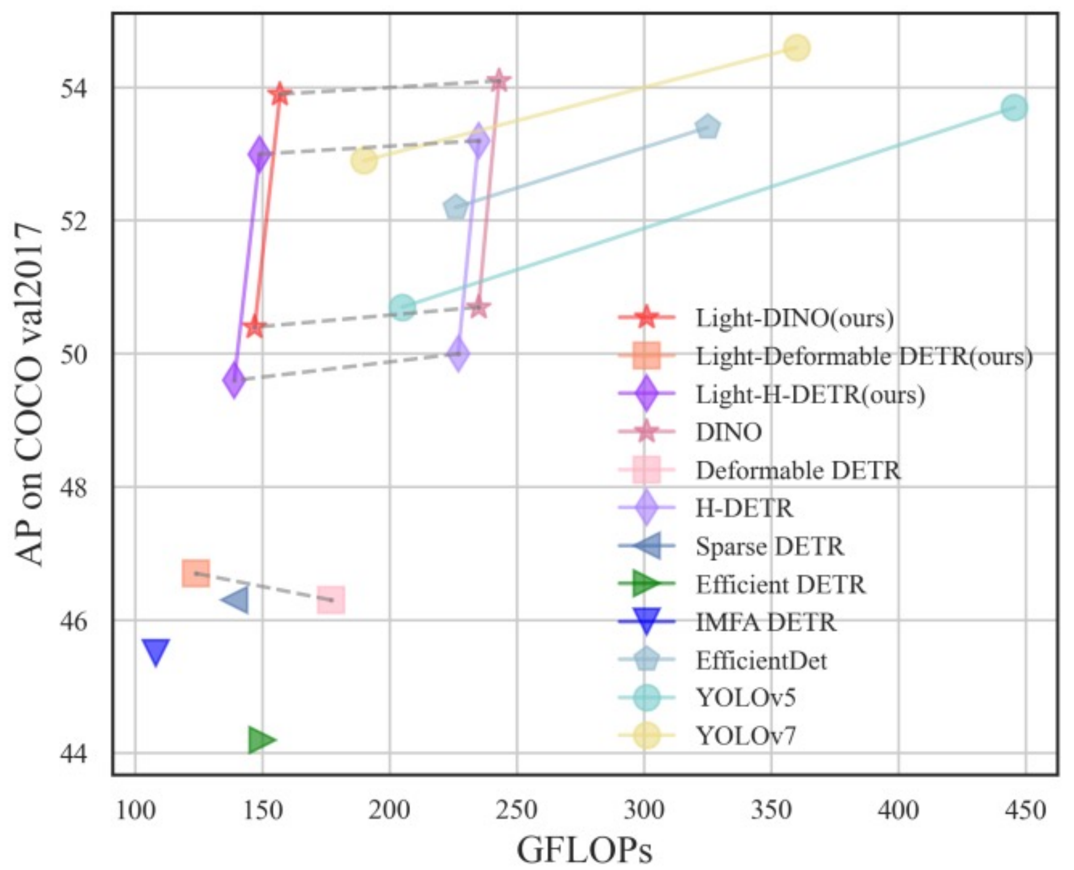
Lite DETR : An Interleaved Multi-Scale Encoder for Efficient DETR

Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, Lionel M. Ni.

Highlights:

- Encoder in DETR-like models that use multi-scale features account for **75%** of the total computational cost because of **excessive tokens in high-resolution feature maps**.
- We propose to update high-resolution and low-resolution maps in an **interleaved way** to save cost.
- We also propose **a key-aware deformable attention** to predict more reliable weights.
- Efficient encoder design to reduce computational cost
  - **Simple**. Dozens of lines code change (if not consider pluggable key-aware attention).
  - **Effective**. Reduce encoder cost by 50% while preserve most of the original performance.
  - **General**. Validated on a series of DETR models (Deformable DETR, H-DETR, DINO).



