Enhanced Training of Query-Based Object Detection via Selective Query Recollection

Fangyi Chen[1] Han Zhang[1] Kai Hu[1] Yu-Kai Huang[1] Chenchen Zhu[2] Marios Savvides[1]
Carnegie Mellon University[1] Meta AI[2]
{fangyic,hanz3,kaihu,yukaih2,marioss}@andrew.cmu.edu chenchenz@fb.com

Kunyang Han

Motivation

- Common issue of query-based object detectors (DETR)
- Limitation
  - lack of training emphasis
  - cascading errors from decoding sequence

DETR

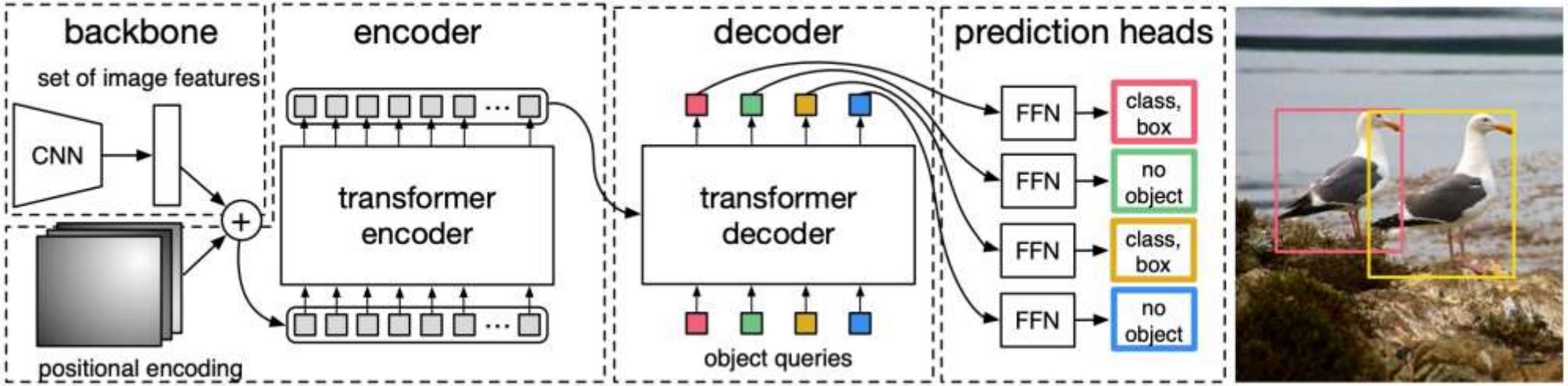


Fig. 2: DETR uses a conventional CNN backbone to learn a 2D representation of an input image. The model flattens it and supplements it with a positional encoding before passing it into a transformer encoder. A transformer decoder then takes as input a small fixed number of learned positional embeddings, which we call *object queries*, and additionally attends to the encoder output. We pass each output embedding of the decoder to a shared feed forward network (FFN) that predicts either a detection (class and bounding box) or a "`no object`" class.

