Sparse DETR:

Efficient End-to-End Object Detection with Learnable Sparsity

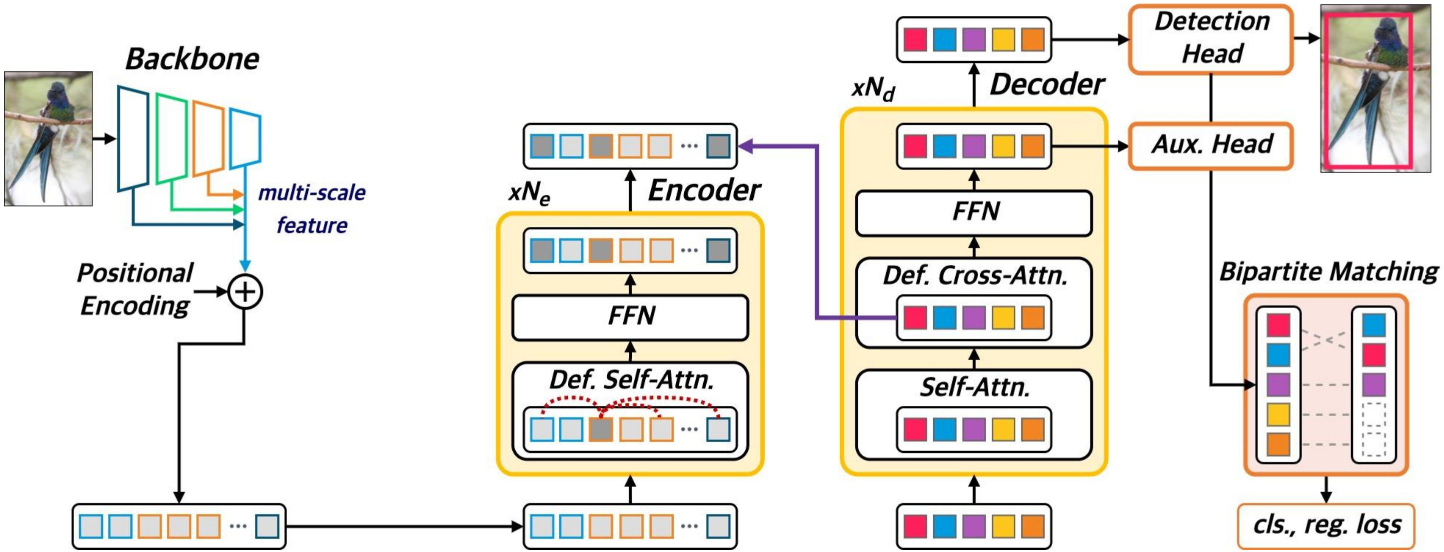
Byungseok Roh*, Jaewoong Shin*, Wuhyun Shin*, Saehoon Kim

Kakao Brain

*equal contribution

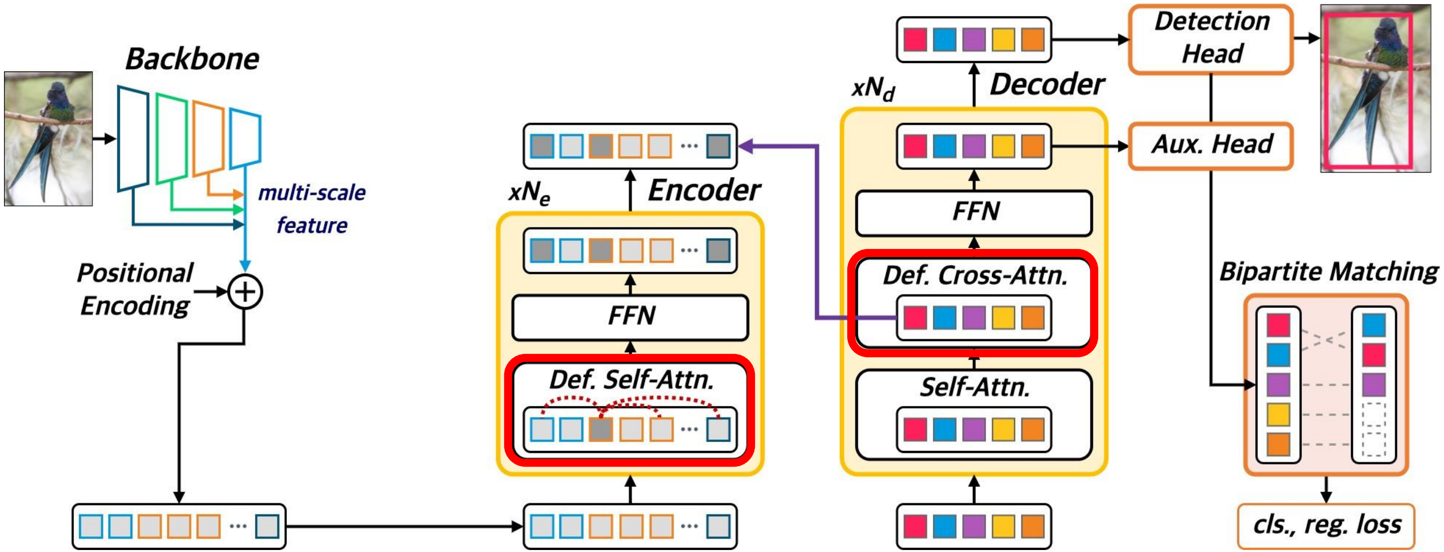
Motivation

- Deformable DETR introduces **deformable attention** which **reduces computation cost** from **quadratic to linear** complexity



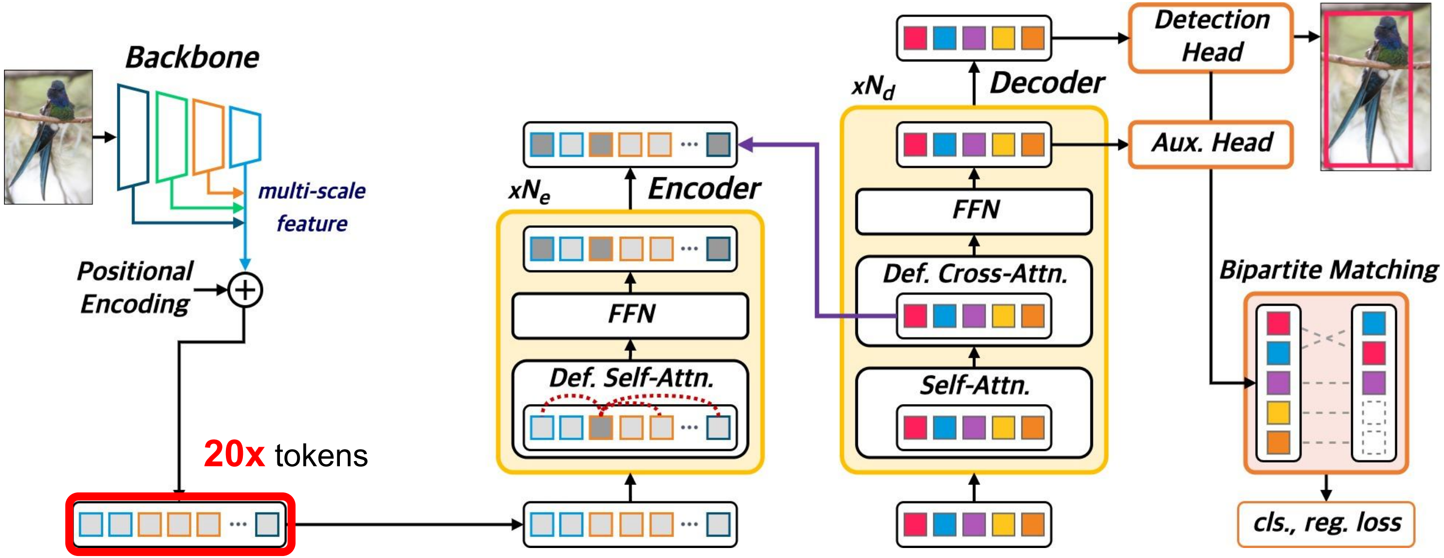
Motivation

- Deformable DETR introduces **deformable attention** which **reduces computation cost** from **quadratic to linear** complexity



