Accelerating DETR Convergence via Semantic-Aligned Matching

Gongjie Zhang[1] Zhipeng Luo[1,2] Yingchen Yu[1] Kaiwen Cui[1] Shijian Lu*[1]
[1]Nanyang Technological University, Singapore [2]SenseTime Research
{gongjiezhang, shijian.lu}@ntu.edu.sg {zhipeng001, yingchen001, kaiwen001}@e.ntu.edu.sg

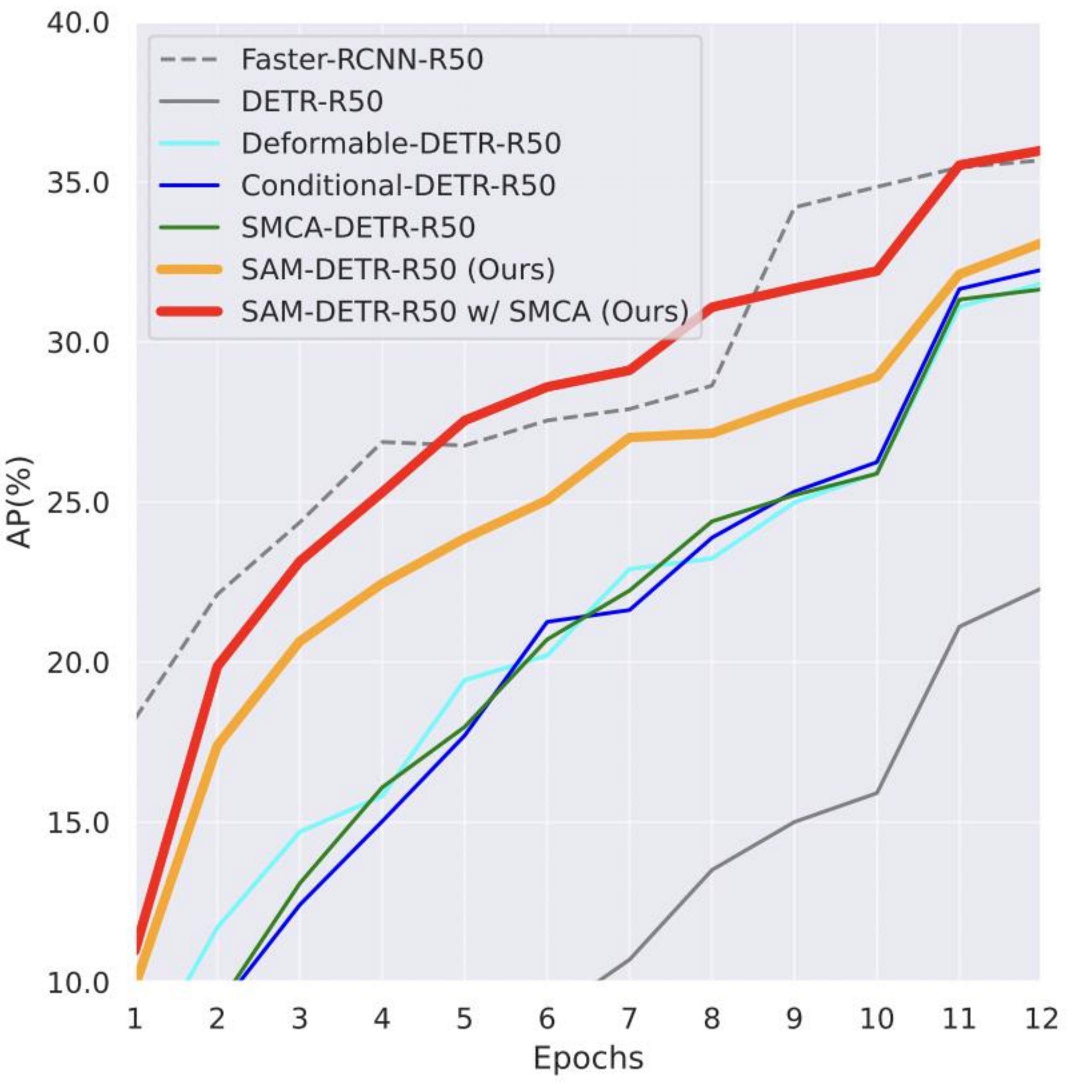
CVPR 2022

Motivation

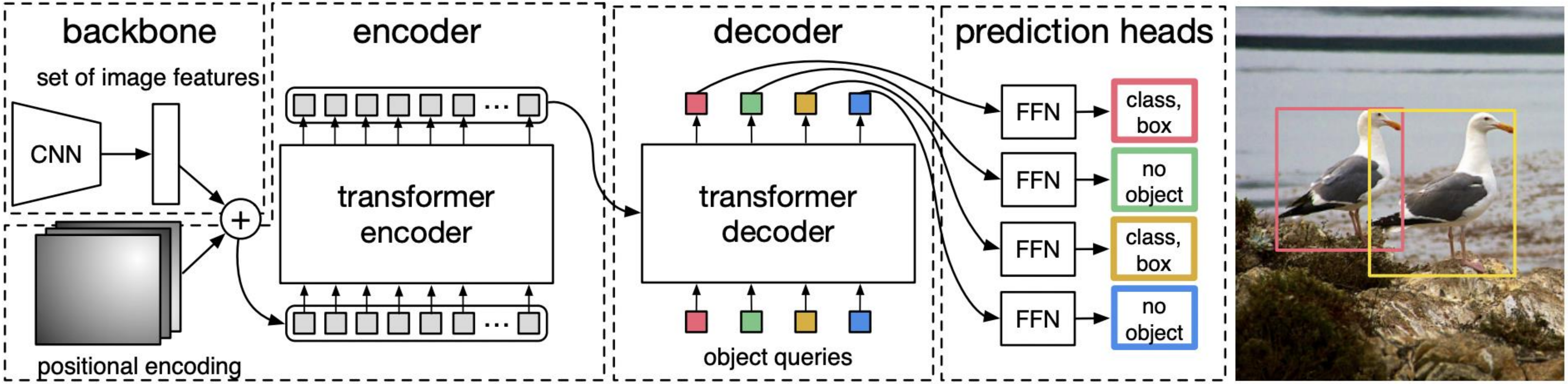
- DETR suffers from extremely slow convergence
- The difficulty in matching queries with features

