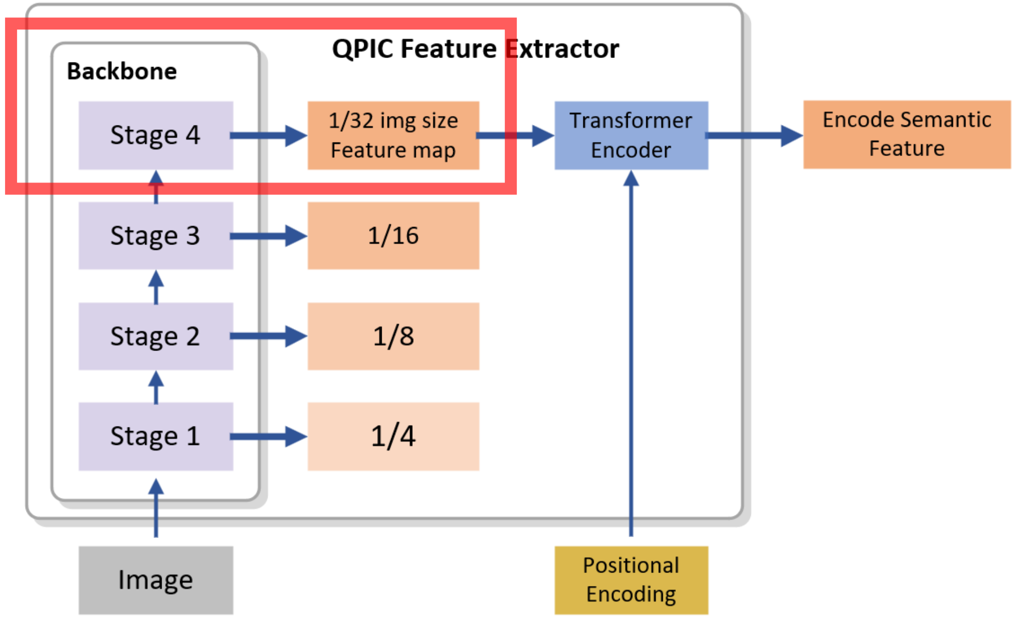
QAHOI: Multi-scale Feature Extractor

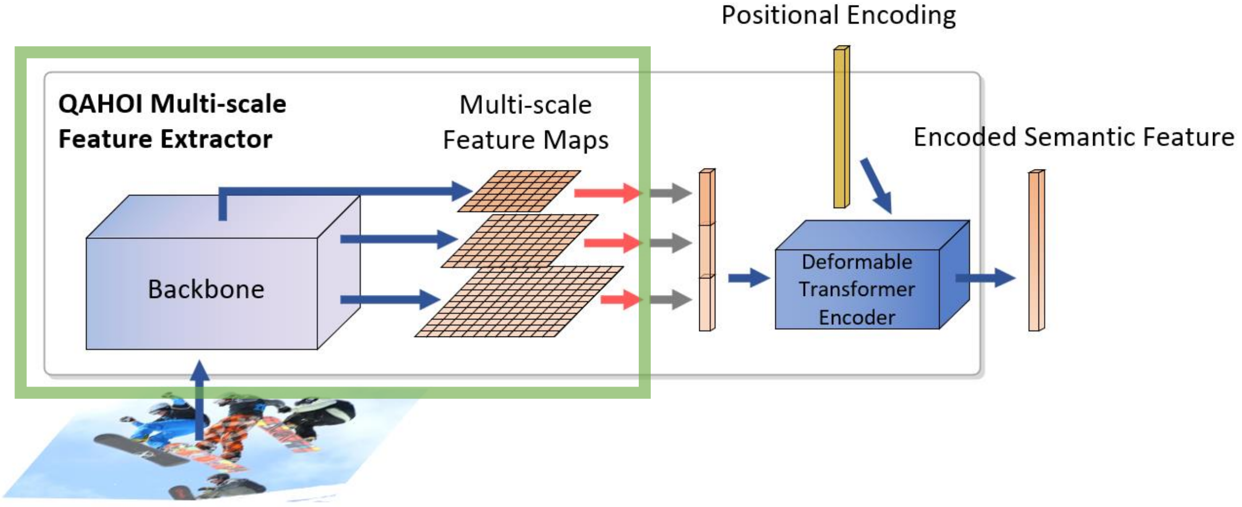
□ Feature Extractor of QPIC

- 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.

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 *Full*	Default *Rare*	Default *Non-Rare*	Known Object *Full*	Known Object *Rare*	Known Object *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**

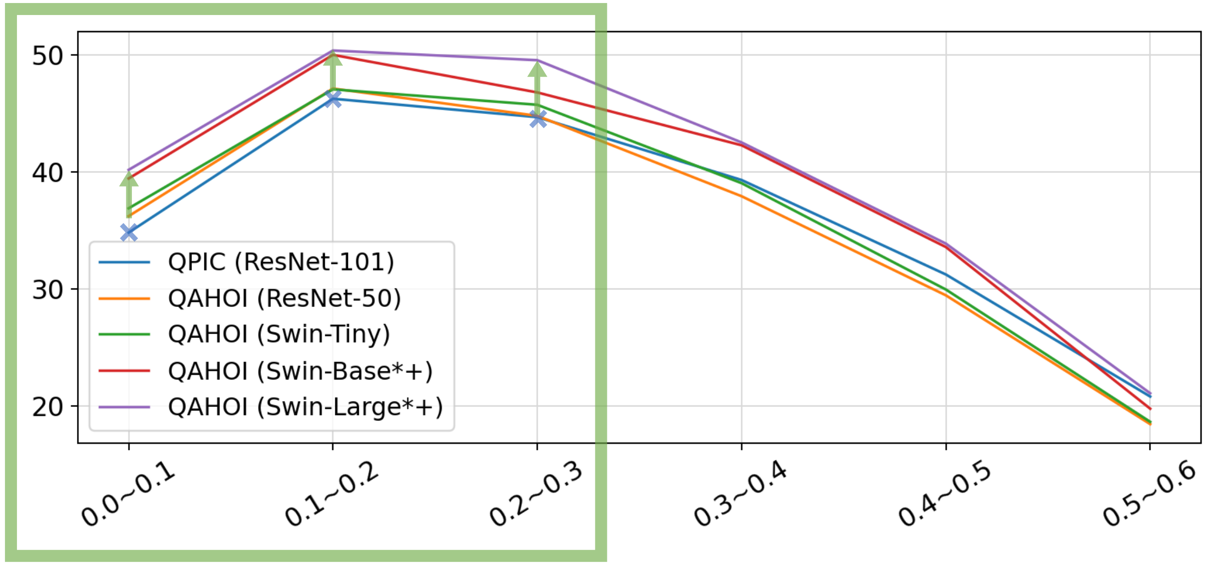
+5.88
(19.7%)

+4.1
(13.9%)