$$\frac{\mathrm{TP}}{TP + FP} = 0.5 \qquad \frac{\mathrm{TP}}{TP + FN} = 0.25$$

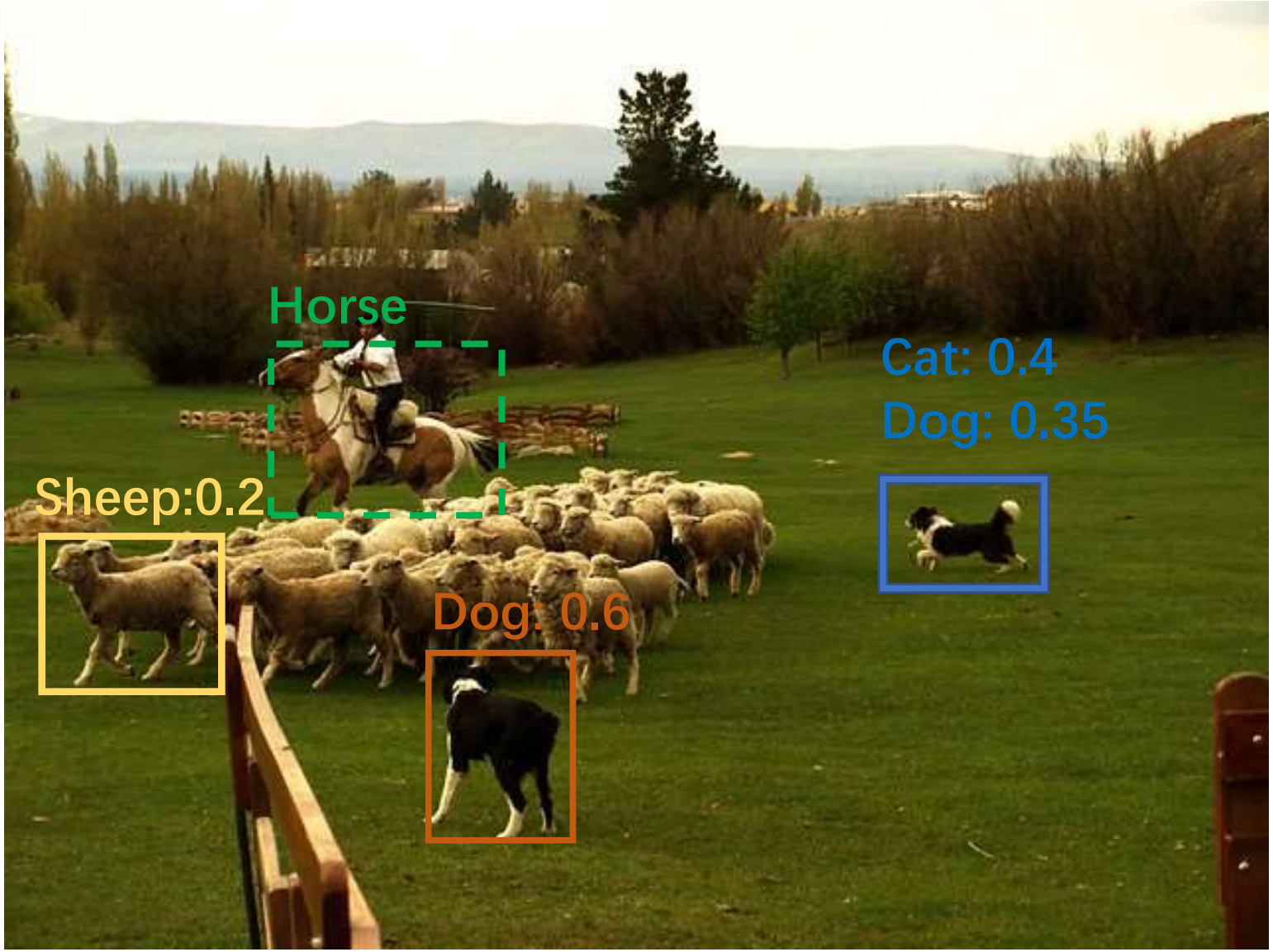
Motivation

Consist with GT

Prediction

FP	TP
FN	TN

Threshold = 0.3



$$\frac{\mathrm{TP}}{TP + FP} = 0.67 \qquad \frac{\mathrm{TP}}{TP + FN} = 0.5$$

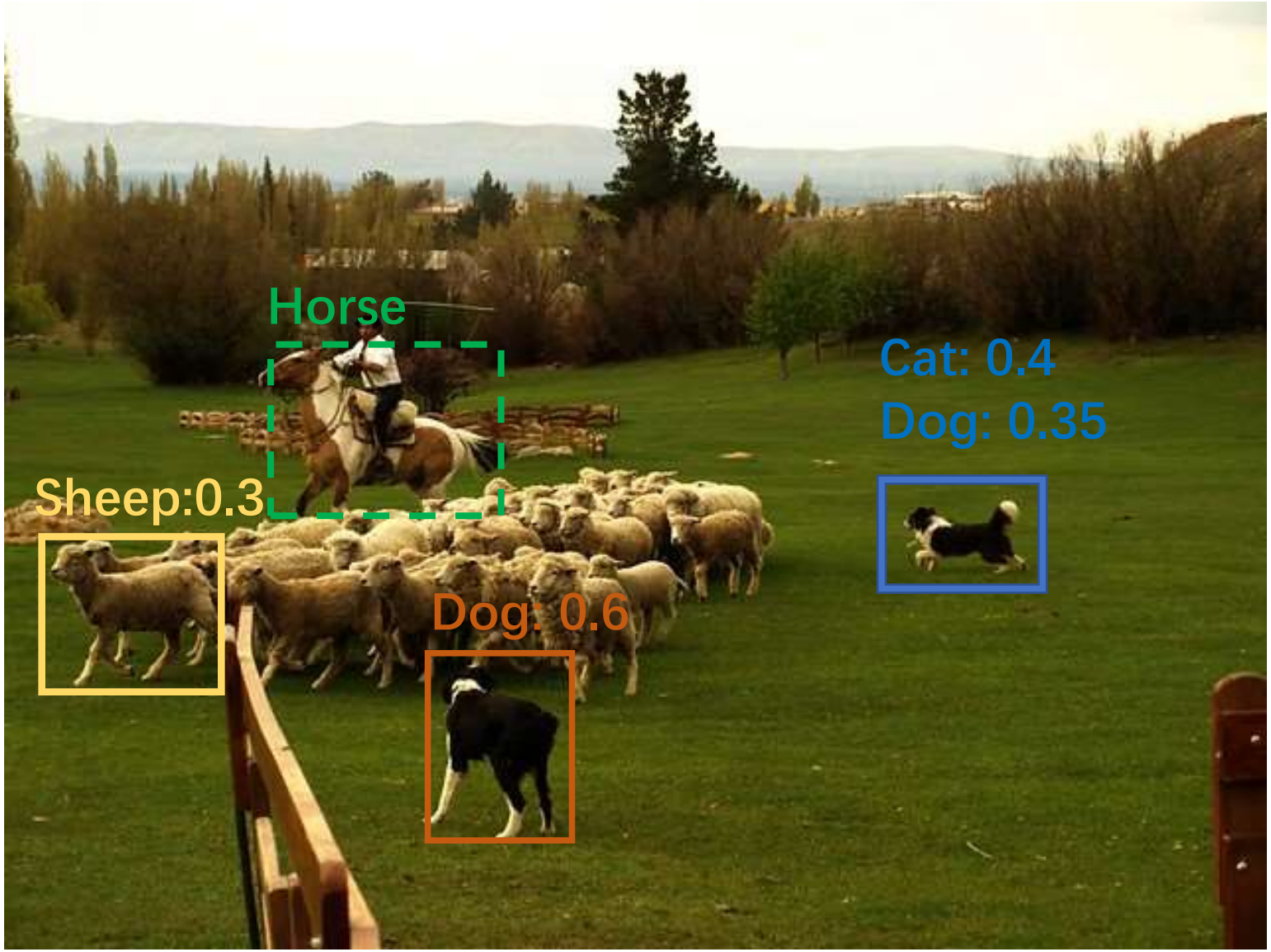
Motivation

Consist with GT

Prediction

FP	TP
FN	TN

Threshold = 0.3