Contribution

- A plug-and-play module
- Semantically aligning object queries with features
- Explicitly searching for objects' salient points
- Scalability

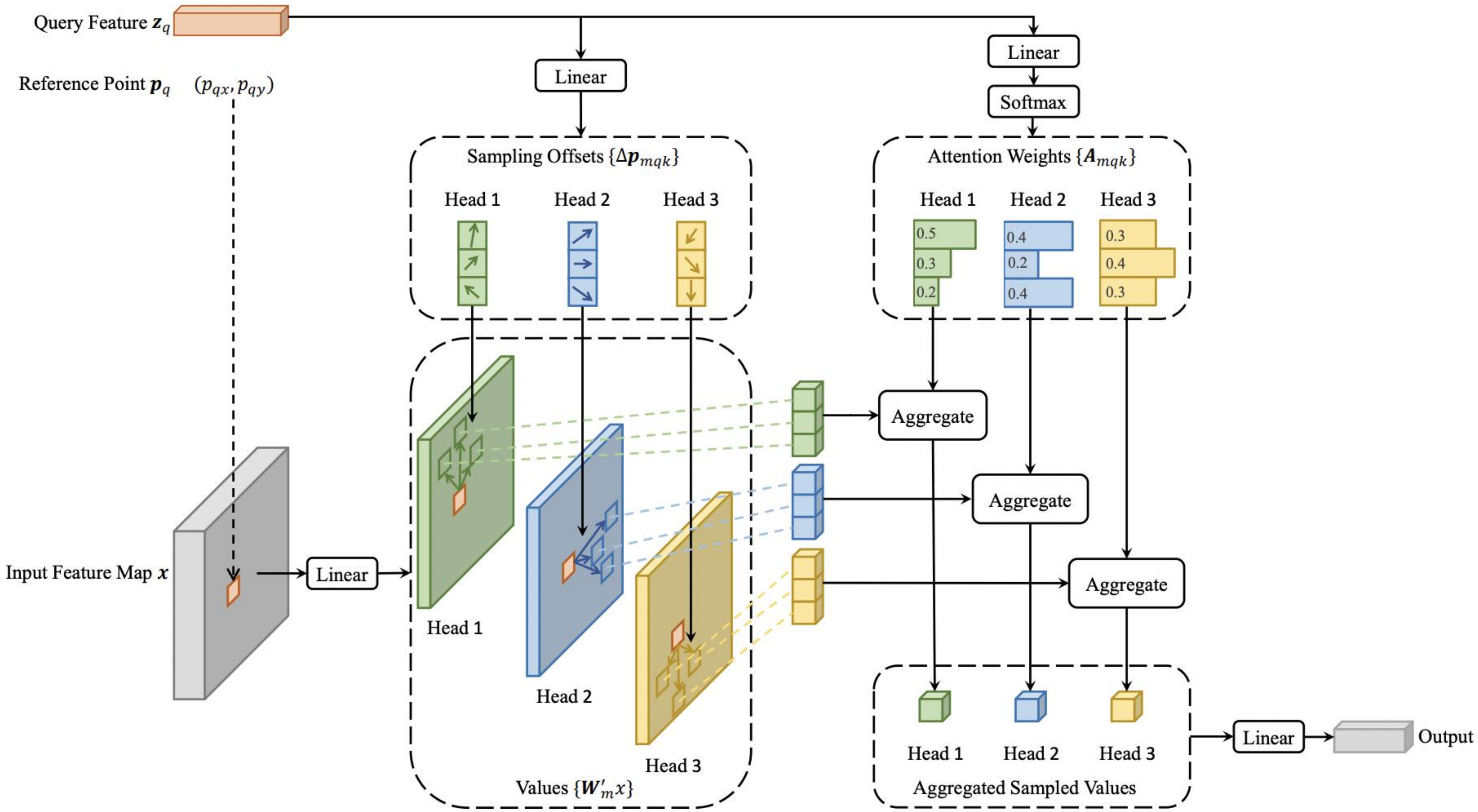


Related Work



DETR: Object query matched to all spatial locations because of its random initialization

Related Work


Query Feature $\boldsymbol{z}_q$
Reference Point $\boldsymbol{p}_q$
(p_{qx}, p_{qy})
Linear
Linear
Softmax
Sampling Offsets $\{\Delta\boldsymbol{p}_{mqk}\}$
Head 1
Head 2
Head 3
Attention Weights $\{\boldsymbol{A}_{mqk}\}$
Head 1
Head 2
Head 3
0.5
0.3
0.2
0.4
0.2
0.4
0.3
0.4
0.3
Aggregate
Aggregate
Aggregate
Input Feature Map $\boldsymbol{x}$
Linear
Head 1
Head 2
Head 3
Values $\{\boldsymbol{W}'_m x\}$
Head 1
Head 2
Head 3
Aggregated Sampled Values
Linear
Output

Deformable DETR: Replacing the original dense attention with deformable attention