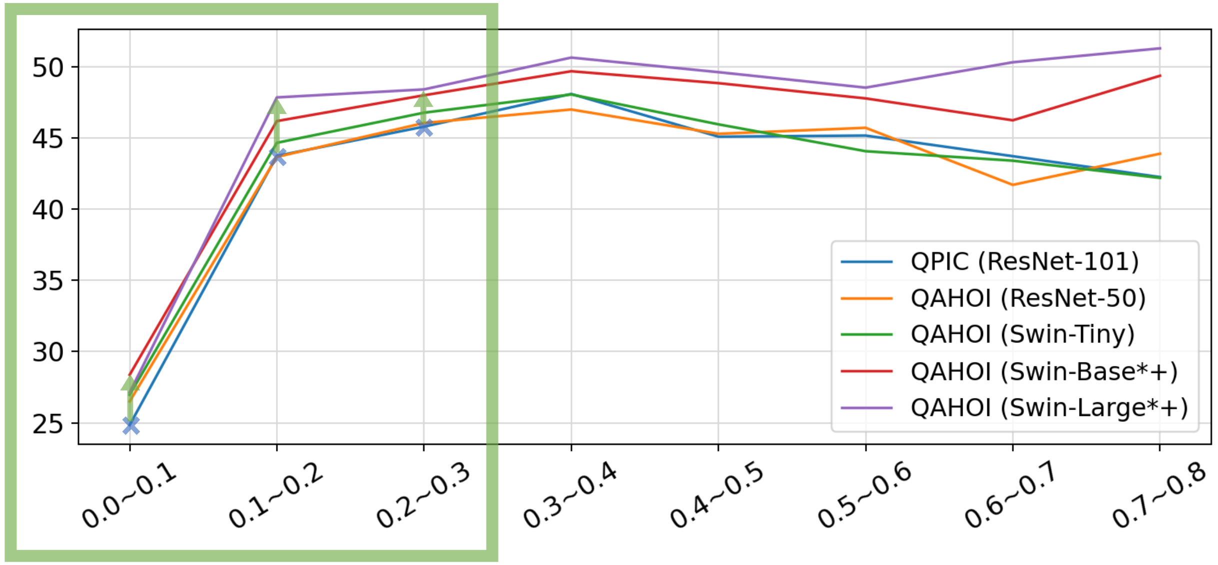
[6] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

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



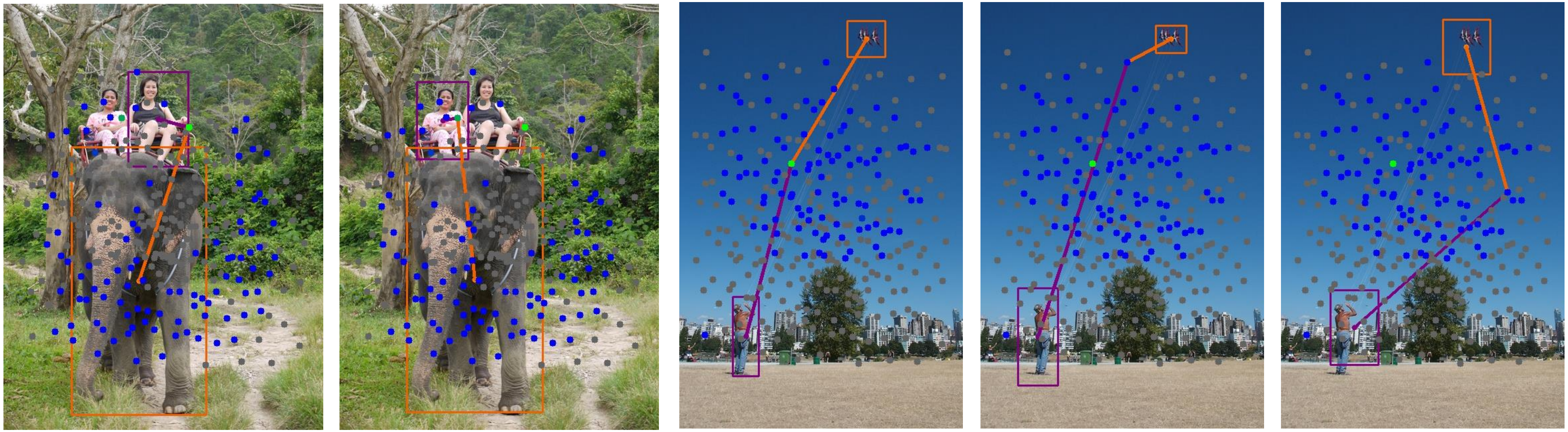
(a) AP results on different large areas.



(b) AP results on different center distances.

The flexibility of Query-Based anchors

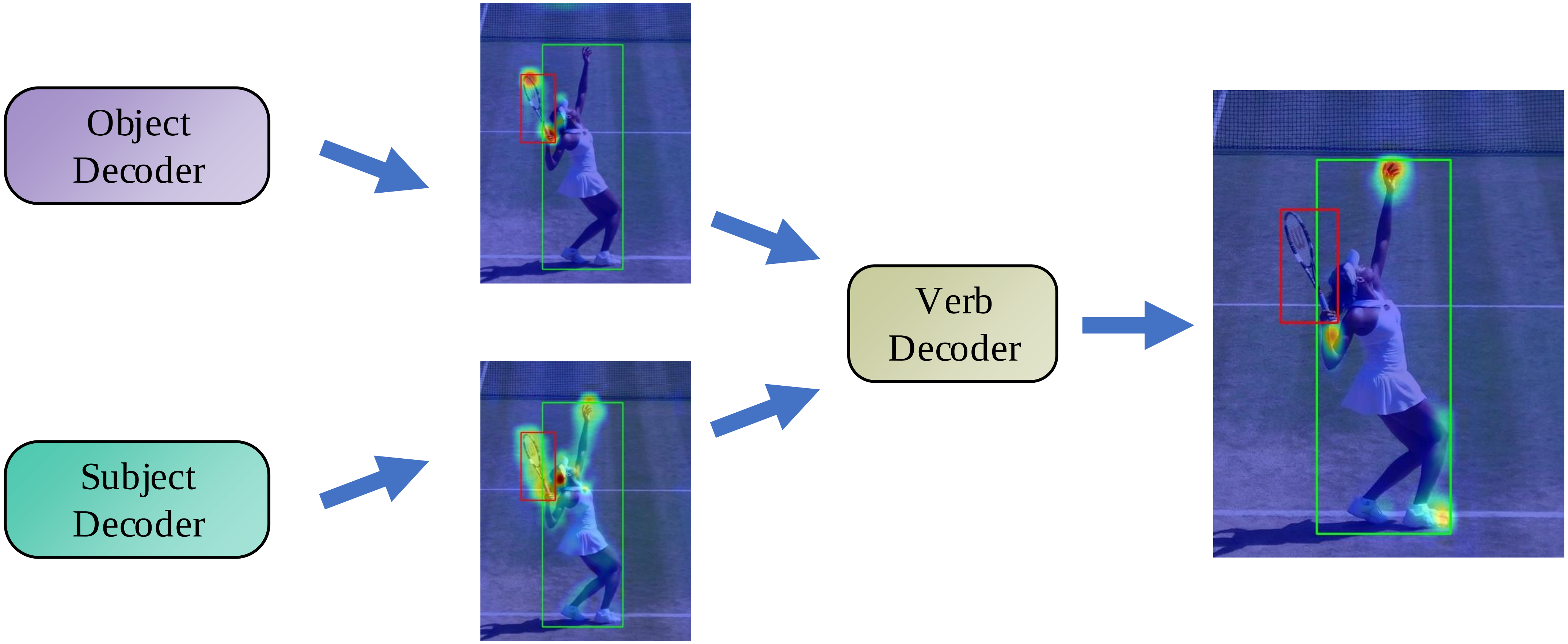
- Far from center
- Close to person or object