Motivation

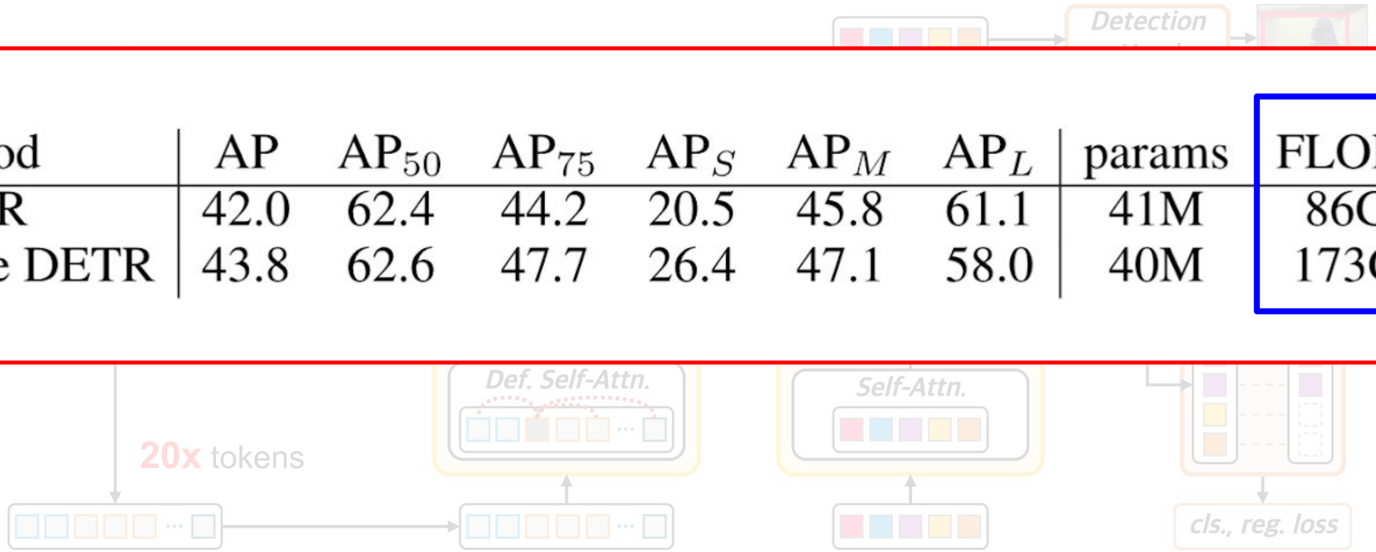
- Using **the multi-scale features** as an encoder input **increases** the **number of tokens** to be processed by about **20 times**



Motivation

- Using **the multi-scale features** as an encoder input **increases** the **number of tokens** to be processed by about **20 times**

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	FPS
DETR	42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	28
Deformable DETR	43.8	62.6	47.7	26.4	47.1	58.0	40M	173G	19



Characteristic of Images for Object Detection

- On average, **only 30%** of the entire image is the foreground pixel.

MS COCO dataset

ICLR

kakaobrain

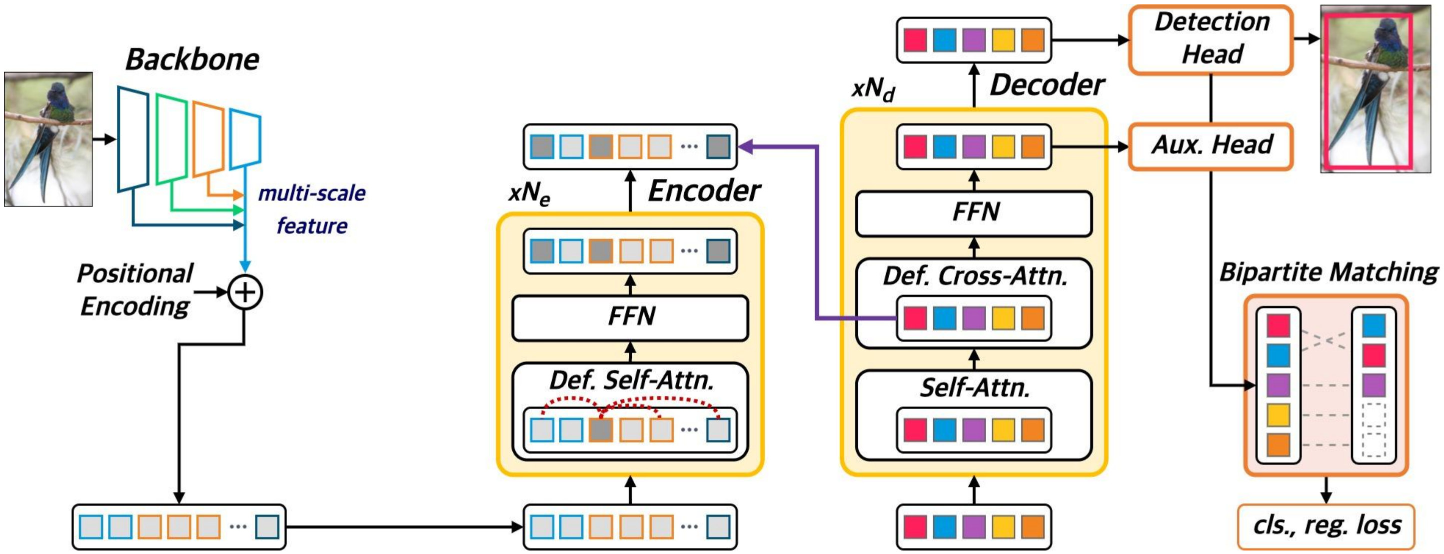
Characteristic of Images for Object Detection

- On average, **only 30%** of the entire image is the foreground pixel.

MS COCO dataset

- Do we need to compute **the entire token** in the encoder block?