Related Work

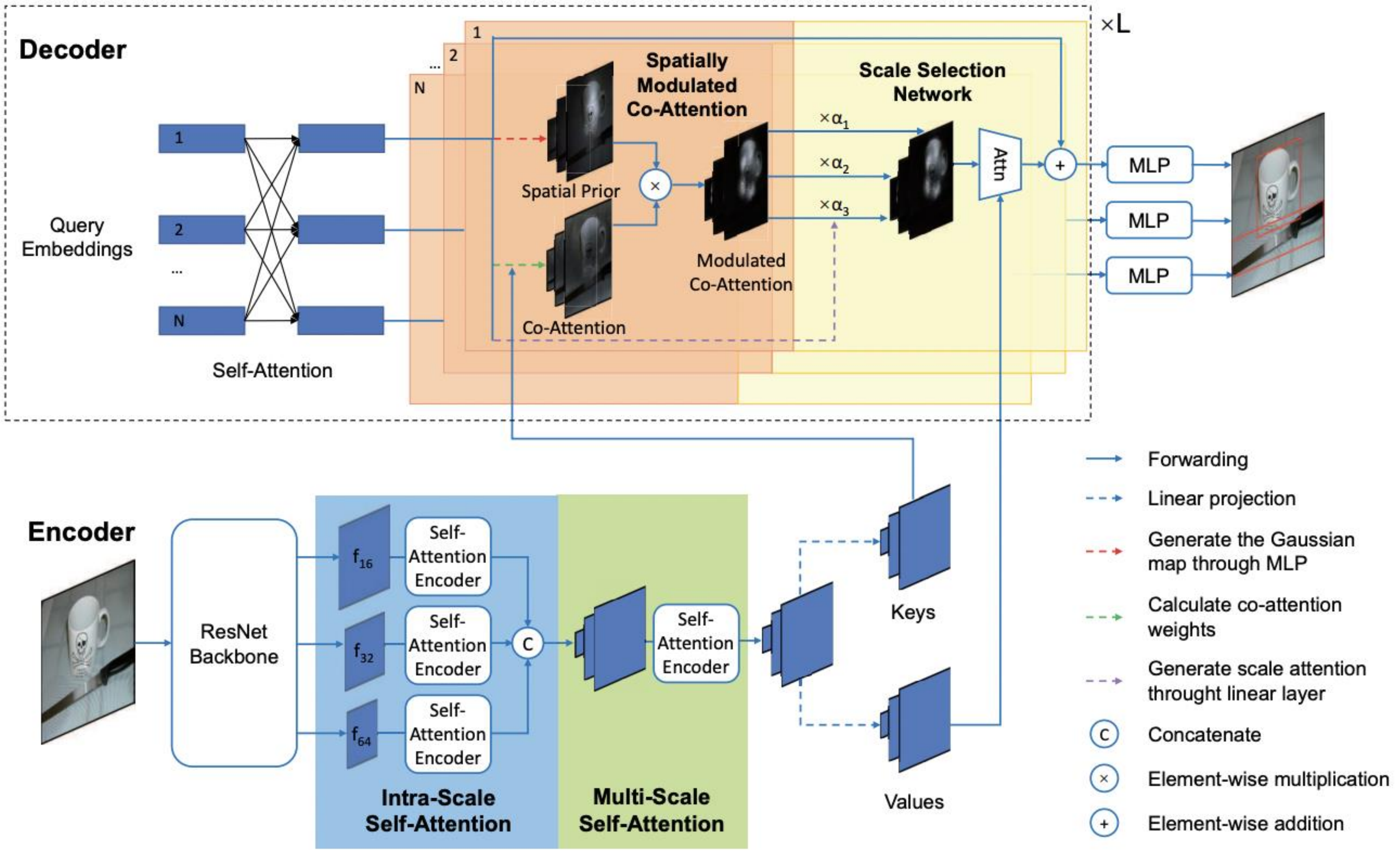
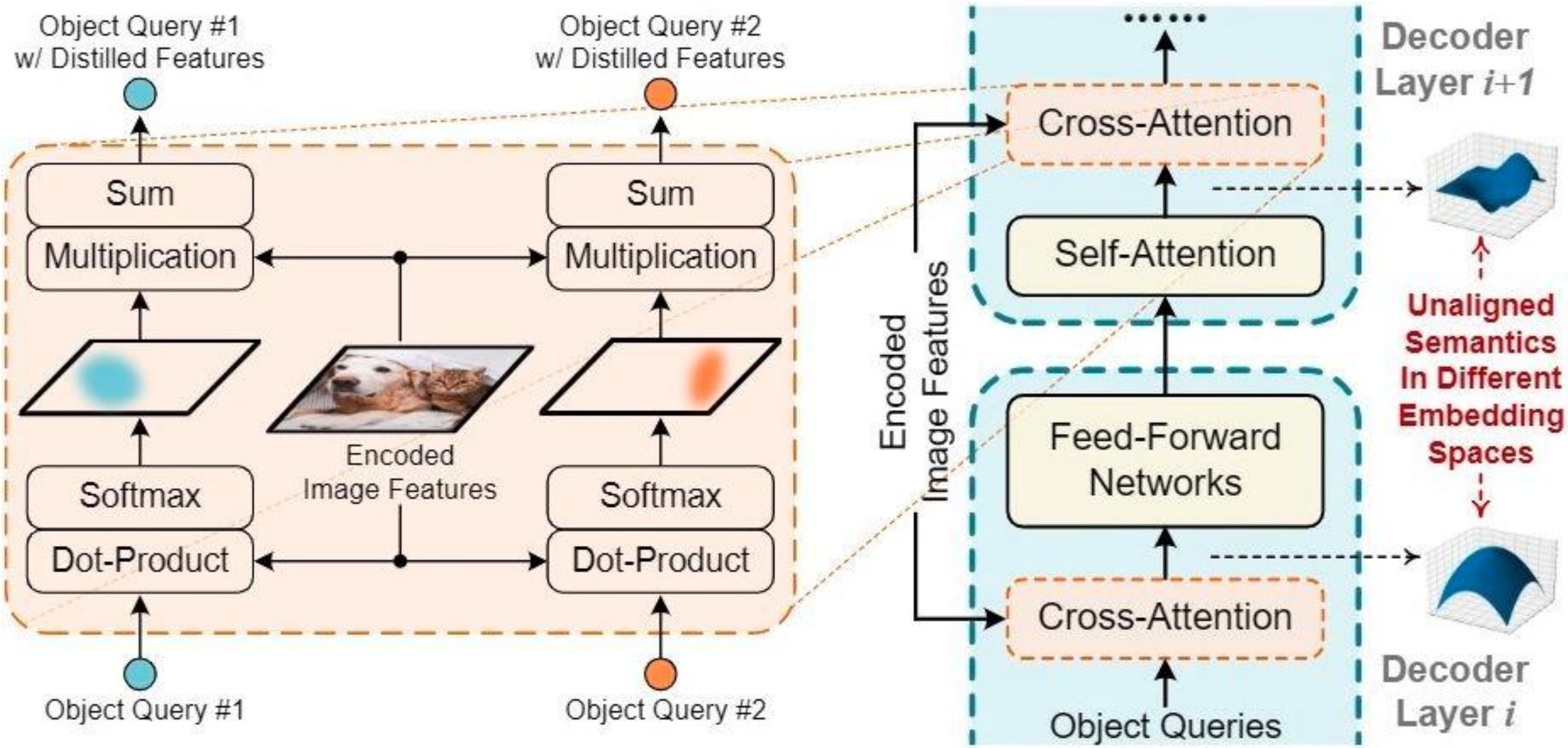