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

The flexibility of the anchors. Anchors Top100 Anchors Highest Score Anchor

Parallel Queries for Human-Object Interaction Detection

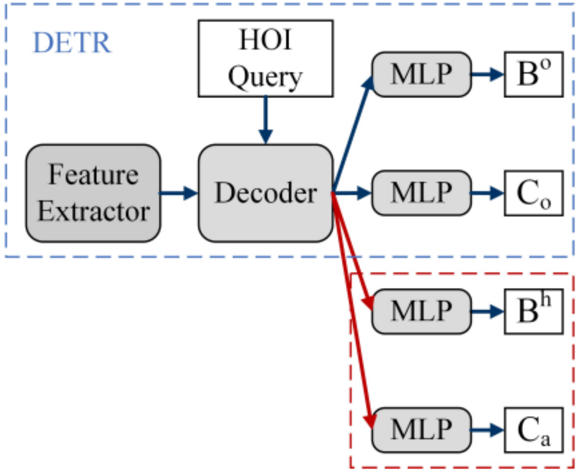


Motivation: More Accuracy and Faster Convergence

□ Problems of the previous methods

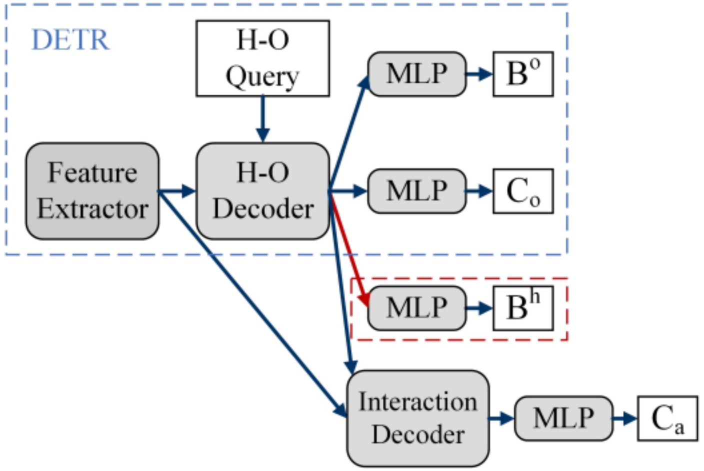
- Transformer-based one-stage methods
  - DETR [Carion et al. ECCV2020] is applied to the HOI task
  - The decoding target of DETR is changed

All of the elements are predicted by the same decoder



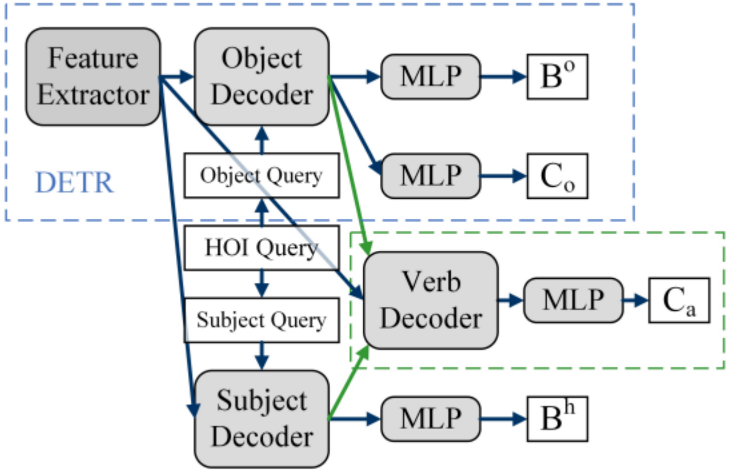
QPIC [Tamura et al. CVPR2021]

Human and object prediction are tangled in the H-O decoder



CDN [Zhang et al. NIPS2021]

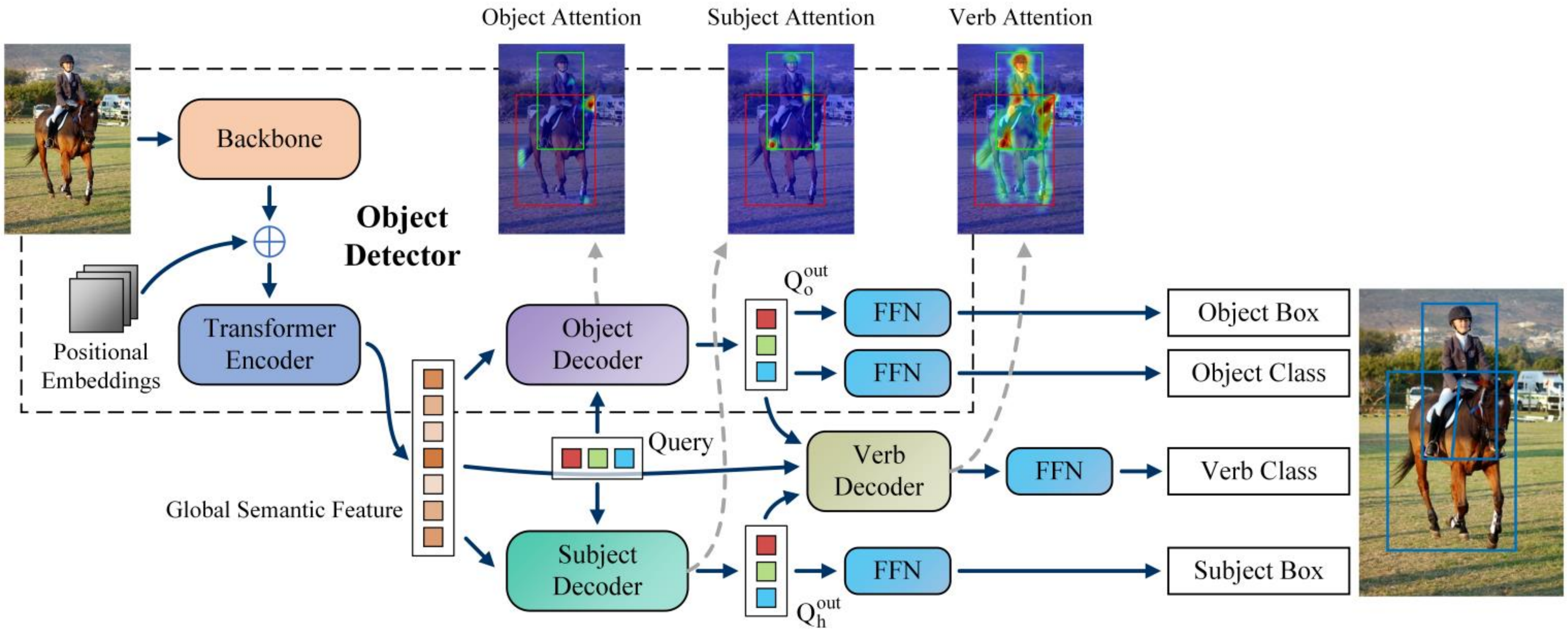
- Human prediction is disentangled
- Maintaining the targets of the object detector