$$\frac{\mathrm{TP}}{TP + FP} = 1 \qquad \frac{\mathrm{TP}}{TP + FN} = 0.5$$

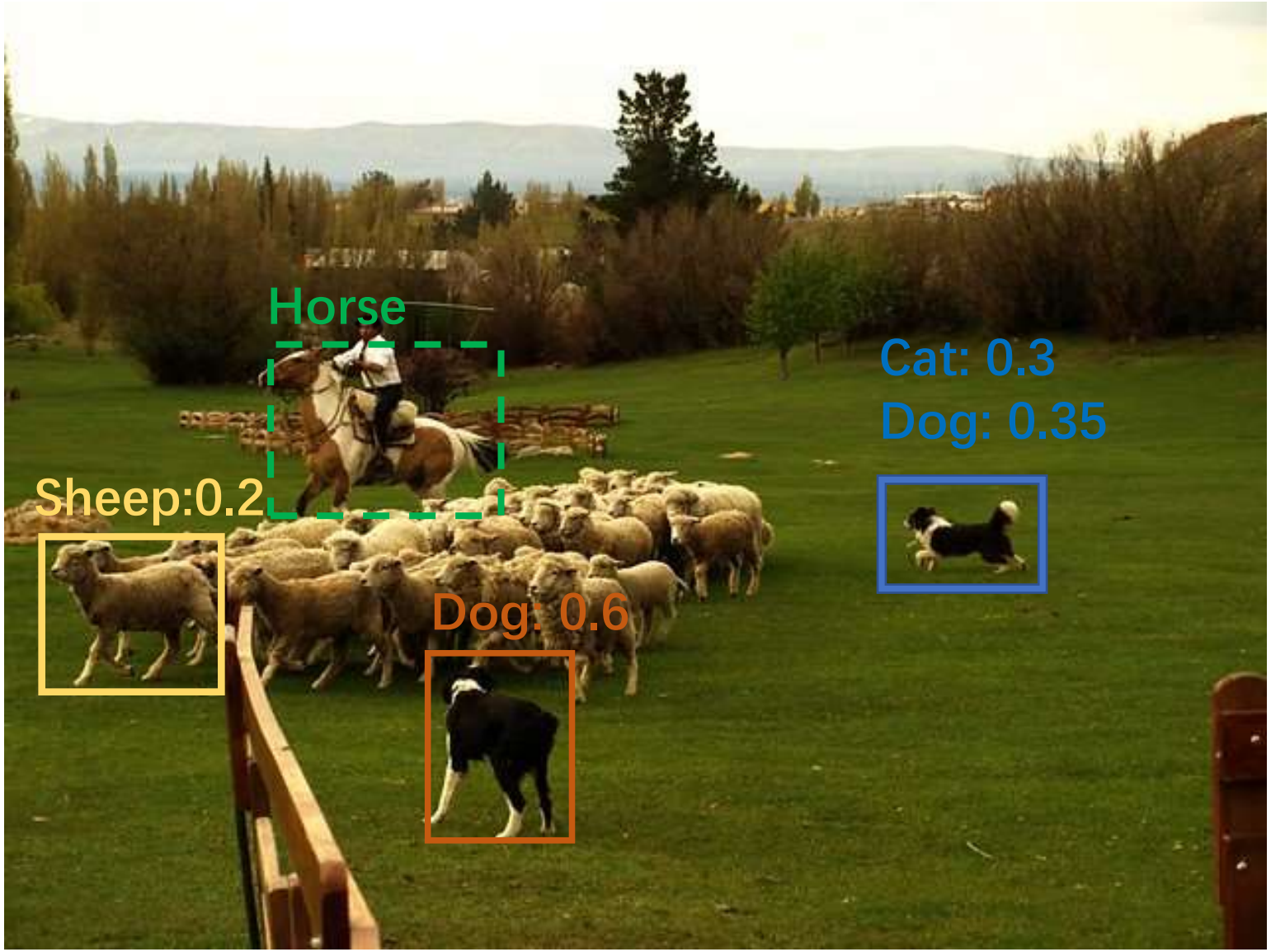
Motivation

Consist with GT

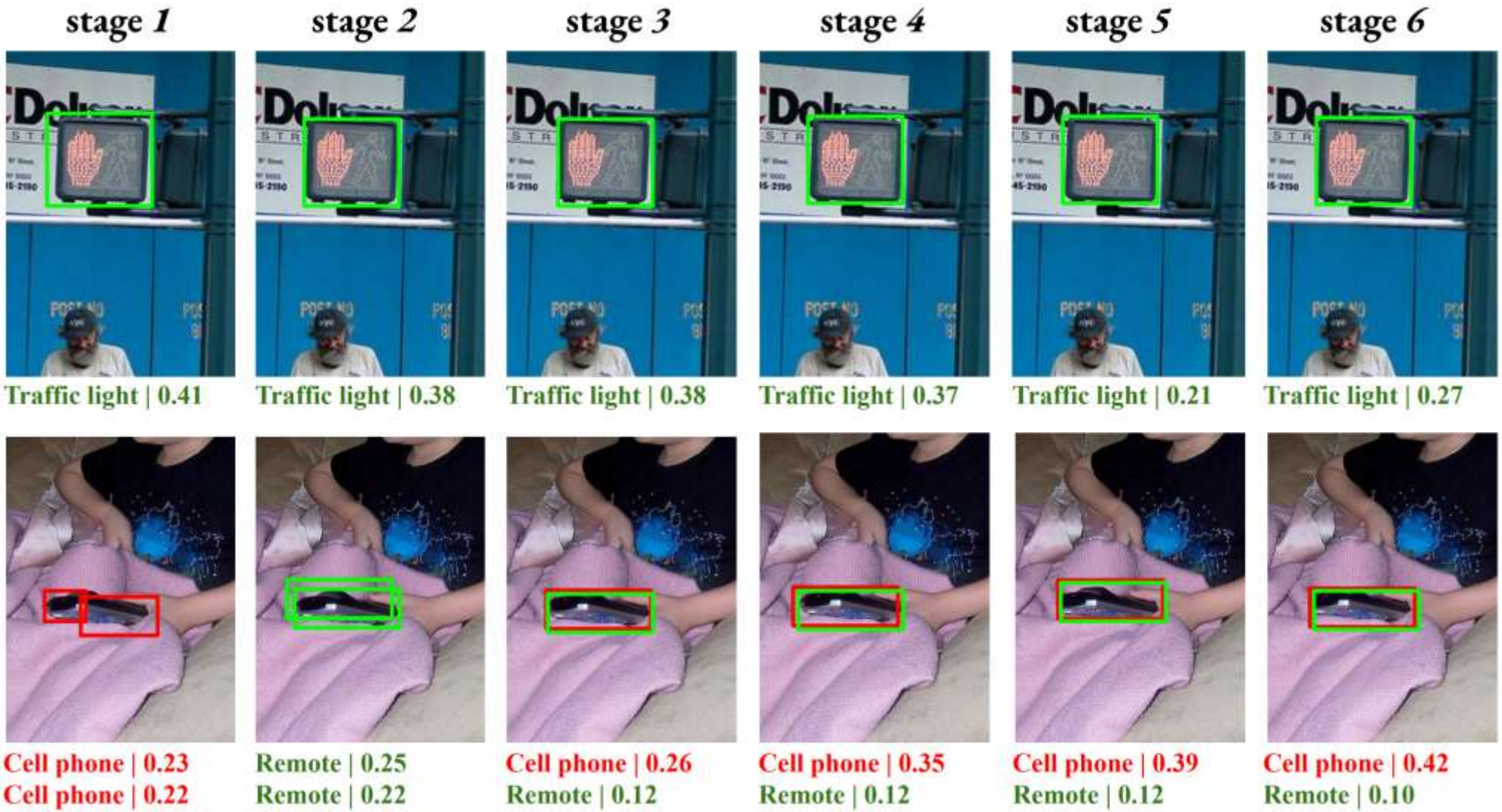
Prediction

FP	TP
FN	TN

Threshold = 0.3



Motivation


stage 1
stage 2
stage 3
stage 4
stage 5
stage 6
Traffic light | 0.41
Traffic light | 0.38
Traffic light | 0.38
Traffic light | 0.37
Traffic light | 0.21
Traffic light | 0.27
Cell phone | 0.23
Cell phone | 0.22
Remote | 0.25
Remote | 0.22
Cell phone | 0.26
Remote | 0.12
Cell phone | 0.35
Remote | 0.12
Cell phone | 0.39
Remote | 0.12
Cell phone | 0.42
Remote | 0.10

Motivation

- TP fading rate
  - P_i^6 is TP
  - P_i^{1-5} has higher IoU & higher category score
- FP exacerbation rate
  - P_i^6 is FP
  - P_i^{1-5} is FP but with lower category score

Model	TP Threshold	TP F Rate	FP E Rate
Deformable DETR	IOU>0.50	51.4%	55.7%
	IOU>0.75	49.5%	55.9%
Adamixer	IOU>0.50	28.6%	50.8%
	IOU>0.75	26.7%	51.2%

Method - Expectancy

- Uneven supervision
  - Enhancing later stages for better final outcomes
- Early-stage queries directly introduced to later ×
  - Mitigate the impact of cascading errors