Figure 2. The overall pipeline of Spatially Modulated Co-Attention (SMCA) with intra-scale self-attention, multi-scale self-attention, spatial modulation, and scale-selection attention modules. Each object query performs spatially modulated co-attention and then predicts the target bounding boxes and their object categories. *N* stands for the number of object queries. *L* stands for the layers of decoder.

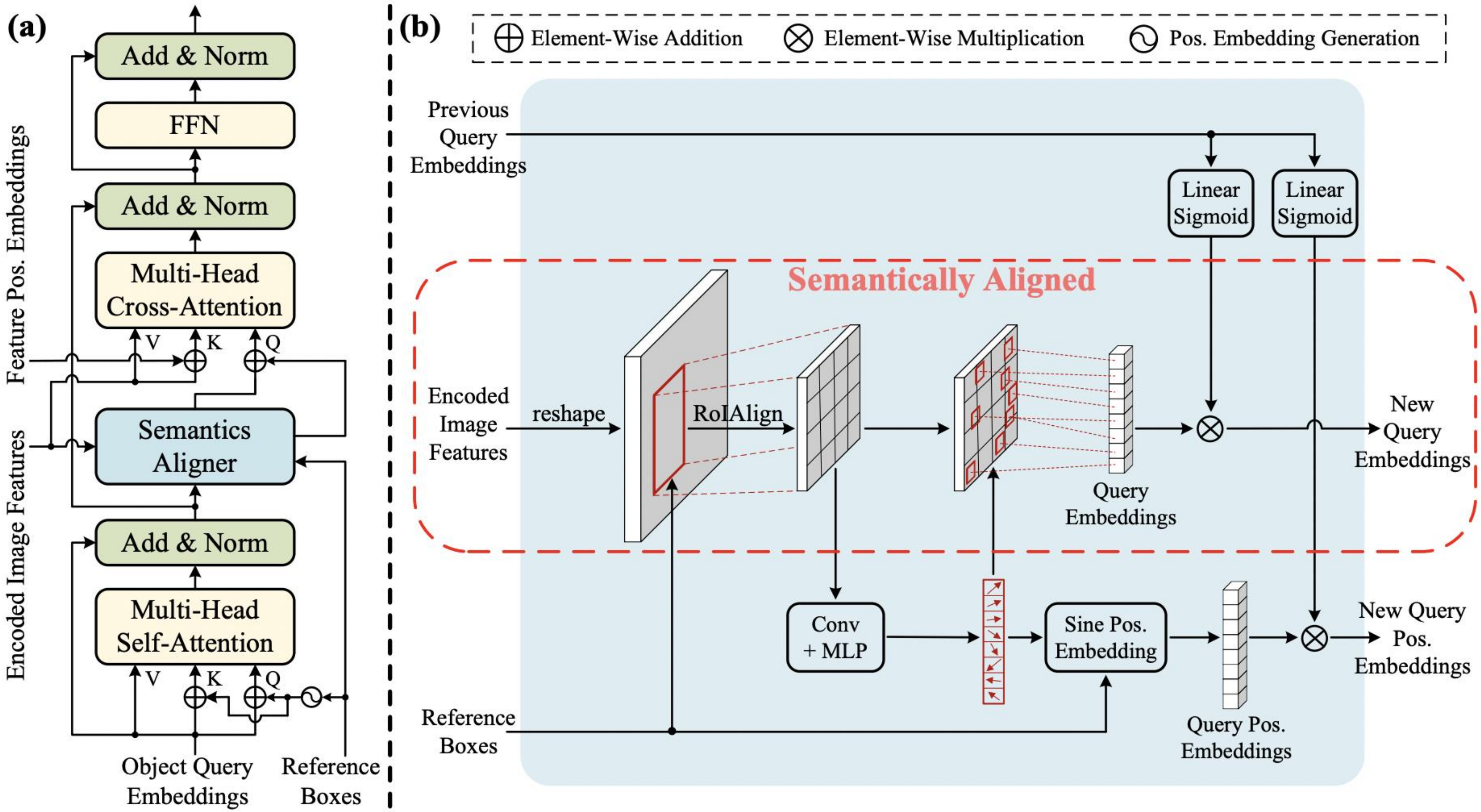
SMCA DETR: Replacing the cross-attention module to impose spatial constraints

Observation



The cross-attention module in DETR's decoder can be interpreted as a 'matching and feature distillation' process. Each object query first matches its own relevant regions in encoded image features, and then distills instance-level features from the matched regions for subsequent prediction. However, modules between cross-attentions perform projections on object queries, which leads to unaligned semantics between object queries and encoded image features.