Proposed method: PQNet

Parallel Queries for Human-Object Interaction Detection

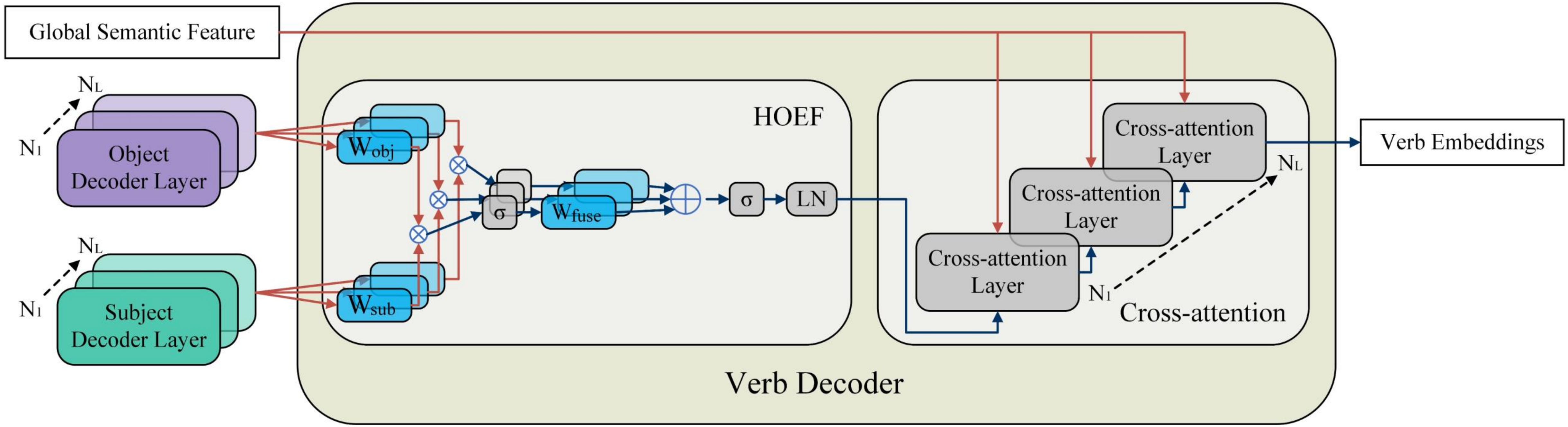
◻ Overview



- Parallel queries are used to split the detecting process
- The verb decoder focuses on extracting the verb representations

Verb Decoder

◻ Human-object Embedding Fusion



- Two kinds of attention mechanisms
  - The HOEF module is used to form the verb embedding
  - The cross-attention module is used to extract verb information from the context

Experiments

◻ Compare with current state-of-the-art (SOTA) methods

Method	Fine-tuned Detector	Backbone	Feature	Default *Full*	Default *Rare*	Default *Non-Rare*	Known Object *Full*	Known Object *Rare*	Known Object *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**

+3.06
(10.5%)

+0.35
(1.1%)

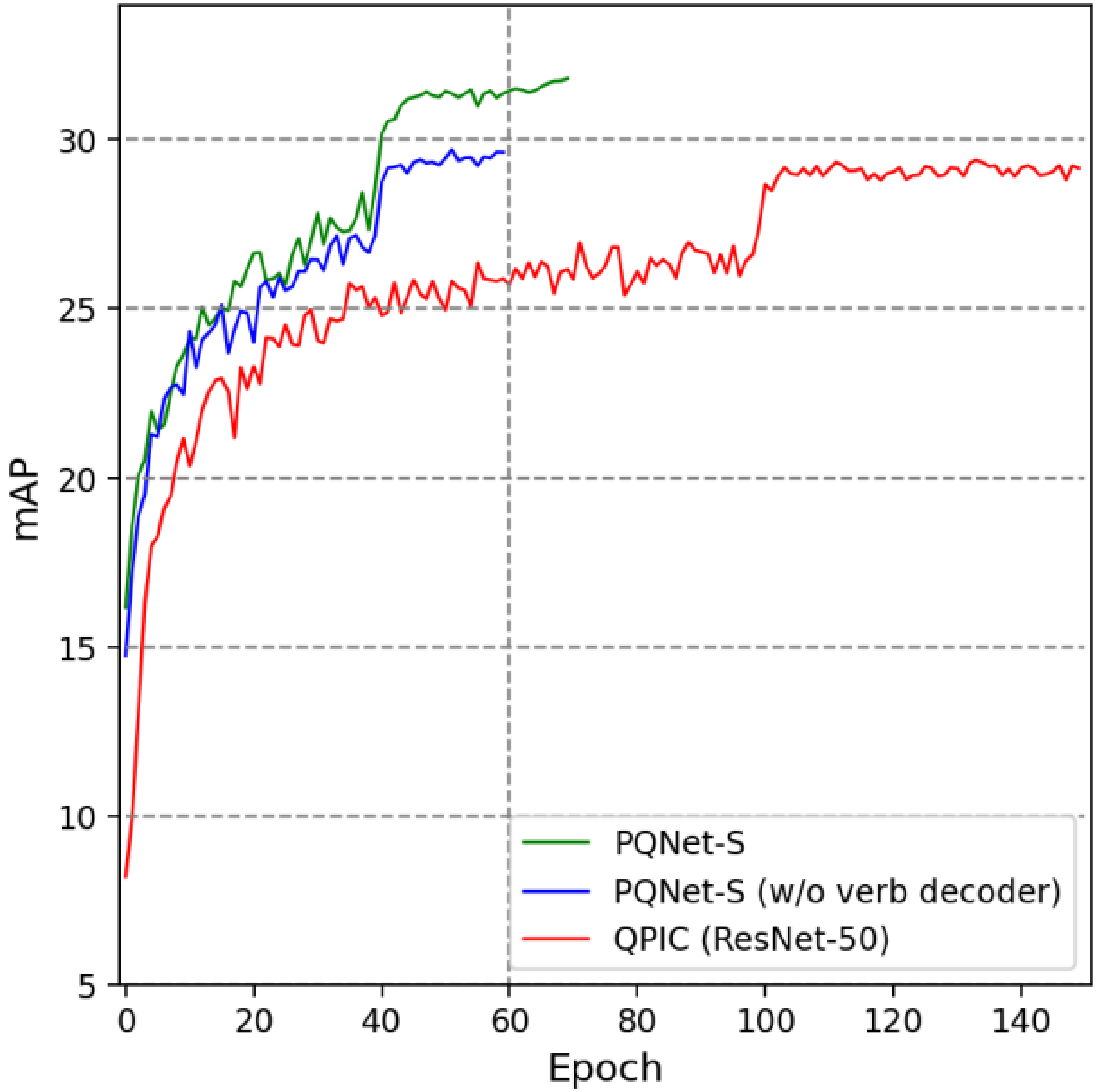
+0.80
(2.5%)