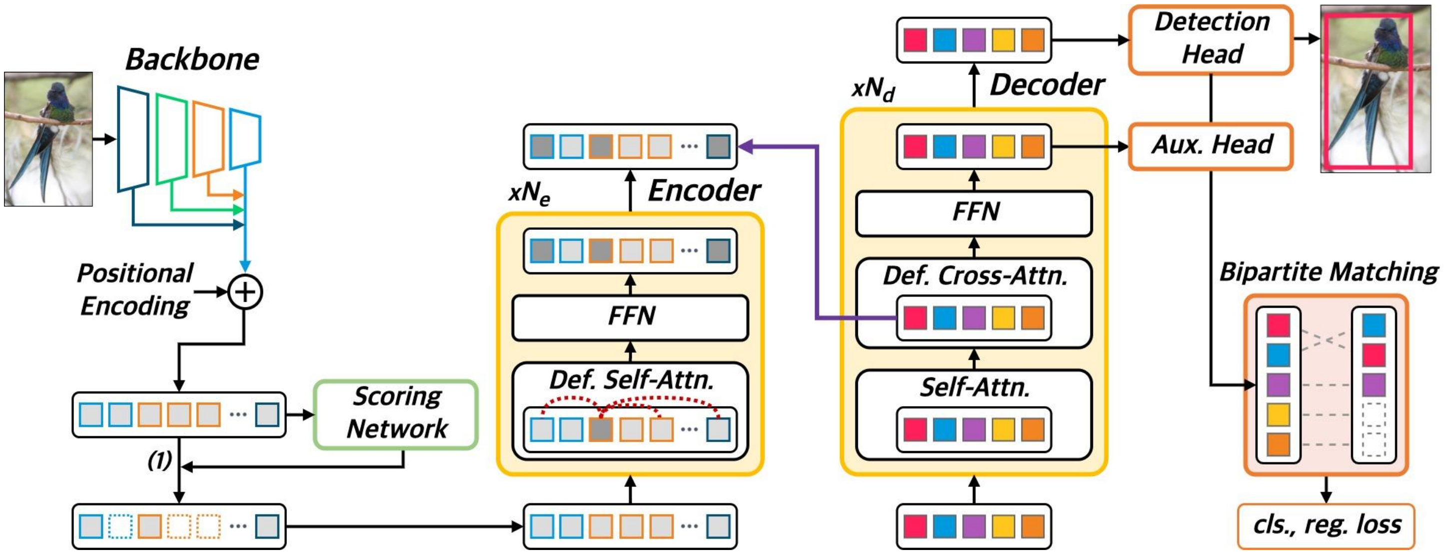
Architecture



ICLR

kakaobrain

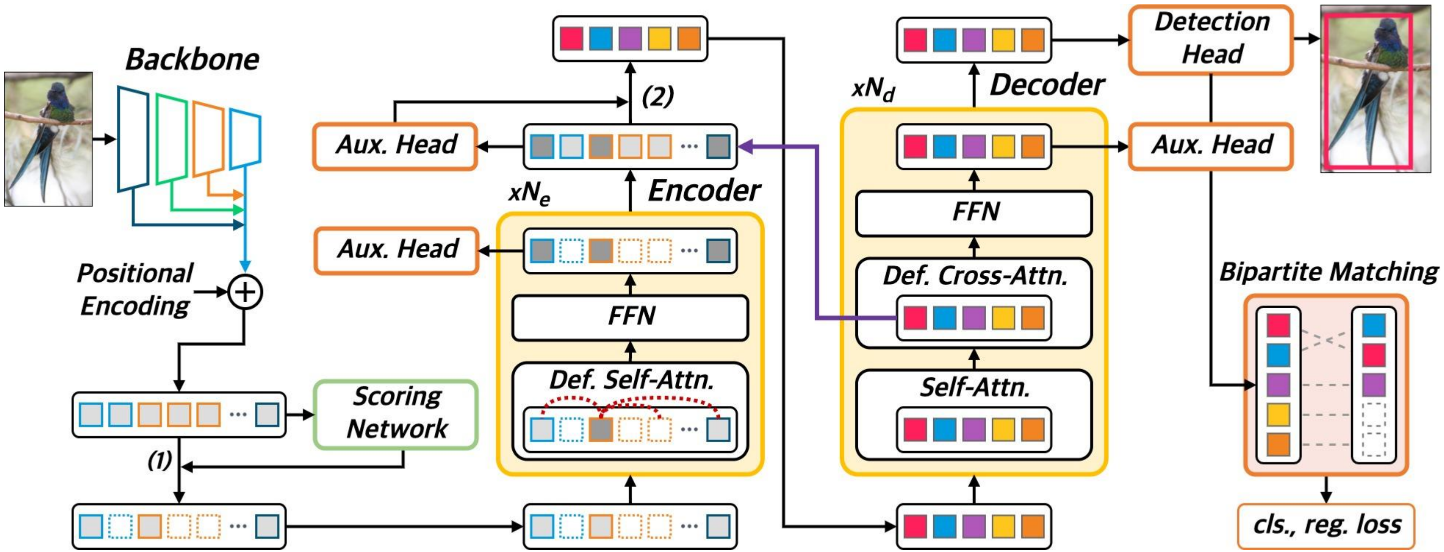
Architecture



ICLR

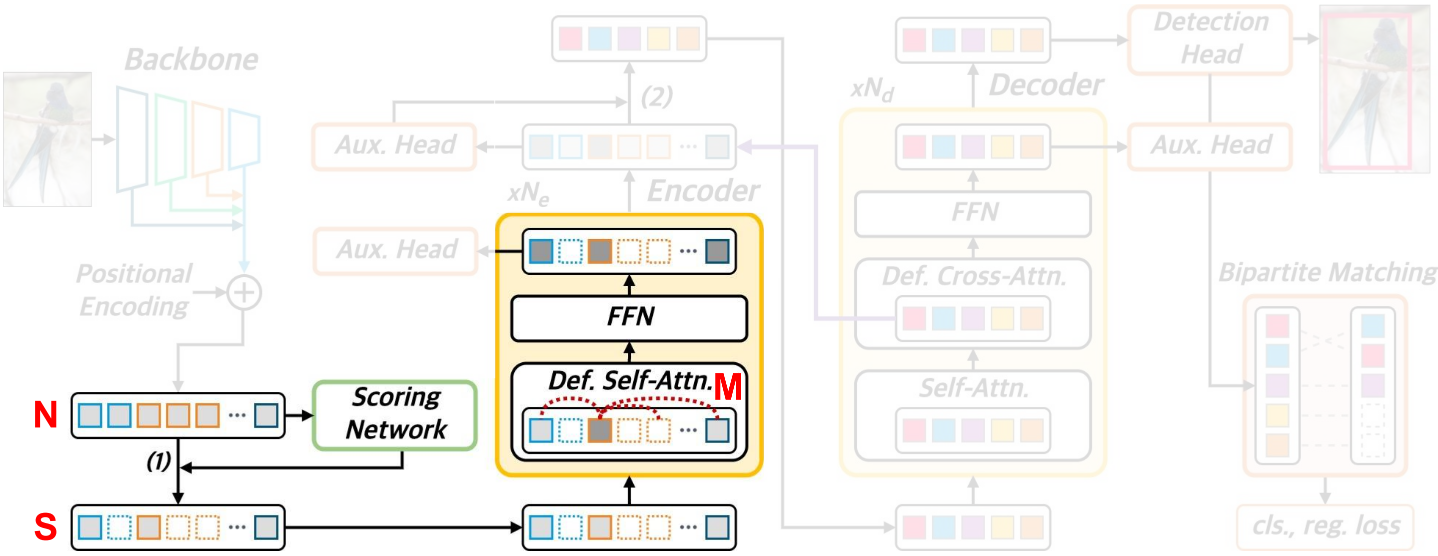
kakaobrain

Architecture



Encoder Complexity

encoder complexity	N: # of tokens, M: # of sampling points (**4**) S: # of sampled tokens in encoder (**0.1N**)