Method



Step 1

$$\mathbf{F}_{\mathrm{R}} = \mathrm{RoIAlign}(\mathbf{F}, \mathbf{R}_{\mathrm{box}})$$

$$\mathbf{Q}^{\mathrm{new}}, \mathbf{Q}^{\mathrm{new}}_{\mathrm{pos}} = \mathrm{Resample}(\mathbf{F}_{\mathrm{R}}, \mathbf{R}_{\mathrm{box}}, \mathbf{Q})$$

Step 2

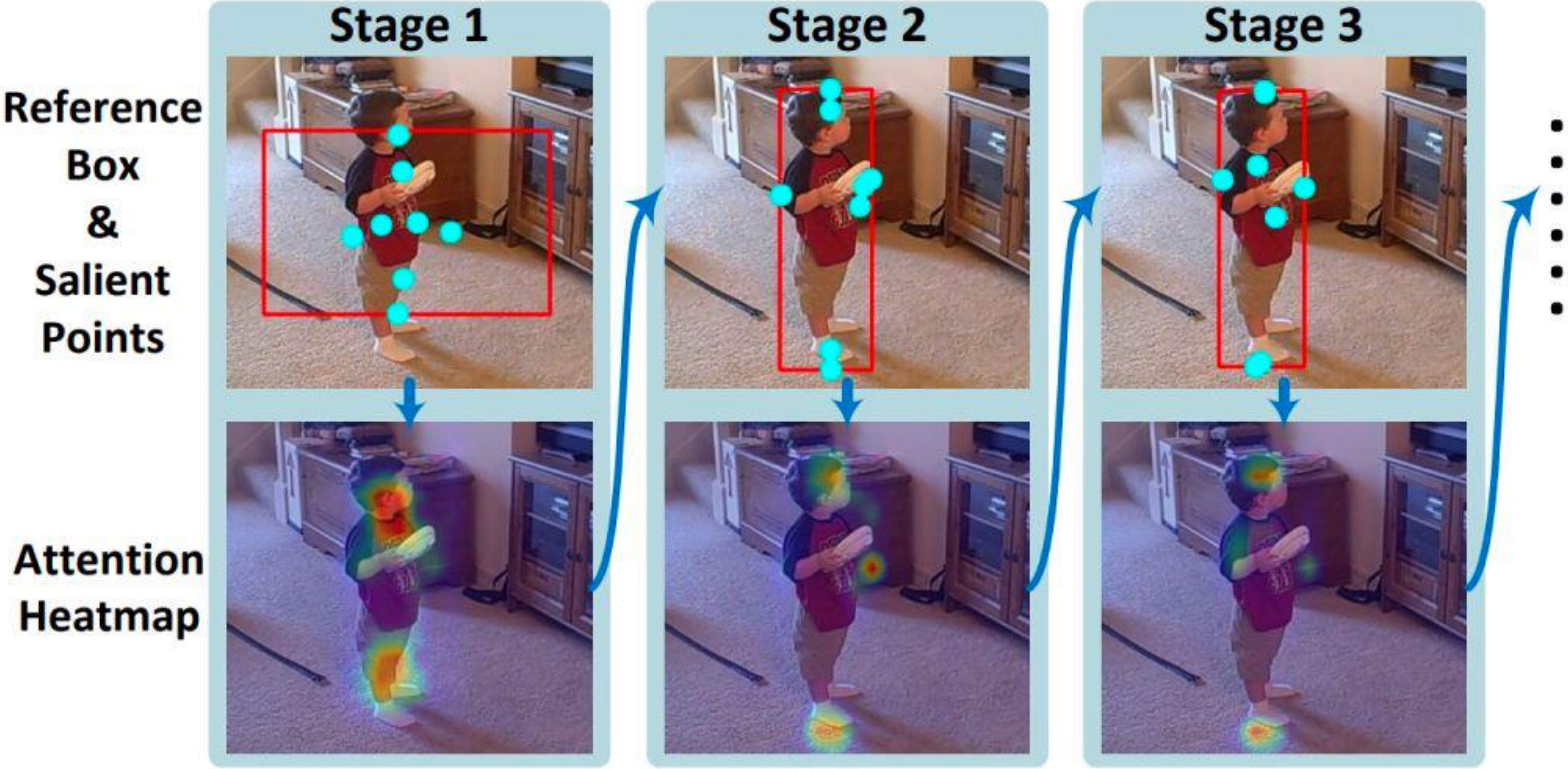
$$\mathbf{R}_{\mathrm{SP}} = \mathrm{MLP}(\mathrm{ConvNet}(\mathbf{F}_{\mathrm{R}}))$$