Experiments

☐ The training convergence

- Parallel queries & decoders
  - **Improve the model's performance**
  - **Accelerate the convergence**

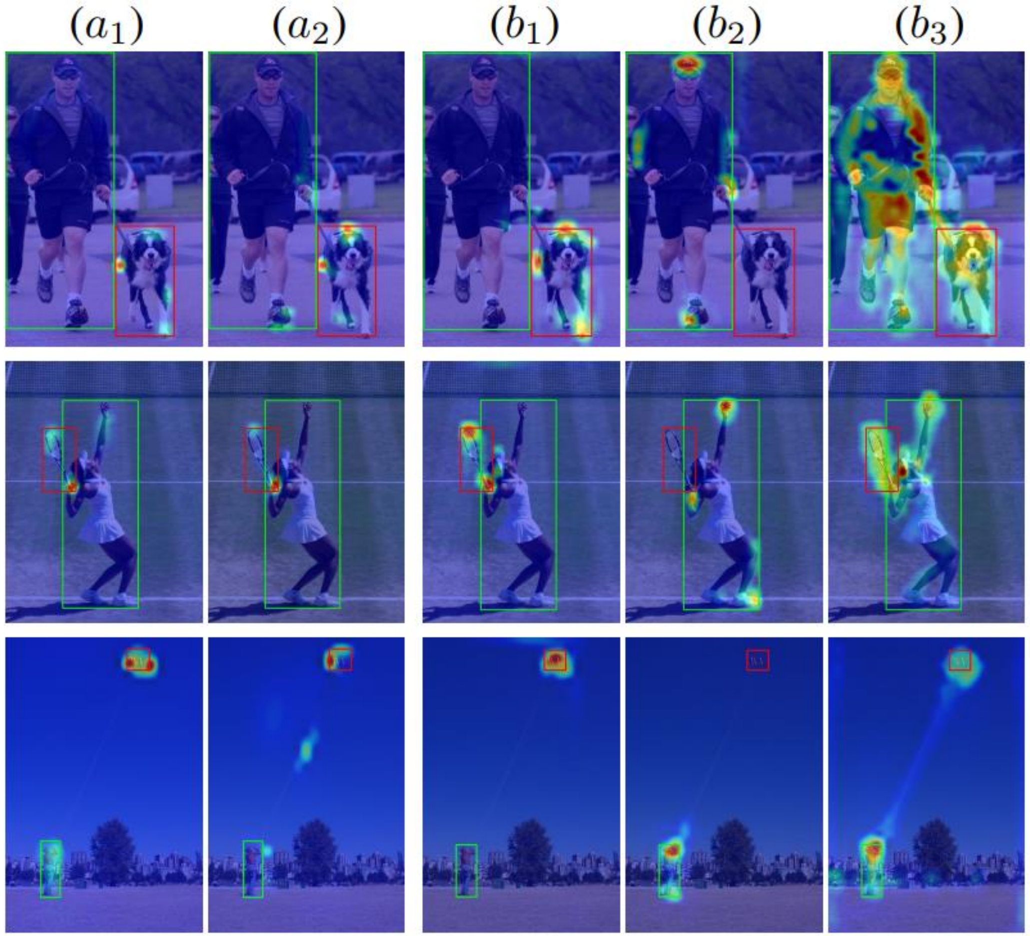
- Compare to previous SOTA
  - 2× mAP at the first epoch
  - Fast convergence in the first 40 epochs