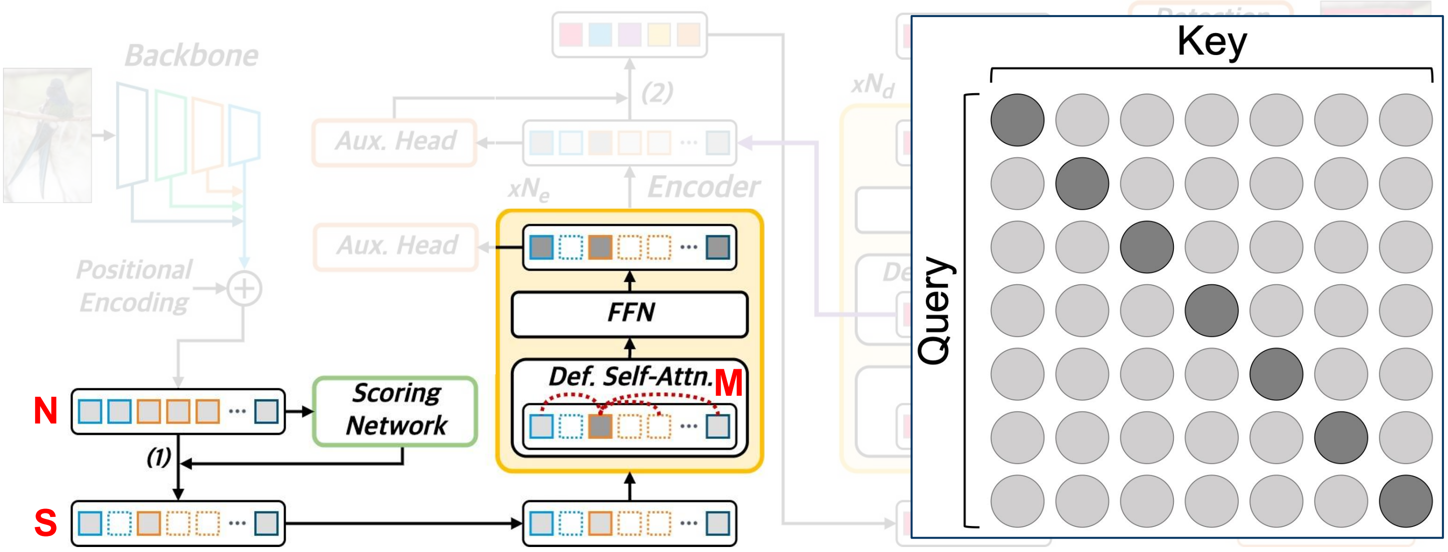


ICLR

kakaobrain

Encoder Complexity

encoder complexity	N: # of tokens, M: # of sampling points (4) S: # of sampled tokens in encoder (0.1N)
DETR	
N x N	



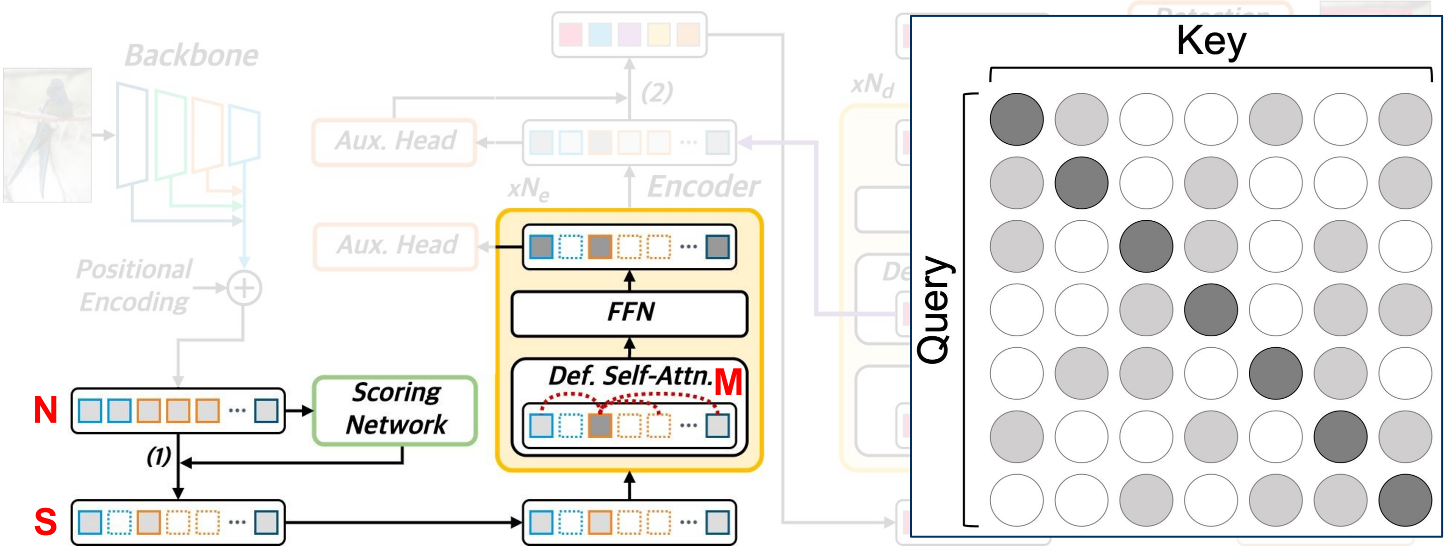
ICLR

kakaobrain

Encoder Complexity

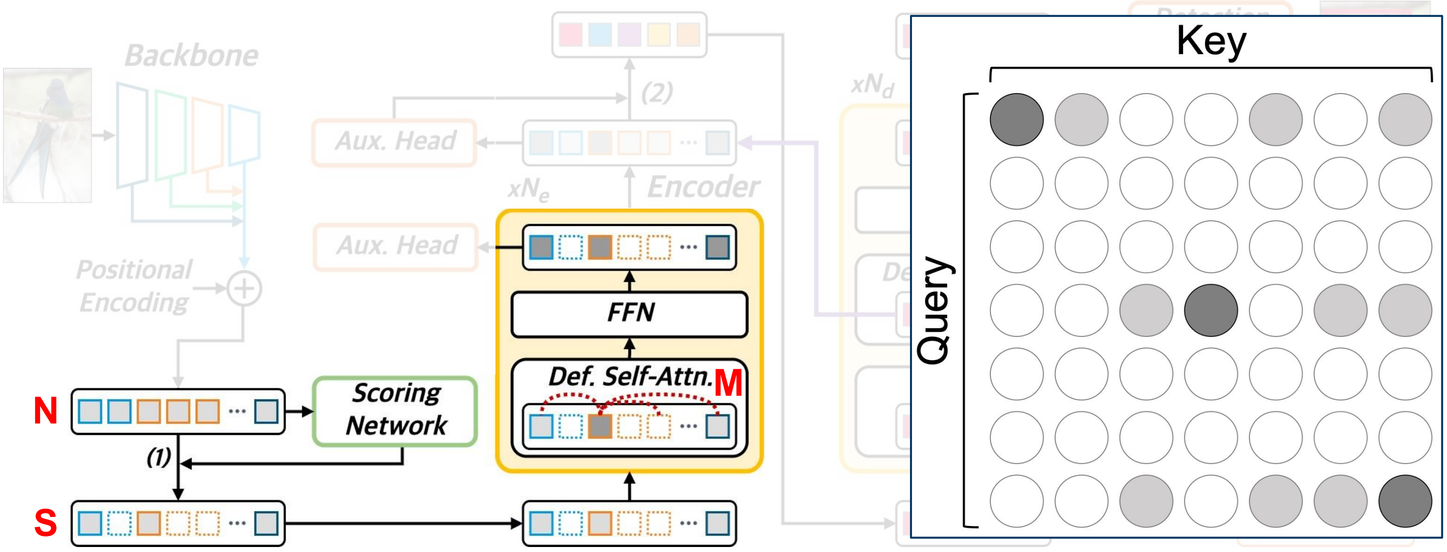
encoder complexity	N: # of tokens, M: # of sampling points (4) S: # of sampled tokens in encoder (0.1N)