Step 3

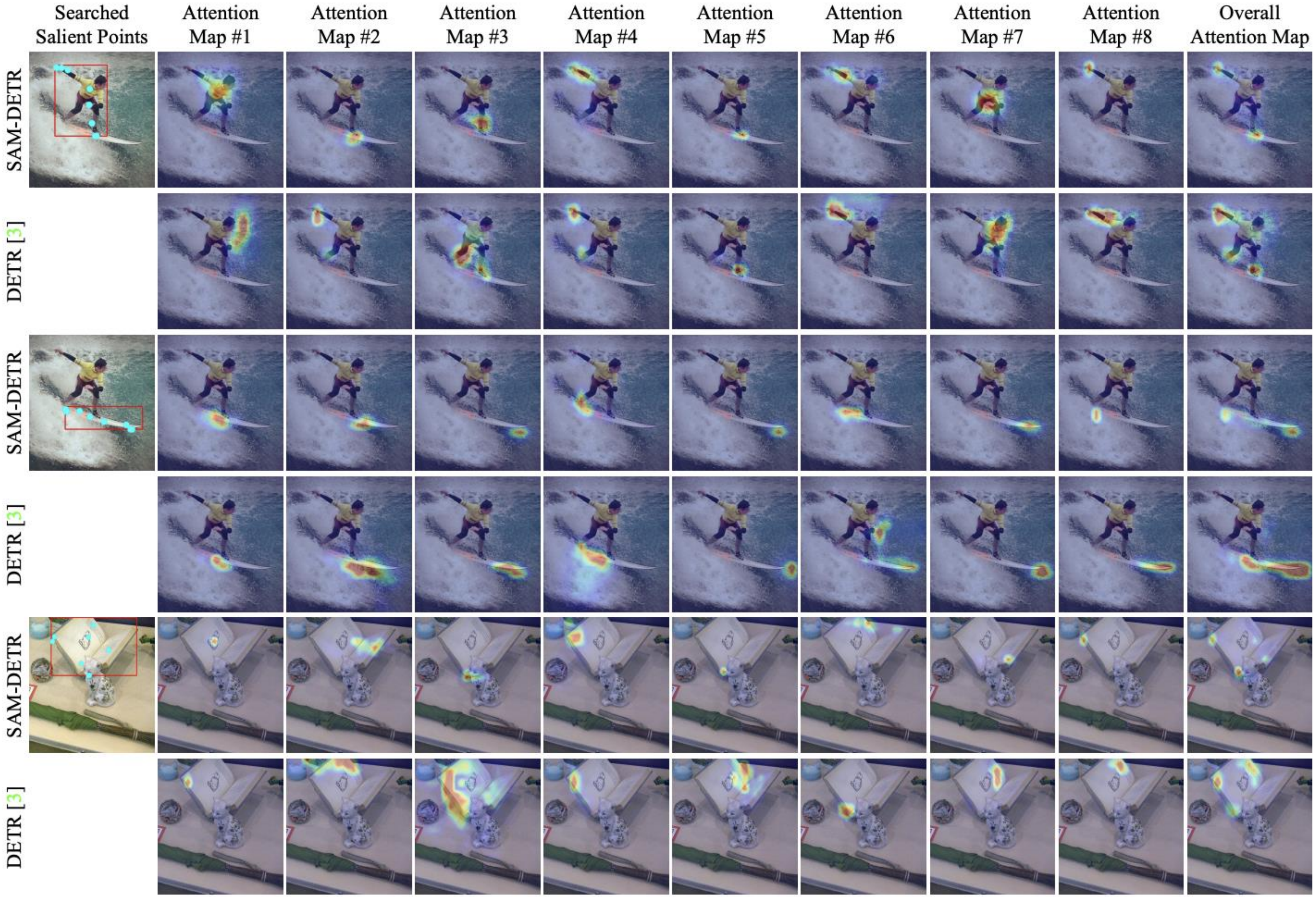
$$\mathbf{Q}^{\mathrm{new}} = \mathbf{Q}^{\mathrm{new}\prime} \otimes \sigma(\mathbf{Q}\mathbf{W}_{\mathrm{RW1}})$$

$$\mathbf{Q}^{\mathrm{new}}_{\mathrm{pos}} = \mathbf{Q}^{\mathrm{new}\prime}_{\mathrm{pos}} \otimes \sigma(\mathbf{Q}\mathbf{W}_{\mathrm{RW2}}),$$