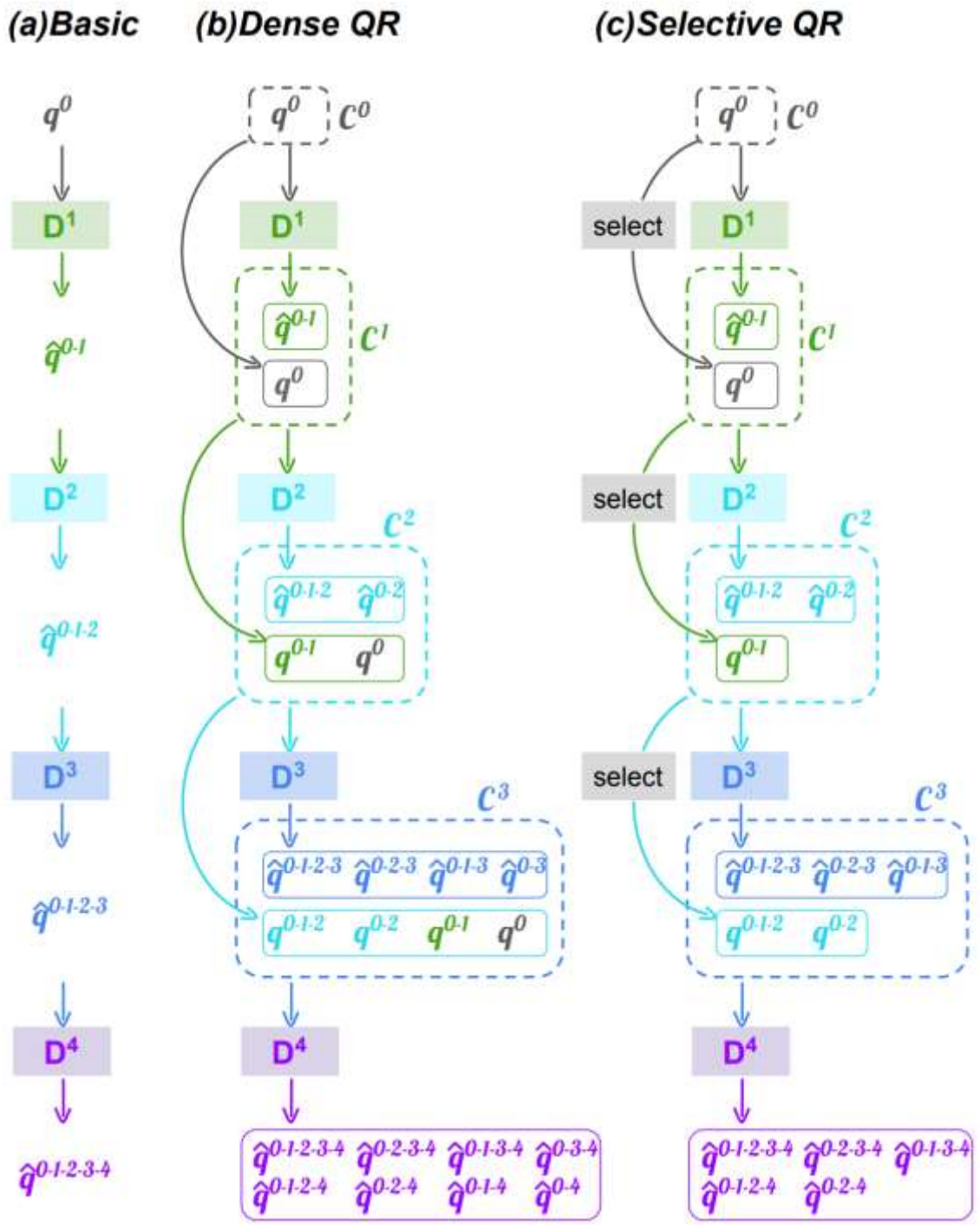
$$q^0 - D^1 \rightarrow q^1 - D^2 \rightarrow q^2 - D^3 \rightarrow q^3$$

$$q^1 = D^1(q^0) \times$$

$$q^1 = \{q^0, D^1(q^0)\}$$

Method

- Dense QR
  - $q^n = \{q^{n-1}, D^n(q^{n-1})\}$
- Selective QR
  - $q^n = \{D^{n-1}(q^{n-2}), D^n(q^{n-1})\}$



Method

- Dense QR
  - $q^n = \{q^{n-1}, D^n(q^{n-1})\}$
- Selective QR
  - $q^n = \{D^{n-1}(q^{n-2}), D^n(q^{n-1})\}$

Methods	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Baseline	42.5	61.5	45.6	24.6	45.1	59.2
DQR	44.2	62.8	47.9	26.7	46.9	60.5
SQR	44.4	63.2	47.8	25.7	47.4	60.2

Table 4. AP comparison among Baseline, DQR, and SQR

Experiment

Method	Start Stage	Train Time	AP	AP_{50}
Baseline	-	1x(5hours)	42.5	61.5
Baseline	-	2x	42.5	61.3
Baseline	-	3x	42.5	61.4
DQR	-	2.24x	44.2	62.8
SQR	1	1.57x	44.4	63.2
SQR	2	1.34x	44.2	63.0
SQR	3	1.18x	43.8	62.3
SQR	4	1.07x	42.9	61.4

Table 5. Further comparison among Baseline, DQR, and SQR with different starting stage in terms of training time and AP.