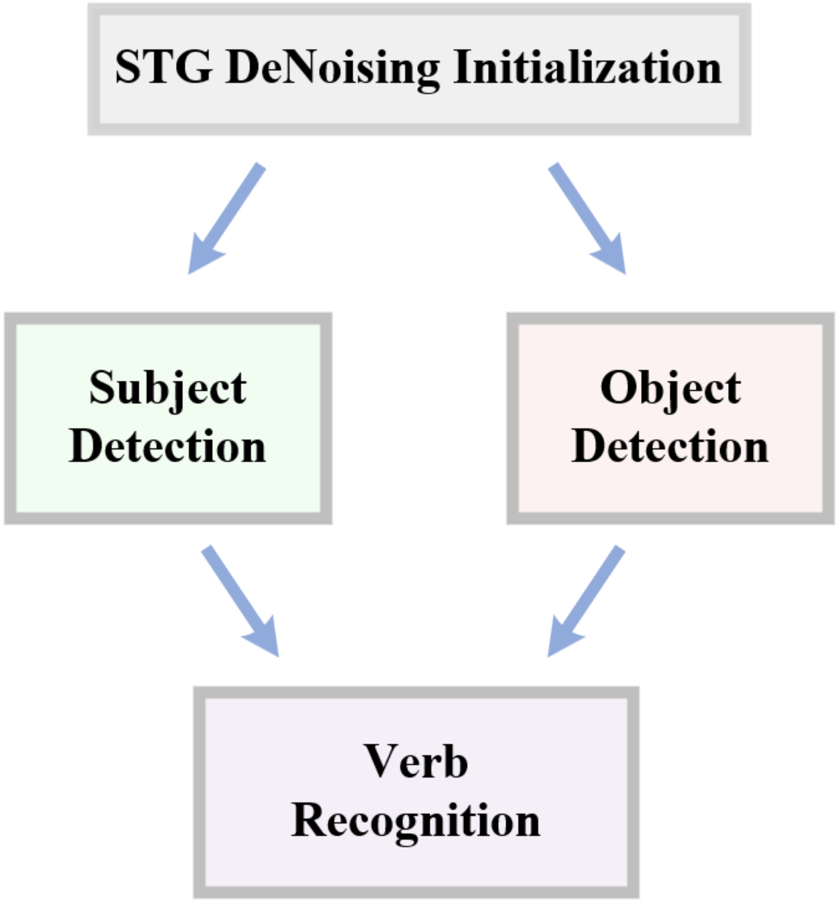
Qualitative Analysis

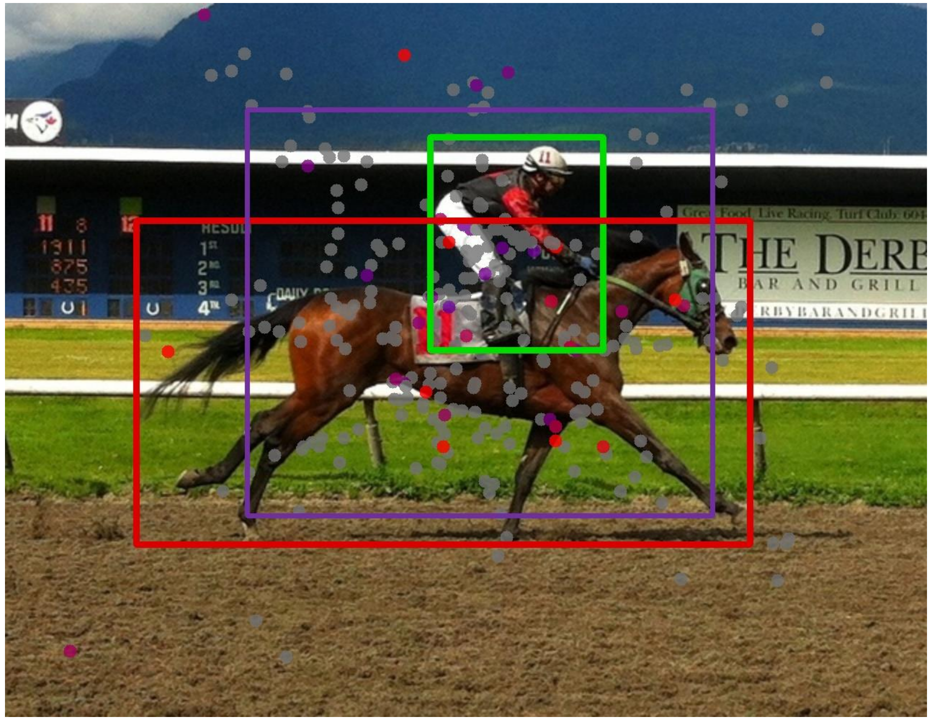
◻ The visualization of attention maps

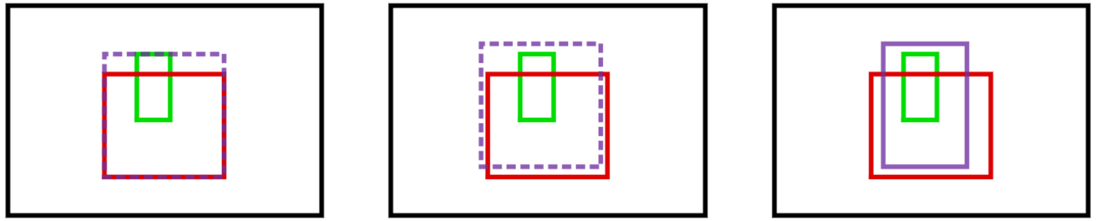
- CDN concentrate on **the object more than the human**
- PQNet learned to focus on **the extreme points of the target**
  - The verb decoder focuses on the **whole part of the human and object** but **pays more attention to the interaction regions**



Subject Object Verb (SOV) Decoders with Specific Target Guided (STG)