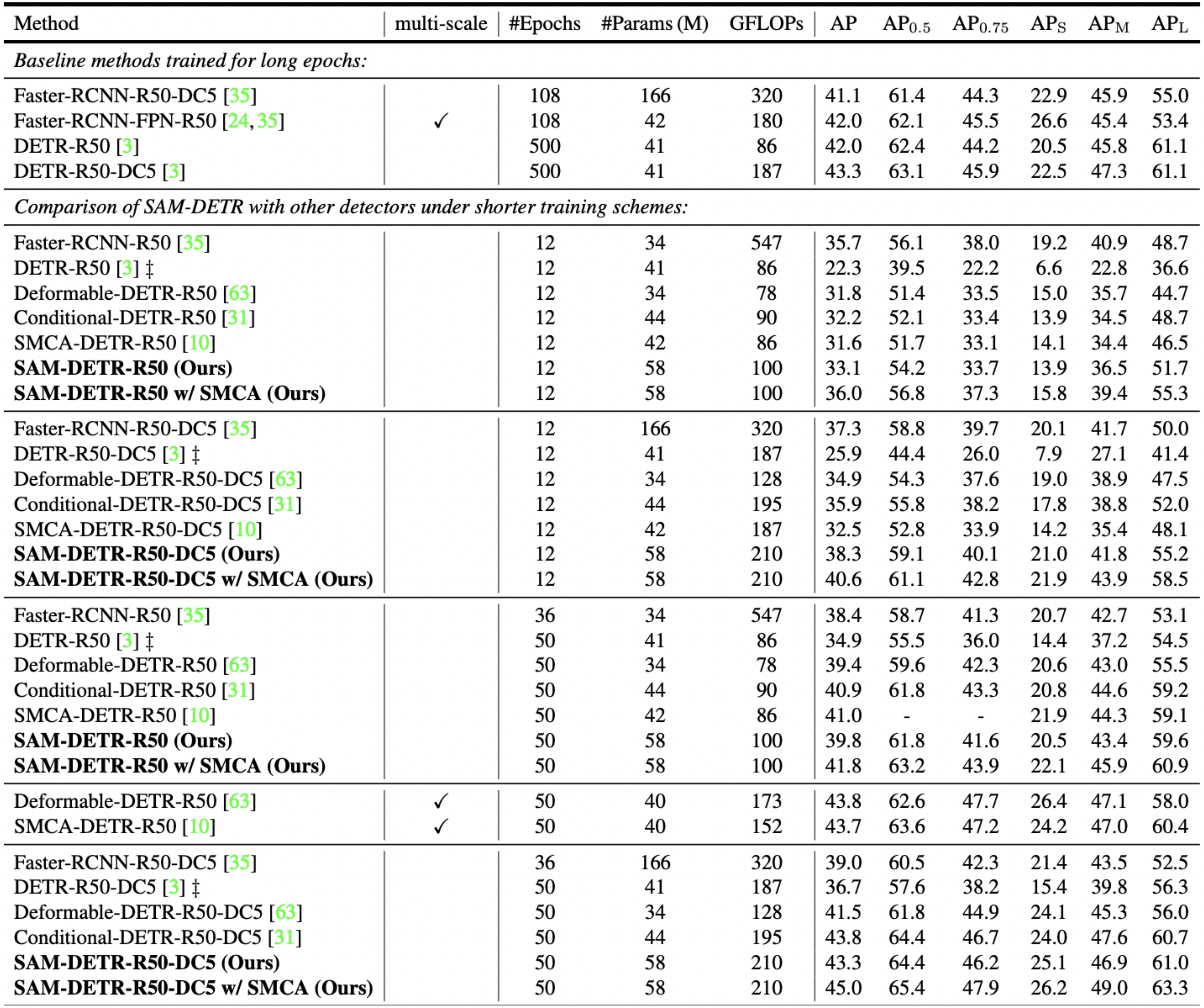
Visualization



Visualization

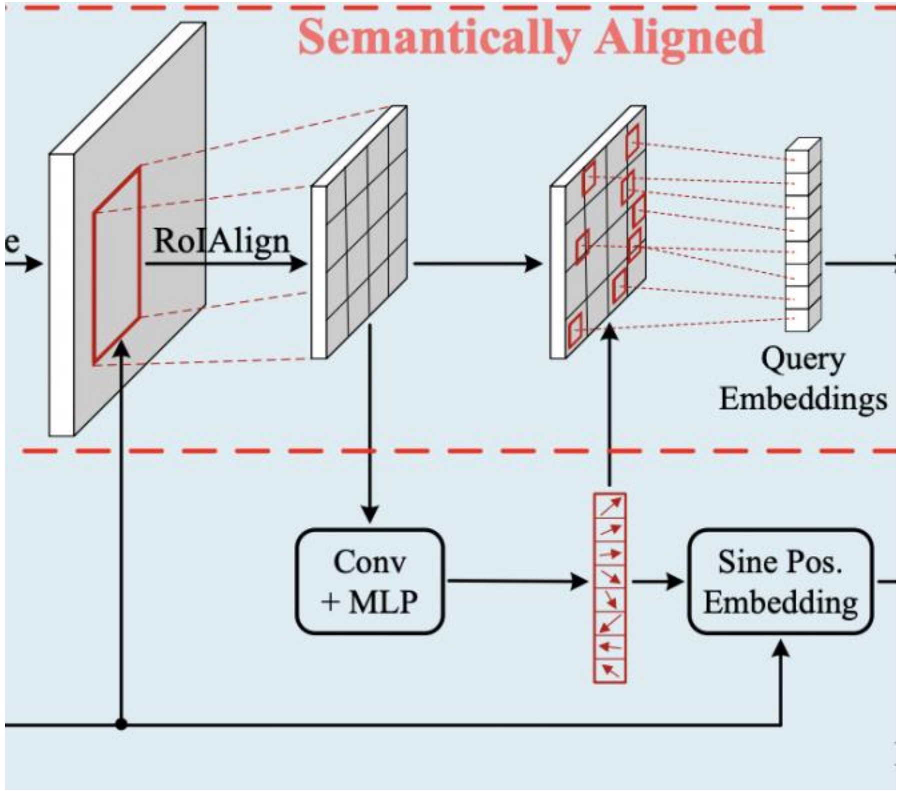


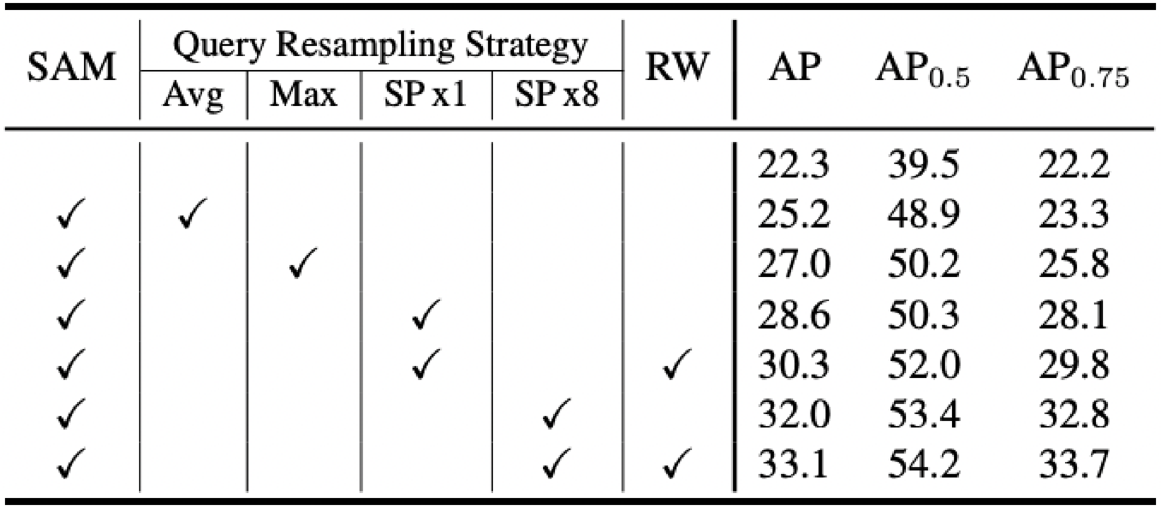
Experiment



Method	multi-scale	#Epochs	#Params (M)	GFLOPs	AP	AP$_{0.5}$	AP$_{0.75}$	AP$_S$	AP$_M$	AP$_L$
Baseline methods trained for long epochs:										
Faster-RCNN-R50-DC5 [35]		108	166	320	41.1	61.4	44.3	22.9	45.9	55.0
Faster-RCNN-FPN-R50 [24, 35]	✓	108	42	180	42.0	62.1	45.5	26.6	45.4	53.4
DETR-R50 [3]		500	41	86	42.0	62.4	44.2	20.5	45.8	61.1
DETR-R50-DC5 [3]		500	41	187	43.3	63.1	45.9	22.5	47.3	61.1
Comparison of SAM-DETR with other detectors under shorter training schemes:										
Faster-RCNN-R50 [35]		12	34	547	35.7	56.1	38.0	19.2	40.9	48.7
DETR-R50 [3] ‡		12	41	86	22.3	39.5	22.2	6.6	22.8	36.6
Deformable-DETR-R50 [63]		12	34	78	31.8	51.4	33.5	15.0	35.7	44.7
Conditional-DETR-R50 [31]		12	44	90	32.2	52.1	33.4	13.9	34.5	48.7
SMCA-DETR-R50 [10]		12	42	86	31.6	51.7	33.1	14.1	34.4	46.5
SAM-DETR-R50 (Ours)		12	58	100	33.1	54.2	33.7	13.9	36.5	51.7
SAM-DETR-R50 w/ SMCA (Ours)		12	58	100	36.0	56.8	37.3	15.8	39.4	55.3
Faster-RCNN-R50-DC5 [35]		12	166	320	37.3	58.8	39.7	20.1	41.7	50.0
DETR-R50-DC5 [3] ‡		12	41	187	25.9	44.4	26.0	7.9	27.1	41.4
Deformable-DETR-R50-DC5 [63]		12	34	128	34.9	54.3	37.6	19.0	38.9	47.5
Conditional-DETR-R50-DC5 [31]		12	44	195	35.9	55.8	38.2	17.8	38.8	52.0
SMCA-DETR-R50-DC5 [10]		12	42	187	32.5	52.8	33.9	14.2	35.4	48.1
SAM-DETR-R50-DC5 (Ours)		12	58	210	38.3	59.1	40.1	21.0	41.8	55.2
SAM-DETR-R50-DC5 w/ SMCA (Ours)		12	58	210	40.6	61.1	42.8	21.9	43.9	58.5
Faster-RCNN-R50 [35]		36	34	547	38.4	58.7	41.3	20.7	42.7	53.1
DETR-R50 [3] ‡		50	41	86	34.9	55.5	36.0	14.4	37.2	54.5
Deformable-DETR-R50 [63]		50	34	78	39.4	59.6	42.3	20.6	43.0	55.5