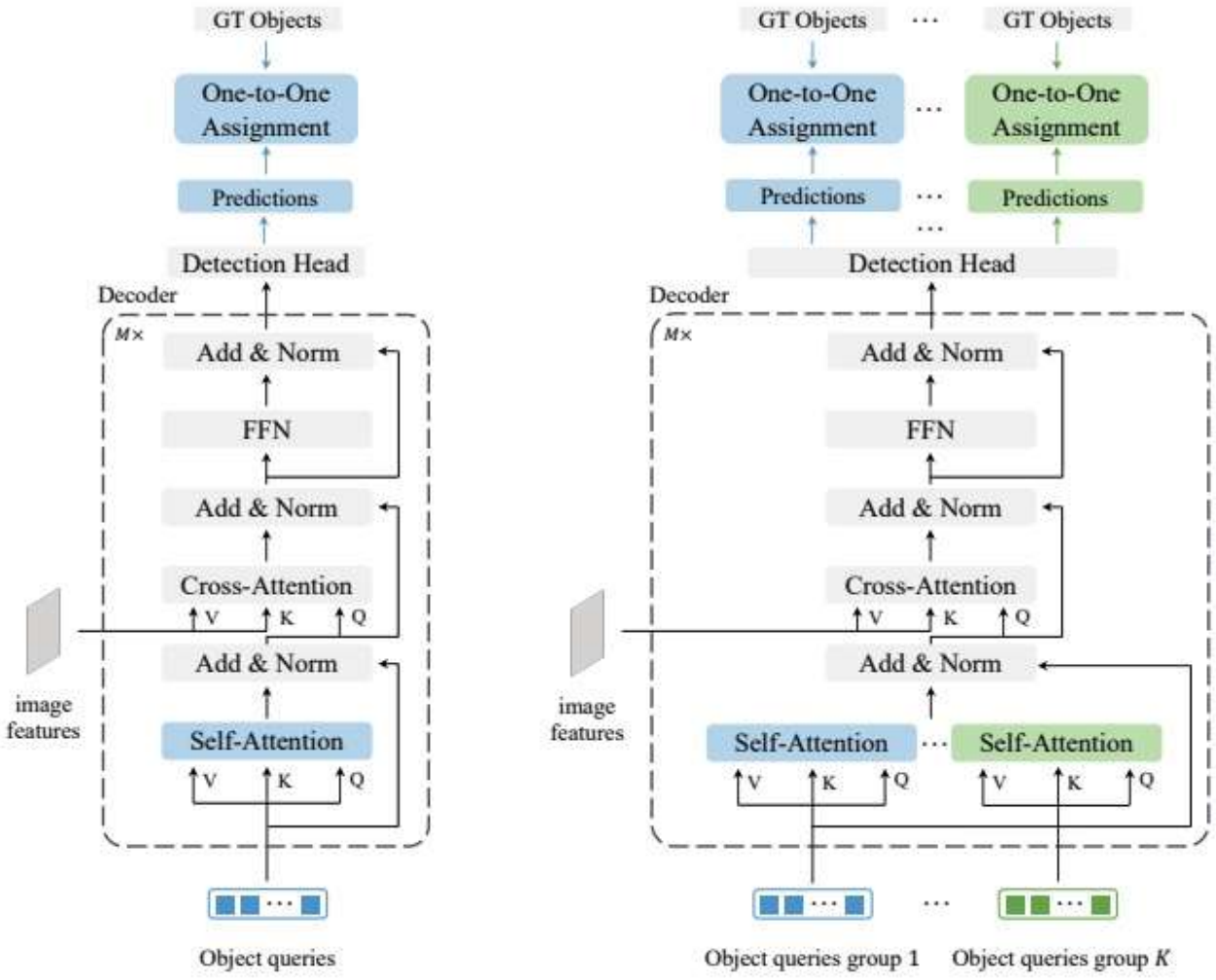
Method	TP Threshold	TP F Rate	FP E Rate
Baseline	IOU>0.50	28.6%	50.8%
SQR	IOU>0.50	23.3 %	47.3 %
Baseline	IOU>0.75	26.7%	51.2%
SQR	IOU>0.75	21.1%	47.0%

Table 6. Baseline vs. SQR on true-positive fading rate and false-positive exacerbation rate.