The reason of high computational cost: encoder

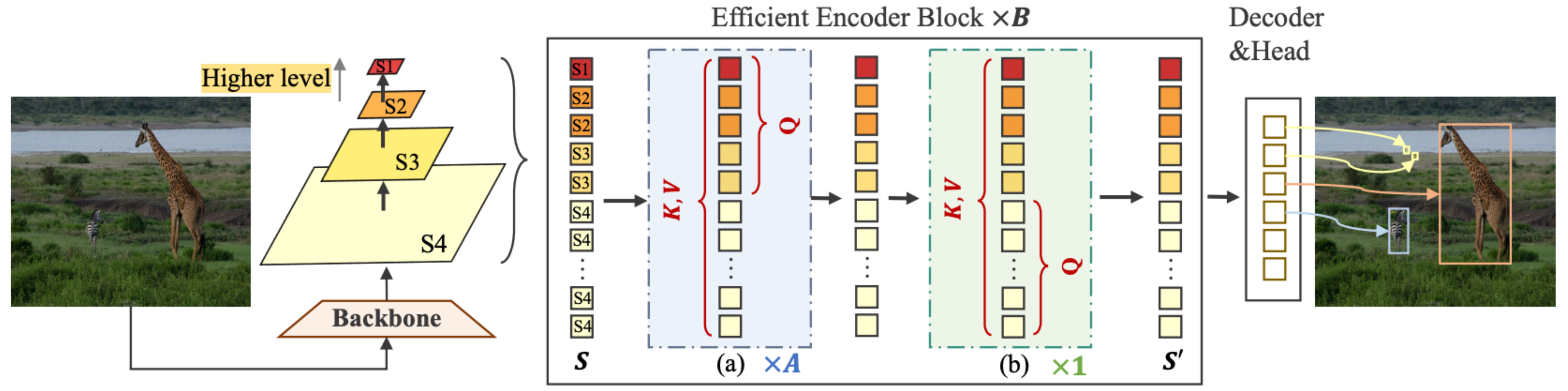
Model	Total GFLOPs	Backbone	(M.S.) Encoder	Decoder	Total Train Mem	AP	AP_s	AP_L
DINO-4scale (100%)	235	70	**137**	28	32G	50.7	**33.5**	64.7
DINO-3scale (25%)	122	70	31	21	13G	48.2	**30.1**	63.9

The reason of high encoder cost: high-resolution maps

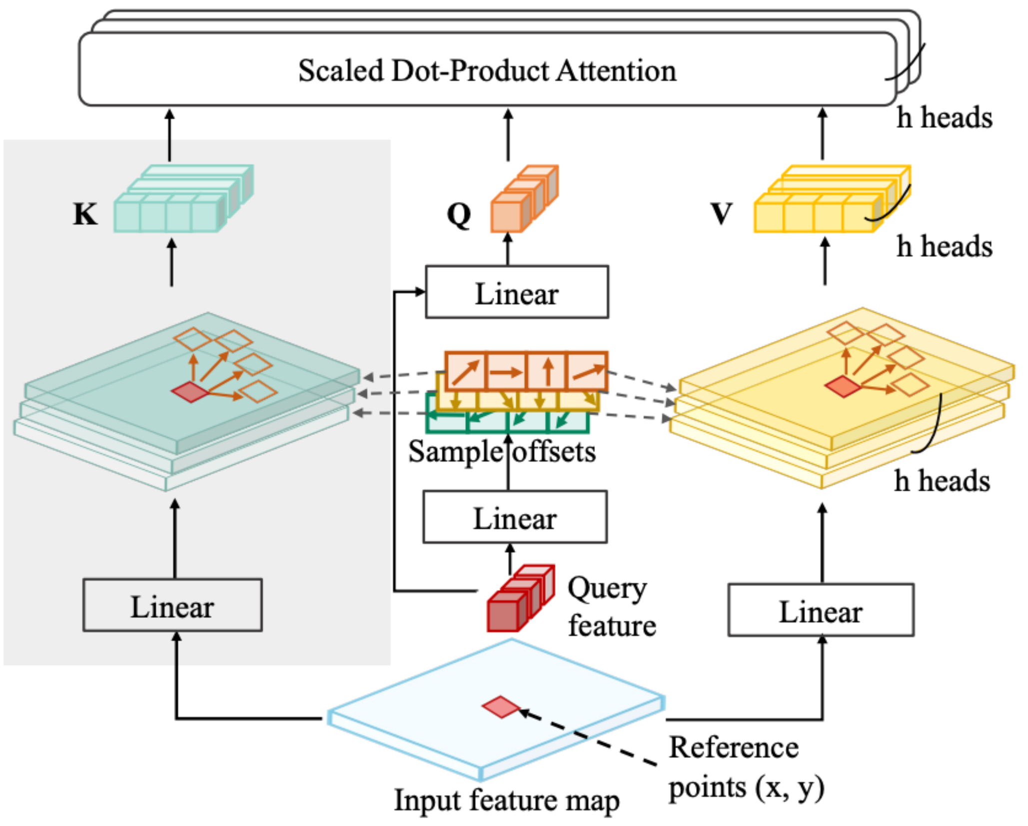
Feature Scale (S)	S1 ($\frac{1}{64}$)	S2 ($\frac{1}{32}$)	S3 ($\frac{1}{16}$)	S4 ($\frac{1}{8}$)
Token Ratio	1.17%	4.71%	18.8%	75.3%

Model design:

- Interleaved update:
  - **Excessive high-resolution features**, and most of which are not informative but contain local details for small objects
  - Update **high-resolution** features at lower frequency



- **Key-aware** deformable attention
  - Deformable attention: regress attention weights directly from query
  - We calculate attention weights with sampled query and key.