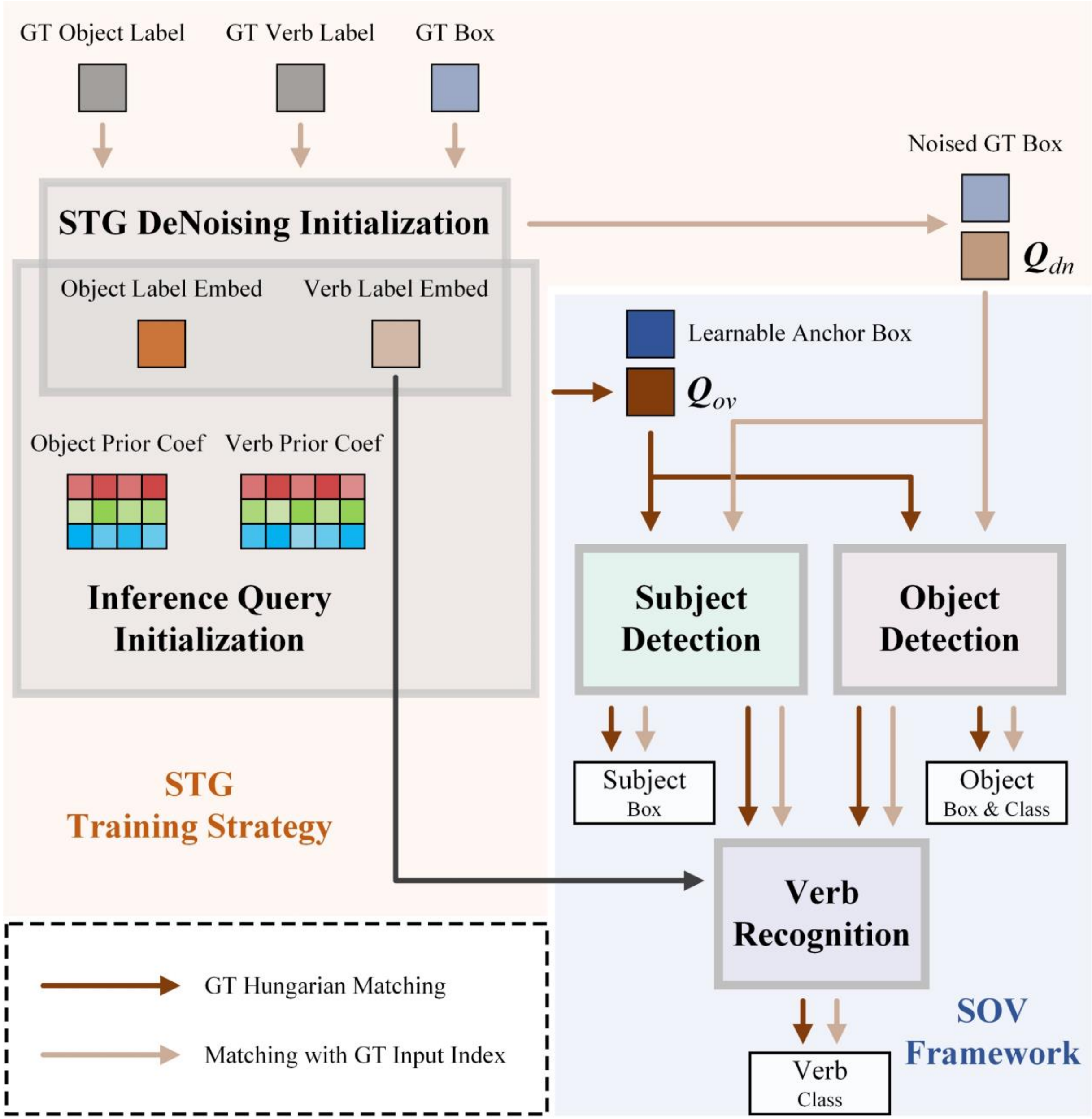


MBR   Shifted MBR   Adaptive Shifted MBR

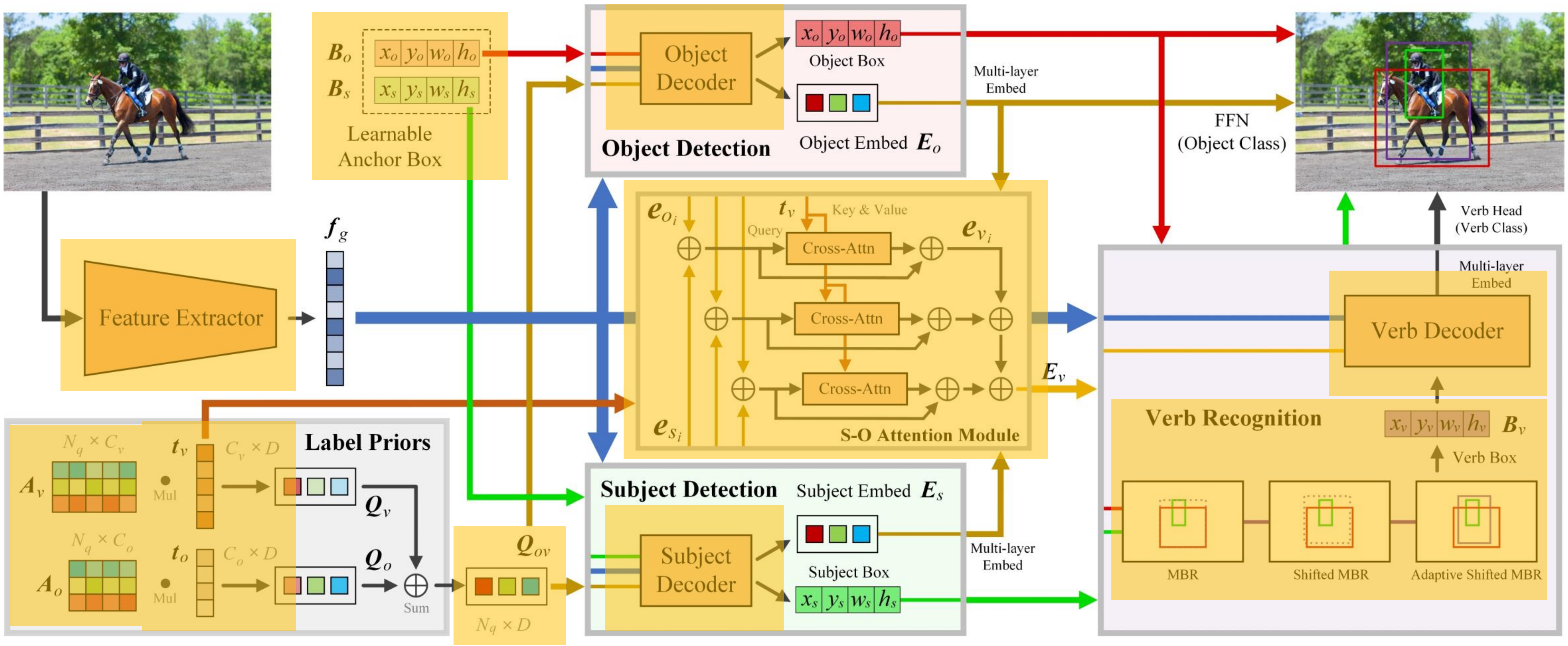
SOV-STG: Focusing on what to decode and what to train

☐ End-to-end training pipeline

- SOV framework splits the decoding process into three parts
- STG training strategy efficiently transfers the ground-truth information



SOV-STG: Overview



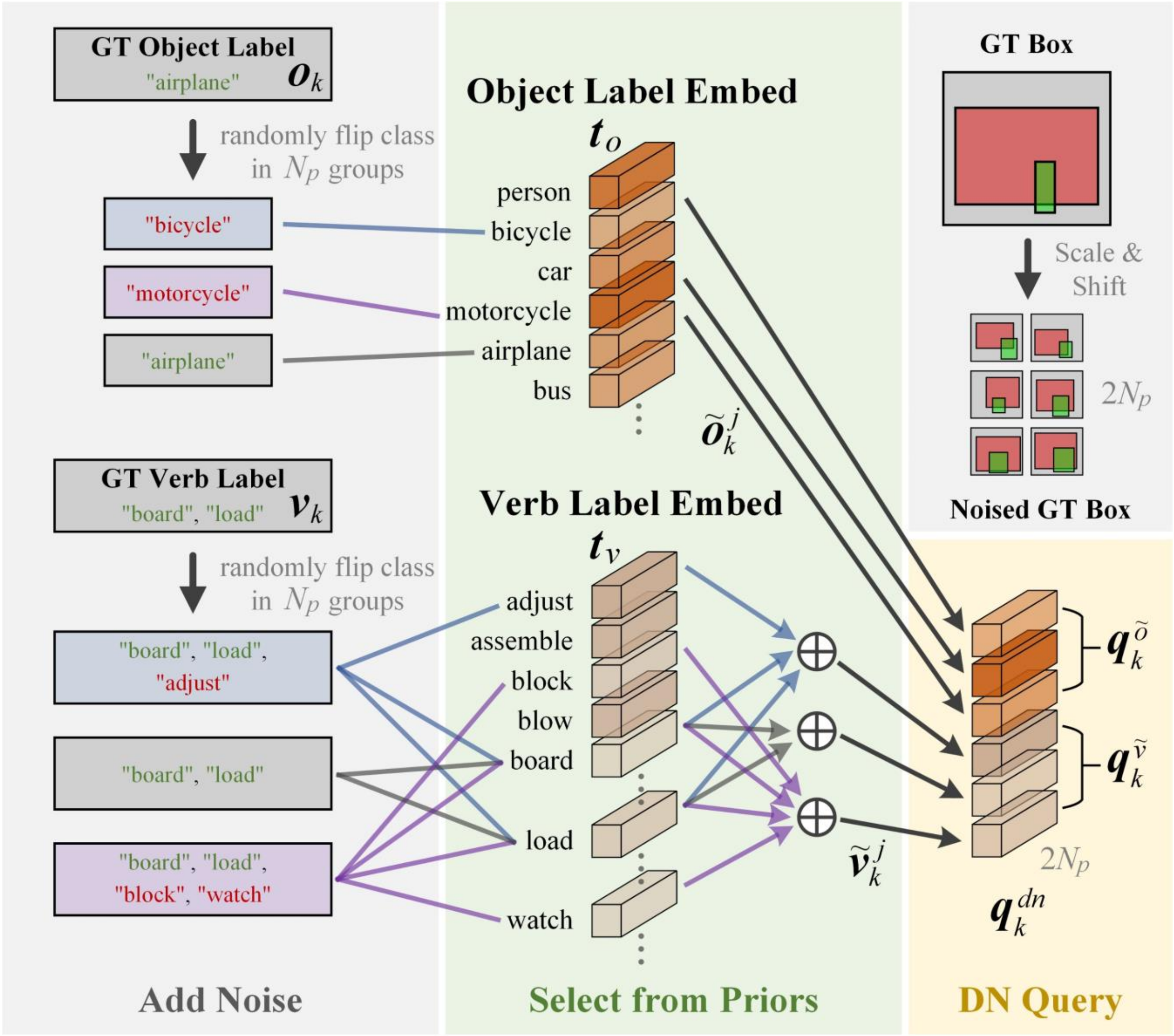
- The position information is separated from the context query
- Multi-scale feature extractor and SOV decoders
- Learnable anchor boxes and label embeddings provide prior knowledge for inference and noise removal learning