Experiment

Design	#*Supv* / stage	#*Supv*	AP
I (Group DETR)	3,3,3,3,3,3	18	43.4
II	4,4,4,3,2,1	18	43.0
III	1,2,3,4,4,4	18	43.7
IV (SQR)	1,1,2,3,5,8	20	44.2
V (Group DETR)	6,6,6,6,6,6	36	43.6
VI (SQR)	1,2,3,5,8,13	32	44.4

Table 7. Results of the 6 designed training strategies on Adamixer to investigate the relation with number of supervision. The inference is untouched. #*Supv* denotes the number of supervision.



(a) Original DETR

(b) Our Group DETR

Model	w/ SQR	#query	#epochs	COCO 2017 validation split					
				AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR-R50 [3]		100	500	42.0	62.4	44.2	20.5	45.8	61.1
Conditional DETR-R50 [24]		300	50	40.9	61.8	43.3	20.8	44.6	60.2
Conditional DETR-R101 [24]		300	50	42.8	63.7	46.0	21.7	46.6	60.9
Anchor-DETR-R50 [31]		300	50	42.1	63.1	44.9	22.3	46.2	60.0
Anchor-DETR-R101 [31]		300	50	43.5	64.3	46.6	23.2	47.7	61.4
SAM-DETR-R50 [32]		300	50	39.8	61.8	41.6	20.5	43.4	59.6
*SMCA-DETR-R50 [8]		300	50	43.7	63.6	47.2	24.2	47.0	60.4
*SMCA-DETR-R50 [8]		300	108	45.6	65.5	49.1	25.9	49.3	62.6
*DN-DAB-DETR-R50 [17]		300	50	44.1	64.4	46.7	22.9	48.0	63.4
*DN-DAB-DETR-R101 [17]		300	50	45.2	65.5	48.3	24.1	49.1	65.1
*DAB-DETR-R50 [21]		300	50	42.2	63.1	44.7	21.5	45.7	60.3
*SQR-DAB-DETR-R50	✓	300	50	**44.5** (+2.3)	64.4	47.5	24.8	48.6	61.7
*DAB-DETR-SwinB [21]		300	50	49.0	71.0	53.0	29.6	53.8	68.3
*SQR-DAB-DETR-SwinB	✓	300	50	**51.6** (+2.6)	72.5	55.9	32.0	56.8	71.0
*Deformable DETR-R50 [36]		300	12	37.2	55.2	40.4	20.6	40.2	50.2
*SQR-Deformable DETR-R50	✓	300	12	**39.9** (+2.7)	58.4	43.7	23.8	43.2	53.3
*Deformable DETR-R50 [36]		300	50	44.5	63.2	48.9	28.0	47.8	58.8
*SQR-Deformable DETR-R50	✓	300	50	**45.9** (+1.4)	64.7	50.2	27.7	49.2	60.5
Adamixer-R50 [9]		100	12	42.5	61.5	45.6	24.6	45.1	59.2
SQR-Adamixer-R50	✓	100	12	**44.4** (+1.9)	63.2	47.8	25.7	47.4	60.2
†Adamixer-R50 [9]		100	12	42.5	61.5	45.8	24.4	45.2	58.7
†SQR-Adamixer-R50	✓	100	12	**45.3** (+2.8)	63.8	49.0	26.8	48.1	62.2
*†Adamixer-R50		100	36	45.1	63.9	48.9	28.3	47.8	60.6
*†SQR-Adamixer-R50	✓	100	36	**46.7** (+1.6)	65.2	50.3	29.4	49.6	62.1
*†Adamixer-R50		300	36	46.6	65.5	50.6	29.3	49.3	62.3
*†SQR-Adamixer-R50	✓	300	36	**48.9** (+2.3)	67.5	53.2	32.0	51.8	63.7
*†Adamixer-R101 [9]		100	36	45.7	64.7	49.6	27.8	49.1	61.2
*†SQR-Adamixer-R101	✓	100	36	**47.3** (+1.6)	66.0	51.3	30.1	50.7	62.2
*†Adamixer-R101 [9]		300	36	47.6	66.7	51.8	29.5	50.5	63.3
*†SQR-Adamixer-R101	✓	300	36	**49.8** (+2.2)	68.8	54.0	32.0	53.4	65.1

Table 8. Comparison results with various query-based detectors on COCO 2017 val. #query: the number of queries used during inference. * indicates models trained with multi-scale augmentation, † marks models with 7 decoder stages.