DETR	Deform. DETR
N x N	N x M

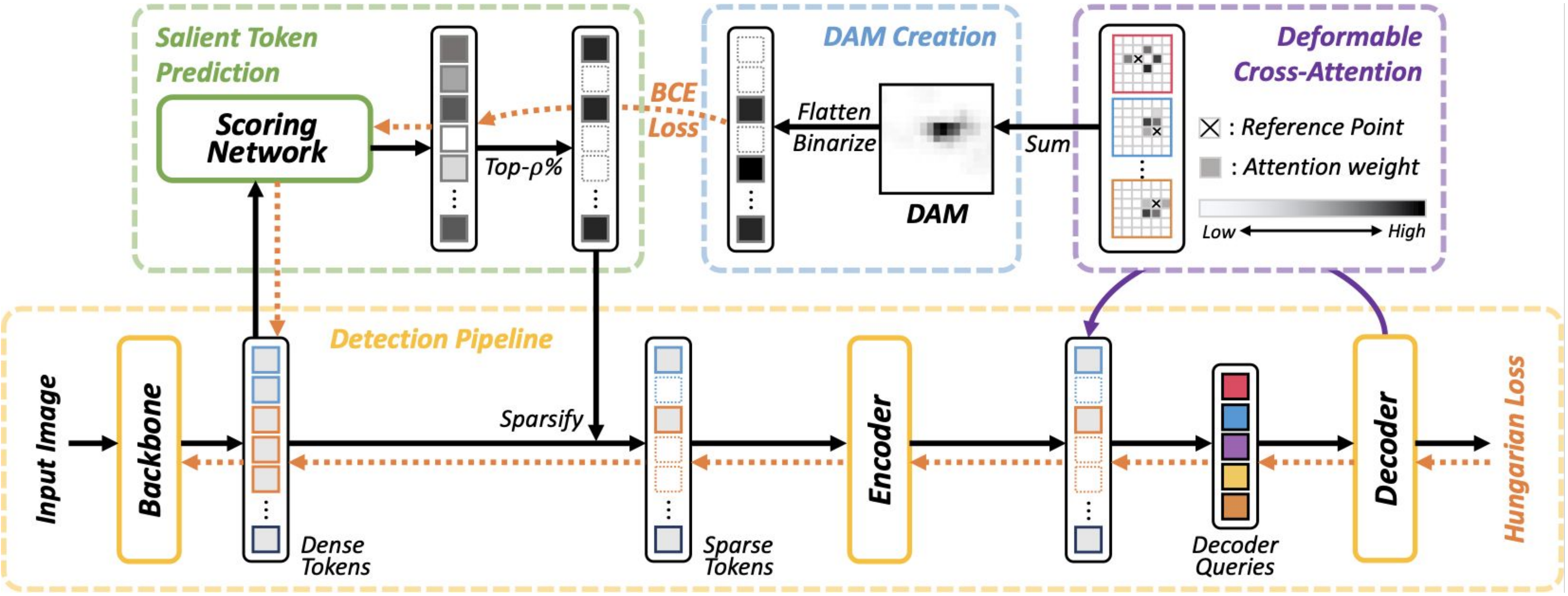


Encoder Complexity

encoder complexity	N: # of tokens, M: # of sampling points (**4**) S: # of sampled tokens in encoder (**0.1N**)	
DETR	**Deform. DETR**	**Sparse DETR**
N x N	**N x M**	**S x M**