Conditional-DETR-R50 [31]		50	44	90	40.9	61.8	43.3	20.8	44.6	59.2
SMCA-DETR-R50 [10]		50	42	86	41.0	-	-	21.9	44.3	59.1
SAM-DETR-R50 (Ours)		50	58	100	39.8	61.8	41.6	20.5	43.4	59.6
SAM-DETR-R50 w/ SMCA (Ours)		50	58	100	41.8	63.2	43.9	22.1	45.9	60.9
Deformable-DETR-R50 [63]	✓	50	40	173	43.8	62.6	47.7	26.4	47.1	58.0
SMCA-DETR-R50 [10]	✓	50	40	152	43.7	63.6	47.2	24.2	47.0	60.4
Faster-RCNN-R50-DC5 [35]		36	166	320	39.0	60.5	42.3	21.4	43.5	52.5
DETR-R50-DC5 [3] ‡		50	41	187	36.7	57.6	38.2	15.4	39.8	56.3
Deformable-DETR-R50-DC5 [63]		50	34	128	41.5	61.8	44.9	24.1	45.3	56.0
Conditional-DETR-R50-DC5 [31]		50	44	195	43.8	64.4	46.7	24.0	47.6	60.7
SAM-DETR-R50-DC5 (Ours)		50	58	210	43.3	64.4	46.2	25.1	46.9	61.0
SAM-DETR-R50-DC5 w/ SMCA (Ours)		50	58	210	45.0	65.4	47.9	26.2	49.0	63.3

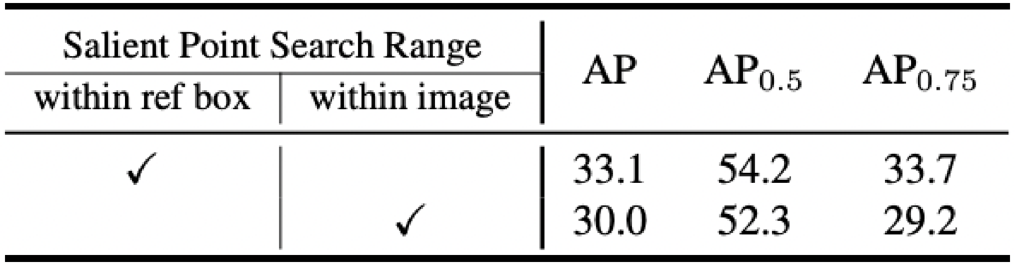
Ablation Study





SAM	Query Resampling Strategy				RW	AP	$AP_{0.5}$	$AP_{0.75}$
	Avg	Max	SP x1	SP x8				
						22.3	39.5	22.2
✓	✓					25.2	48.9	23.3
✓		✓				27.0	50.2	25.8
✓			✓			28.6	50.3	28.1
✓			✓		✓	30.3	52.0	29.8
✓				✓		32.0	53.4	32.8
✓				✓	✓	33.1	54.2	33.7

Table 2. Ablation studies on our proposed design choices. Results are obtained on COCO val 2017. 'SAM' denotes the proposed Semantic-Aligned Matching. 'RW' denotes reweighting by previous query embeddings. Different resampling strategies for SAM are studied, including average-pooling (Avg), max-pooling (Max), one salient point (SP x1), and eight salient points (SP x8).



Salient Point Search Range		AP	$AP_{0.5}$	$AP_{0.75}$
within ref box	within image			
✓		33.1	54.2	33.7
	✓	30.0	52.3	29.2

Table 3. Ablation study on the salient point search range. Results are obtained on COCO val 2017.