SOV-STG: Split Target Guided DeNoising

☐ DN Query

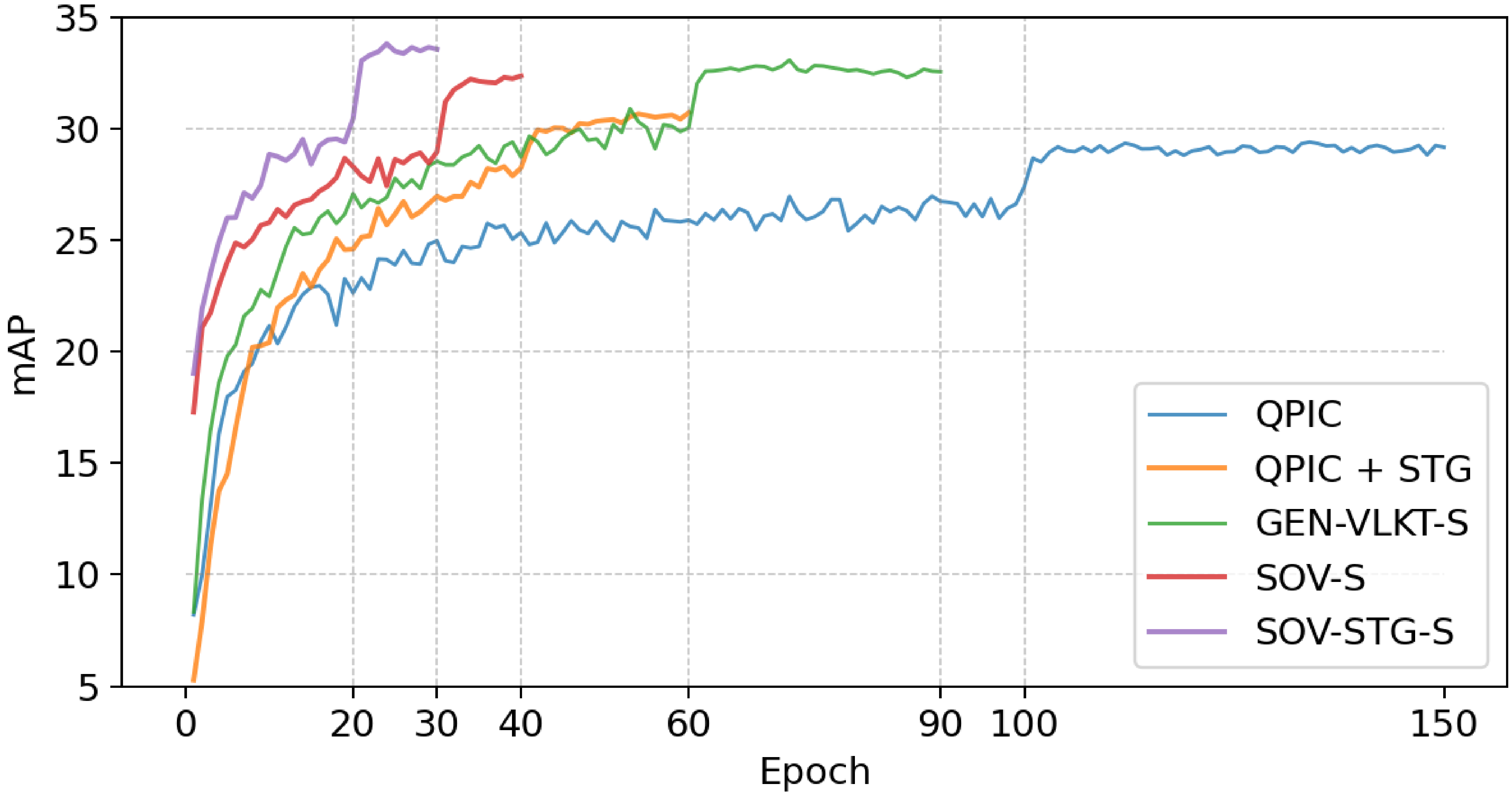
- Two part initialization
  - Object Label DN Query $q_k^{\tilde{o}}$
  - Verb Label DN Query $q_k^{\tilde{v}}$

- Label Priors
  - Learnable Label Embeddings both used in training and inference



Experiments

◻ The training convergence



Experiments

☐ Compare with current state-of-the-art (SOTA) methods

Method	Epoch	Backbone	Default			Known Object		
			Full	*Rare*	*Non-Rare*	*Full*	*Rare*	*Non-Rare*
QPIC [17]	150	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
CDN-S [28]	100	ResNet-50	31.44	27.39	32.64	34.09	29.63	35.42
CDN-B [28]	100	ResNet-50	31.78	27.55	33.05	34.53	29.73	35.96
CDN-L [28]	100	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38
PQNet-S [26]	70	ResNet-50	31.92	28.06	33.08	34.58	30.71	35.74
PQNet-B [26]	100	ResNet-50	32.13	29.43	32.93	34.68	32.06	35.47
PQNet-L [26]	100	ResNet-50	32.45	27.80	33.84	35.28	30.72	36.64
HQM (CDN-S) [35]	80	ResNet-50	32.47	28.15	33.76	35.17	30.73	36.50
RLIP-ParSe [38]	90	ResNet-50	32.84	34.63	26.85	-	-	-
MUREN [39]	100	ResNet-50	32.87	28.67	34.12	35.52	30.88	36.91
DOQ (CDN-S) [34]	80	ResNet-50	33.28	29.19	34.50	-	-	-
GEN-VLKT-S [32]	90	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
HOICLIP [40]	90	ResNet-50	34.69	31.12	35.74	37.61	34.47	38.54
GEN-VLKT-M [32]	90	ResNet-101	34.78	31.50	35.77	38.07	34.94	39.01
GEN-VLKT-L [32]	90	ResNet-101	34.95	31.18	36.08	38.22	34.36	39.37
QAHOI-Swin-L [25]	150	Swin-Large-22K	35.78	29.80	37.56	37.59	31.36	39.36
FGAHOI-Swin-L [41]	190	Swin-Large-22K	37.18	30.71	39.11	38.93	31.93	41.02
DiffHOI-Swin-L [42]	90	Swin-Large-22K	41.50	39.96	41.96	43.62	41.41	44.28
SOV-STG-S	30	ResNet-50	33.80	29.28	35.15	36.22	30.99	37.78
SOV-STG-M	30	ResNet-101	34.87	30.41	36.20	37.35	32.46	38.81
SOV-STG-L	30	ResNet-101	35.01	30.63	36.32	37.60	32.77	39.05
SOV-STG-Swin-L	30	Swin-Large-22K	43.35	42.25	43.69	45.53	43.62	46.11

1/3 epoch

+4.45%

Conclusion

◻ Summary

- A multi-scale transformer-based method, QAHOI for HOI.
- A novel transformer-based one-stage method for HOI detection with parallel queries.
- A new way to represent HOI instances based on query-based anchors

◻ Future Work

- Fast and Powerful
- Improved Prior Knowledge

UEC