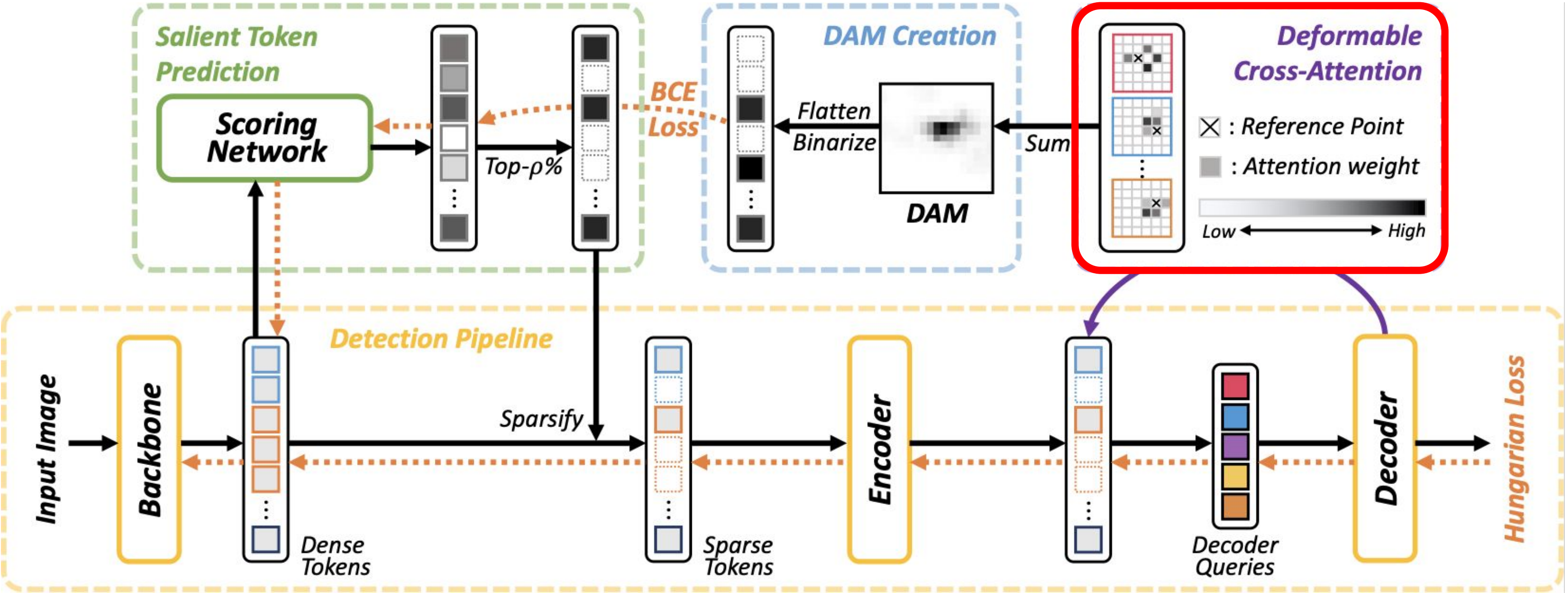
How to Train a Scoring Network



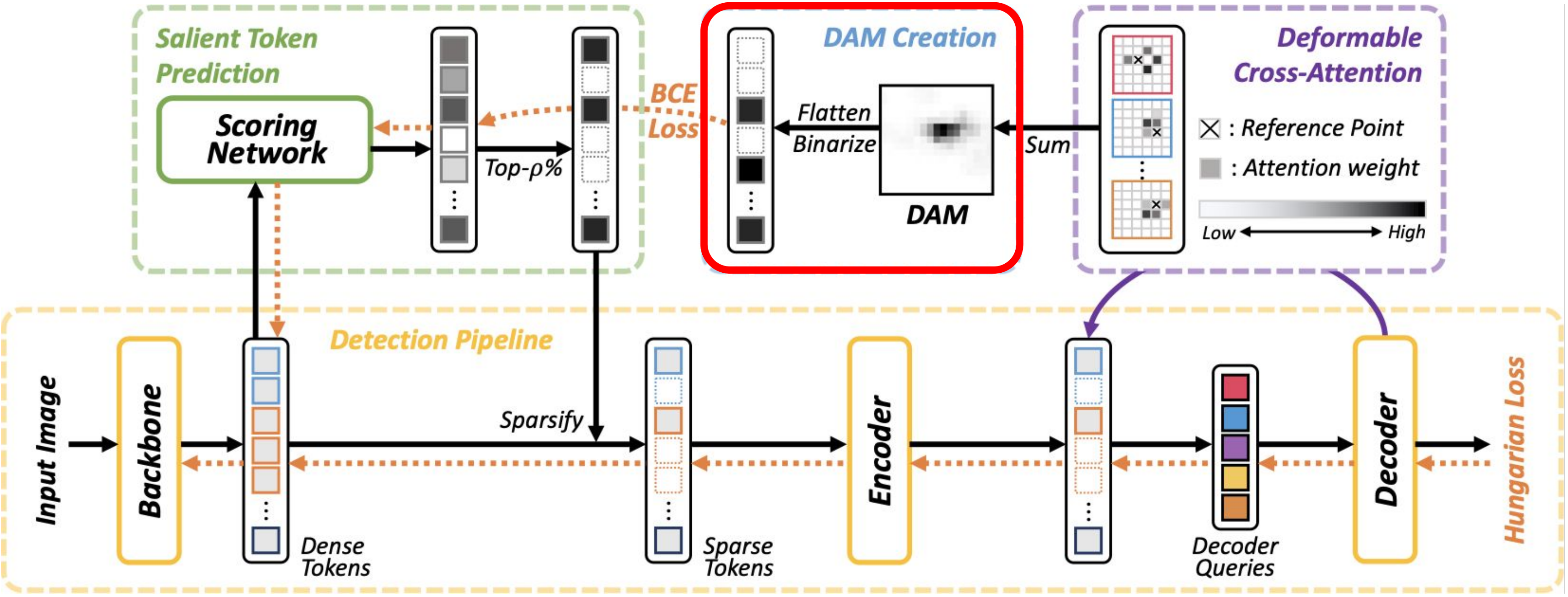
ICLR

kakaobrain

How to Train a Scoring Network



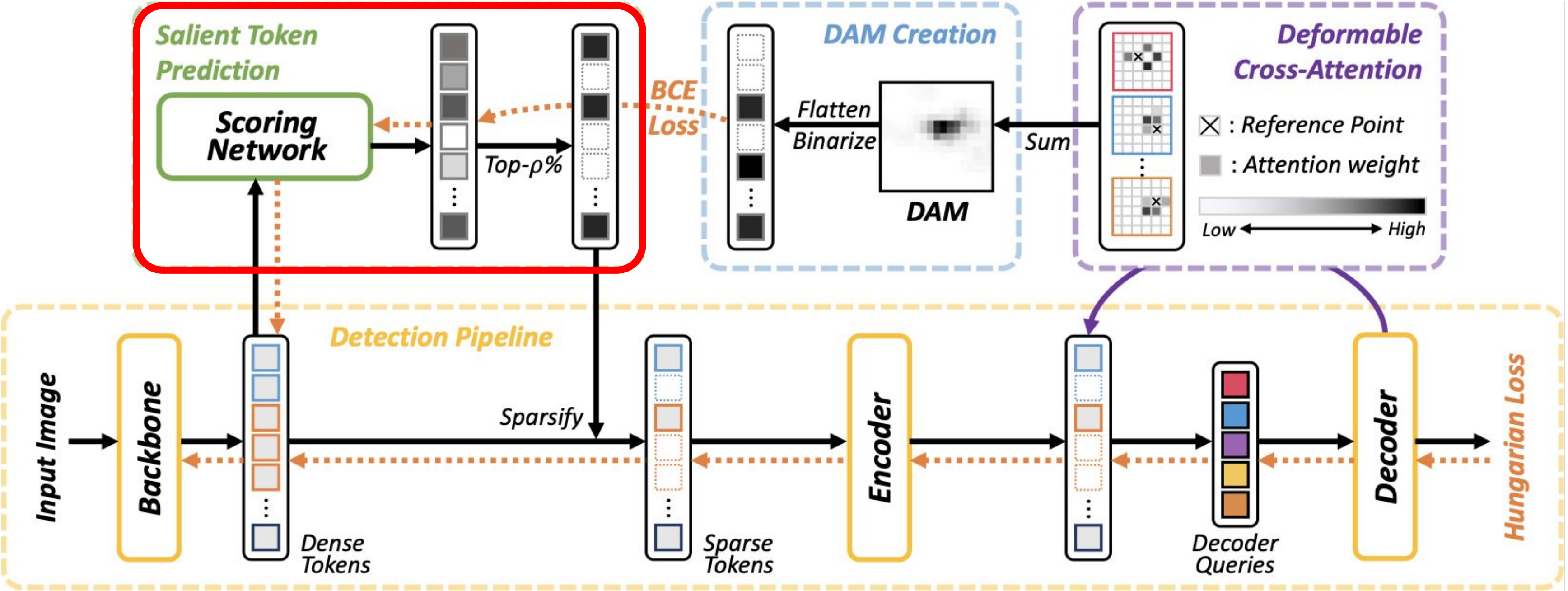
How to Train a Scoring Network



ICLR

kakaobrain

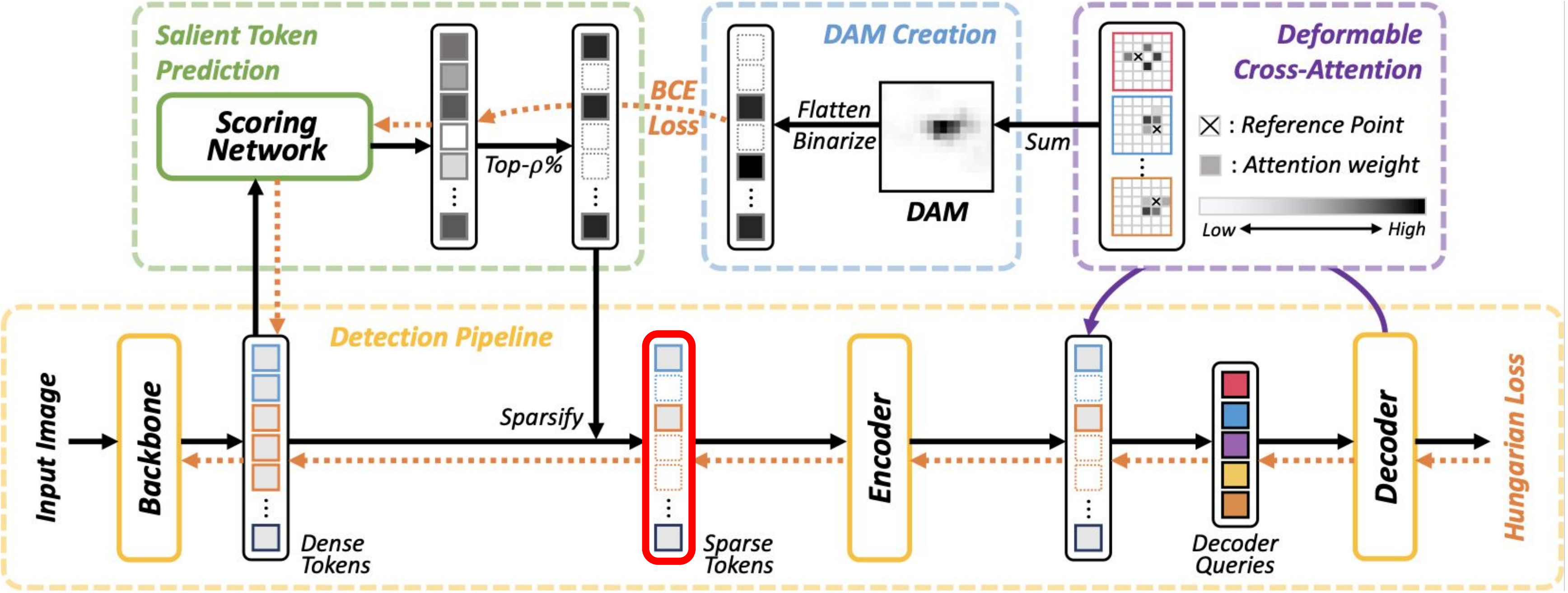
How to Train a Scoring Network



ICLR

kakaobrain

How to Train a Scoring Network



ICLR

kakaobrain

Experiments: ResNet-50

Method	Epochs	Keeping ratio (ρ)	Top-k & BBR	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$	params	FLOPs	FPS
ResNet-50 backbone:												
F-RCNN-FPN†	109	N/A		42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	26
DETR†	500	100%		42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	28
DETR-DC5†	500	100%		43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	12

Experiments: ResNet-50

Method	Epochs	Keeping ratio (ρ)	Top-k & BBR	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$	params	FLOPs	FPS
***ResNet-50** backbone:*												
F-RCNN-FPN†	109	N/A		42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	26
DETR†	500	100%		42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	28
DETR-DC5†	500	100%		43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	12
PnP-DETR‡	500	33%		41.1	61.5	43.7	20.8	44.6	60.0	-	-	-
	500	50%		41.8	62.1	44.4	21.2	45.3	60.8	-	-	-
PnP-DETR-DC5‡	500	33%		42.7	62.8	45.1	22.4	46.2	60	-	-	-
	500	50%		43.1	63.4	45.3	22.7	46.5	61.1	-	-	-

Experiments: ResNet-50

Method	Epochs	Keeping ratio (ρ)	Top-k & BBR	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	FPS
***ResNet-50** backbone:*												
F-RCNN-FPN†	109	N/A		42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	26
DETR†	500	100%		42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	28
DETR-DC5†	500	100%		43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	12
PnP-DETR‡	500	33%		41.1	61.5	43.7	20.8	44.6	60.0	-	-	-
	500	50%		41.8	62.1	44.4	21.2	45.3	60.8	-	-	-
PnP-DETR-DC5‡	500	33%		42.7	62.8	45.1	22.4	46.2	60	-	-	-
	500	50%		43.1	63.4	45.3	22.7	46.5	61.1	-	-	-
Deformable-DETR	50	100%		43.9	62.8	47.8	26.1	47.4	58.0	40M	173G	19.1
	50	100%	✓	46.0	65.2	49.8	28.2	49.1	61.0	41M	177G	18.2

ICLR

kakaobrain

Experiments: ResNet-50

Method	Epochs	Keeping ratio (ρ)	Top-k & BBR	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	FPS