Results:

- Apply to Deformable DETR

Model	#epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	GFLOPs	Encoder GFLOPs	Params
DETR-DC5 [1]	500	43.3	63.1	45.9	22.5	47.3	61.1	187	100	41M
Anchor DETR-DC5 [32]	50	44.2	64.7	47.5	24.7	48.2	60.6	151	70	39M
Conditional DETR-DC5 [21]	50	43.8	64.4	46.7	24.0	47.6	60.7	195	100	44M
DAB-DETR-DC5 [18]	50	44.5	65.1	47.7	25.3	48.2	62.3	202	100	44M
DN-DETR-DC5	50	46.3	66.4	49.7	26.7	50.0	64.3	202	100	44M
Deformable DETR efficient variants										
Deformable DETR† [37]	50	46.8	66.0	50.6	29.8	49.7	62.0	177	90	40M
Lite-Deformable DETR H2L2-(2+1)x3(5%, ours)	50	45.8	65.1	49.3	27.7	49.1	61.1	108	**23(↓ 74%)**	41M
Lite-Deformable DETR H3L1-(6+1)x1(25%, ours)	50	45.9	65.6	49.2	27.9	49.0	61.6	115	**30(↓ 66%)**	41M
Lite-Deformable DETR H3L1-(3+1)x2(25%, ours)	50	46.2	65.5	49.8	28.2	49.2	61.5	119	**35(↓ 61%)**	41M
Lite-Deformable DETR H3L1-(2+1)x3(25%, ours)	50	**46.7**	66.1	50.6	29.1	49.7	62.2	123	**39(↓ 57%)**	41M
Efficient DETR [33]	50	44.2	62.2	48.0	28.4	47.5	56.6	159	79	32M
Sparse DETR*-rho-0.1 [37]	50	45.3	65.8	49.3	28.4	48.3	60.1	111	24	41M
Sparse DETR*-rho-0.2 [37]	50	45.6	65.8	49.6	28.5	48.6	60.4	119	32	41M
Sparse DETR*-rho-0.3 [37]	50	46.0	65.9	49.7	29.1	49.1	60.6	127	40	41M
Sparse DETR*-rho-0.5 [37]	50	46.3	66.0	50.1	29.0	49.5	60.8	141	54	41M

- Apply to DINO and H-DETR

Model	#epochs	AP	AP_{36}	AP_{75}	AP_S	AP_M	AP_L	GFLOPs	Encoder GFLOPs	Params
EfficientDet-D6 [29]	–	51.3	–	–	–	–	–	226	–	52M
YOLOv5-X [12]	–	50.7	–	–	–	–	–	206	–	87M
YOLOv7-X [30]	–	52.9	–	–	–	–	–	190	–	71M
Swin-T backbone										
VIDT+ [27]	50	49.7	67.7	54.2	31.6	53.4	65.9	–	–	38M
D^2ETR [15]	50	49.1	–	–	–	–	–	127	–	46M
DINO [36]	36	54.1	72.0	59.3	38.3	57.3	68.6	243	137	47M
Lite-DINO H2L2-(2+1)x3(5%, ours)	36	53.1	71.4	57.9	36.6	56.0	68.8	138	**30(↓78%)**	47M
Lite-DINO H3L1-(6+1)x1(25%, ours)	36	53.3	71.7	58.2	36.3	56.6	68.7	149	**41(↓70%)**	47M
Lite-DINO H3L1-(2+1)x3(25%, ours)	36	**53.9**	72.0	58.8	37.9	57.0	69.1	159	**53(↓62%)**	47M
H-DETR [11]	36	53.2	71.5	58.2	35.9	56.4	68.2	234	137	47M
Lite-H-DETR H2L2-(2+1)x3(5%, ours)	36	52.3	70.7	57.2	35.9	55.2	67.7	131	30	47M
Lite-H-DETR H3L1-(6+1)x1(25%, ours)	36	52.7	71.5	58.3	35.6	56.0	68.0	142	41	47M
Lite-H-DETR H3L1-(2+1)x3(25%, ours)	36	**53.0**	71.3	58.2	36.3	56.3	68.1	152	53	47M
ResNet-50 backbone										
DFFT [3]	36	46.0	–	–	–	–	–	101	18	–
PnP-DETR [31]	36	43.1	63.4	45.3	22.7	46.5	61.1	104	29	–
AdaMixer [8]	36	47.0	66.0	51.1	30.1	50.2	61.8	132	–	135M
IMFA-DETR [35]	36	45.5	45.0	49.3	27.3	48.3	61.6	108	≈ 20	53M
DINO [36]	36	50.7	68.6	55.4	33.5	54.0	64.8	235	137	47M
Lite-DINO H2L2-(2+1)x3 (ours)	36	49.9	68.2	54.6	32.3	52.9	64.7	130	30	47M
Lite-DINO H3L1-(6+1)x1(ours)	36	50.2	68.6	54.3	33.0	53.4	66.0	141	41	47M
Lite-DINO H3L1-(2+1)x3(ours)	36	**50.4**	68.5	54.6	33.5	53.6	65.5	151	53	47M
H-DETR [11]	36	50.0	68.3	54.4	32.9	52.7	65.3	226	137	47M
Lite-H-DETR H3L1-(2+1)x3 (ours)	36	49.5	67.6	53.9	32.0	52.8	64.0	142	53	47M

- Key-aware deformable attention visualization

(a) Attention map of a single query on all scales (b) Attention map of all queries on scale S3 (c) Attention map of all queries on scale S4