***ResNet-50** backbone:*												
F-RCNN-FPN†	109	N/A		42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	26
DETR†	500	100%		42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	28
DETR-DC5†	500	100%		43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	12
PnP-DETR‡	500	33%		41.1	61.5	43.7	20.8	44.6	60.0	-	-	-
	500	50%		41.8	62.1	44.4	21.2	45.3	60.8	-	-	-
PnP-DETR-DC5‡	500	33%		42.7	62.8	45.1	22.4	46.2	60	-	-	-
	500	50%		43.1	63.4	45.3	22.7	46.5	61.1	-	-	-
Deformable-DETR	50	100%		43.9	62.8	47.8	26.1	47.4	58.0	40M	173G	19.1
	50	100%	✓	46.0	65.2	49.8	28.2	49.1	61.0	41M	177G	18.2
Sparse-DETR	50	10%	✓	45.3	65.8	49.3	28.4	48.3	60.1	41M	105G	25.3
	50	20%	✓	45.6	65.8	49.6	28.5	48.6	60.4	41M	113G	24.8
	50	30%	✓	46.0	65.9	49.7	29.1	49.1	60.6	41M	121G	23.2

AP
0.0

GFLOPs
-56 (-32%)

FPS
+5.0 (22%)

Experiments: ResNet-50

Method	Epochs	Keeping ratio (ρ)	Top-k & BBR	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	FPS
***ResNet-50** backbone:*												
F-RCNN-FPN†	109	N/A		42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	26
DETR†	500	100%		42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	28
DETR-DC5†	500	100%		43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	12
PnP-DETR‡	500	33%		41.1	61.5	43.7	20.8	44.6	60.0	-	-	-
	500	50%		41.8	62.1	44.4	21.2	45.3	60.8	-	-	-
PnP-DETR-DC5‡	500	33%		42.7	62.8	45.1	22.4	46.2	60	-	-	-
	500	50%		43.1	63.4	45.3	22.7	46.5	61.1	-	-	-
Deformable-DETR	50	100%		43.9	62.8	47.8	26.1	47.4	58.0	40M	173G	19.1
	50	100%	✓	46.0	65.2	49.8	28.2	49.1	61.0	41M	177G	18.2
Sparse-DETR	50	10%	✓	45.3	65.8	49.3	28.4	48.3	60.1	41M	105G	25.3
	50	20%	✓	45.6	65.8	49.6	28.5	48.6	60.4	41M	113G	24.8
	50	30%	✓	46.0	65.9	49.7	29.1	49.1	60.6	41M	121G	23.2
	50	40%	✓	46.2	66.0	50.3	28.7	49.0	61.4	41M	128G	21.8
	50	50%	✓	46.3	66.0	50.1	29.0	49.5	60.8	41M	136G	20.5

AP
+0.3

GFLOPs
-41 (-23%)

FPS
+2.3 (13%)

ICLR

Experiments: Swin-T

Method	Epochs	Keeping ratio (ρ)	Top-k & BBR	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	FPS
***Swin-T** backbone:*												
DETR	500	100%		45.4	66.2	48.1	22.9	49.5	65.9	45M	92G	26.8
Deformable-DETR	50	100%		45.7	65.3	49.9	26.9	49.4	61.2	40M	180G	15.9
	50	100%	✓	48.0	68.0	52.0	30.3	51.4	63.7	41M	185G	15.4
Sparse-DETR	50	10%	✓	48.2	69.2	52.3	29.8	51.2	64.5	41M	113G	21.2

AP
+0.2

GFLOPs
-72 (-39%)

FPS
+5.8 (38%)

ICLR

kakaobrain

Experiments: Swin-T

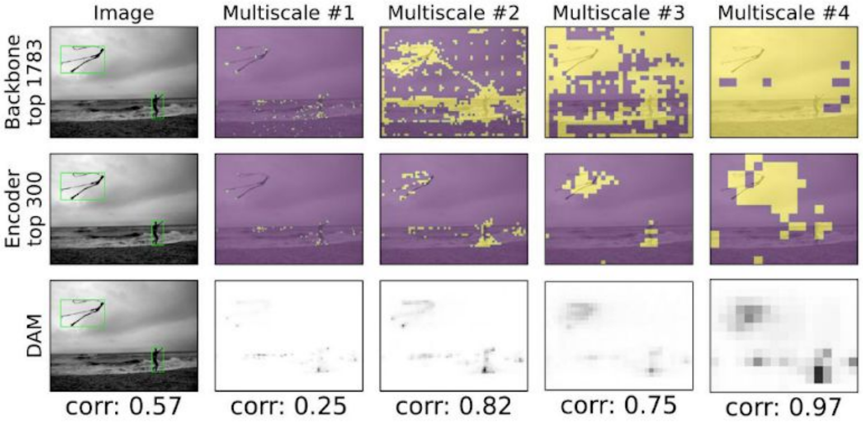
Method	Epochs	Keeping ratio (ρ)	Top-k & BBR	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	FPS
***Swin-T** backbone:*												
DETR	500	100%		45.4	66.2	48.1	22.9	49.5	65.9	45M	92G	26.8
Deformable-DETR	50	100%		45.7	65.3	49.9	26.9	49.4	61.2	40M	180G	15.9
	50	100%	✓	48.0	68.0	52.0	30.3	51.4	63.7	41M	185G	15.4
Sparse-DETR	50	10%	✓	48.2	69.2	52.3	29.8	51.2	64.5	41M	113G	21.2
	50	20%	✓	48.8	69.4	53.0	30.4	51.9	64.8	41M	121G	20.0
	50	30%	✓	49.1	69.5	53.5	31.4	52.5	65.1	41M	129G	18.9
	50	40%	✓	49.2	69.5	53.5	31.4	52.9	64.8	41M	136G	18.0
	50	50%	✓	49.3	69.5	53.3	32.0	52.7	64.9	41M	144G	17.2

AP	GFLOPs	FPS
+1.3	-41 (-22%)	+1.8 (12%)

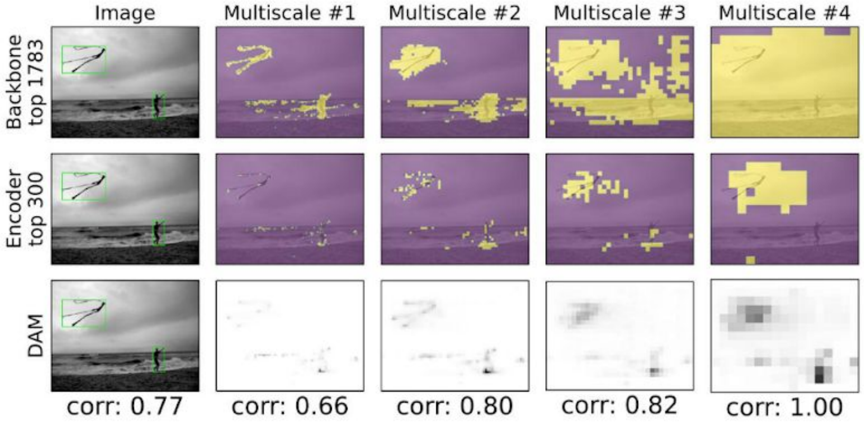
ICLR

kakaobrain

Visualization



(a) OS-based model ($\rho = 0.1$)

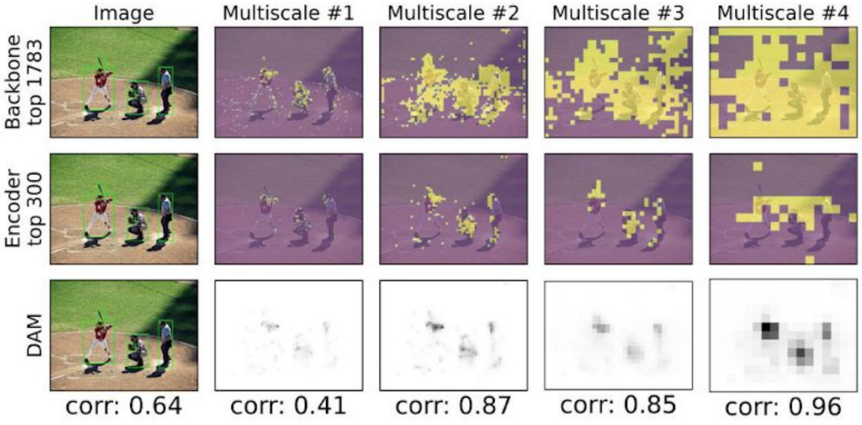


(b) DAM-based model ($\rho = 0.1$)

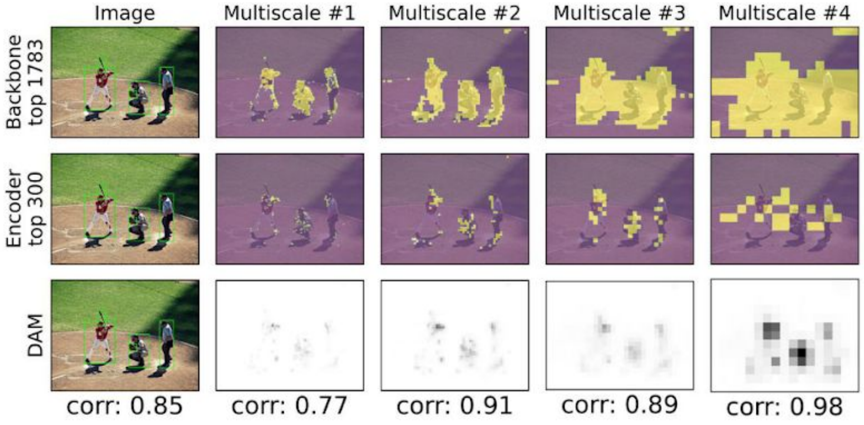
ICLR

kakaobrain

Visualization



(a) OS-based model ($\rho = 0.1$)



(b) DAM-based model ($\rho = 0.1$)

Conclusion

- We propose **the encoder token sparsification method**, which lightens the attention complexity in the encoder.

Conclusion

- We propose **the encoder token sparsification method**, which lightens the attention complexity in the encoder.
- We propose **novel sparsification criteria to sample the informative subset** from the entire token set: *Decoder cross-Attention Map* (DAM)

Conclusion

- We propose **the encoder token sparsification method**, which lightens the attention complexity in the encoder.
- We propose **novel sparsification criteria to sample the informative subset** from the entire token set: *Decoder cross-Attention Map* (DAM)
- Sparse DETR **outperforms the Deformable DETR** even when **using only 10% of the encoder token**, and decreases the overall computation by 38%

ICLR

kakaobrain

Code & models are available now.

https://github.com/kakaobrain/sparse-detr

More experiments and ablation studies can be found in the paper

ICLR

kakaobrain