KOREA UNIVERSITY

고급 컴퓨터 비전 Paper Review

End-to-End Object Detection with Transformer (DETR)

ECCV 2020

DSBA
Data Science & Business Analytics

Industrial & Management Engineering

Data Science & Business Analytics Lab

2019021157 이유경

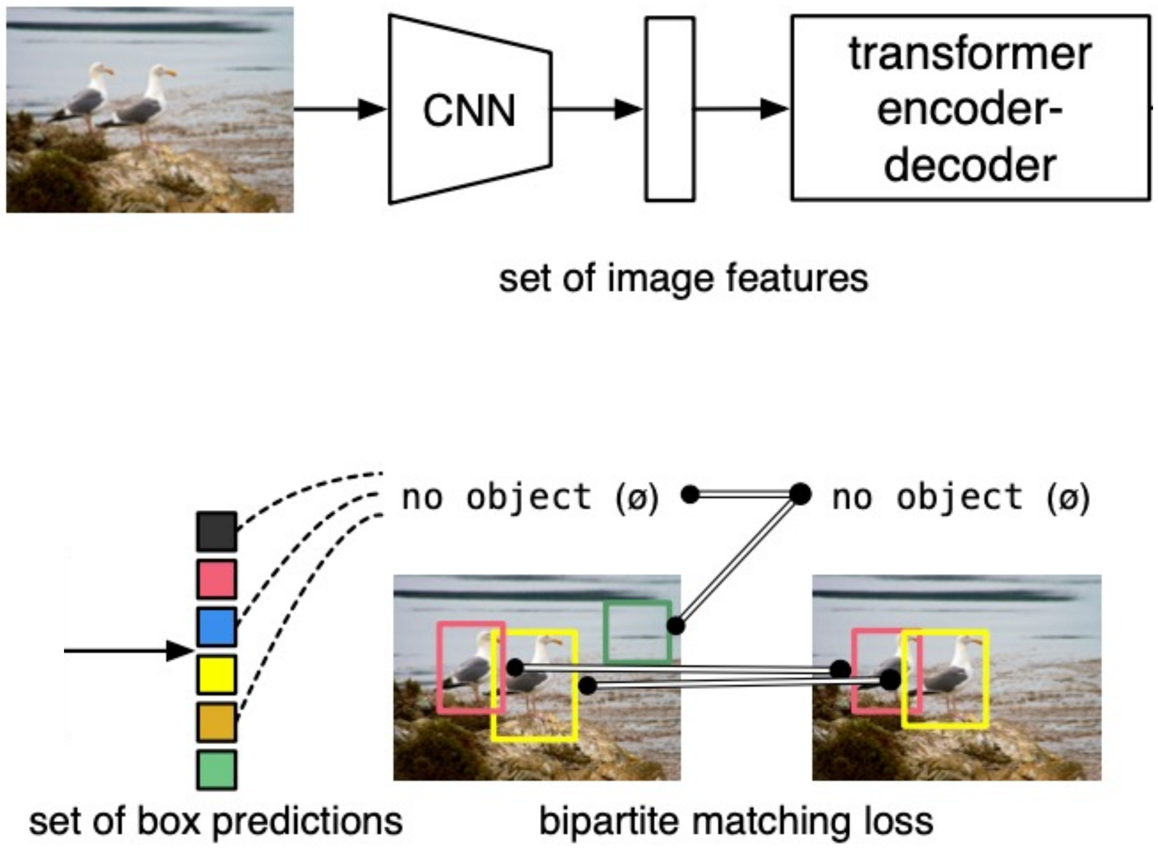
00

발표 목차

Contents

01 DETR

Overview

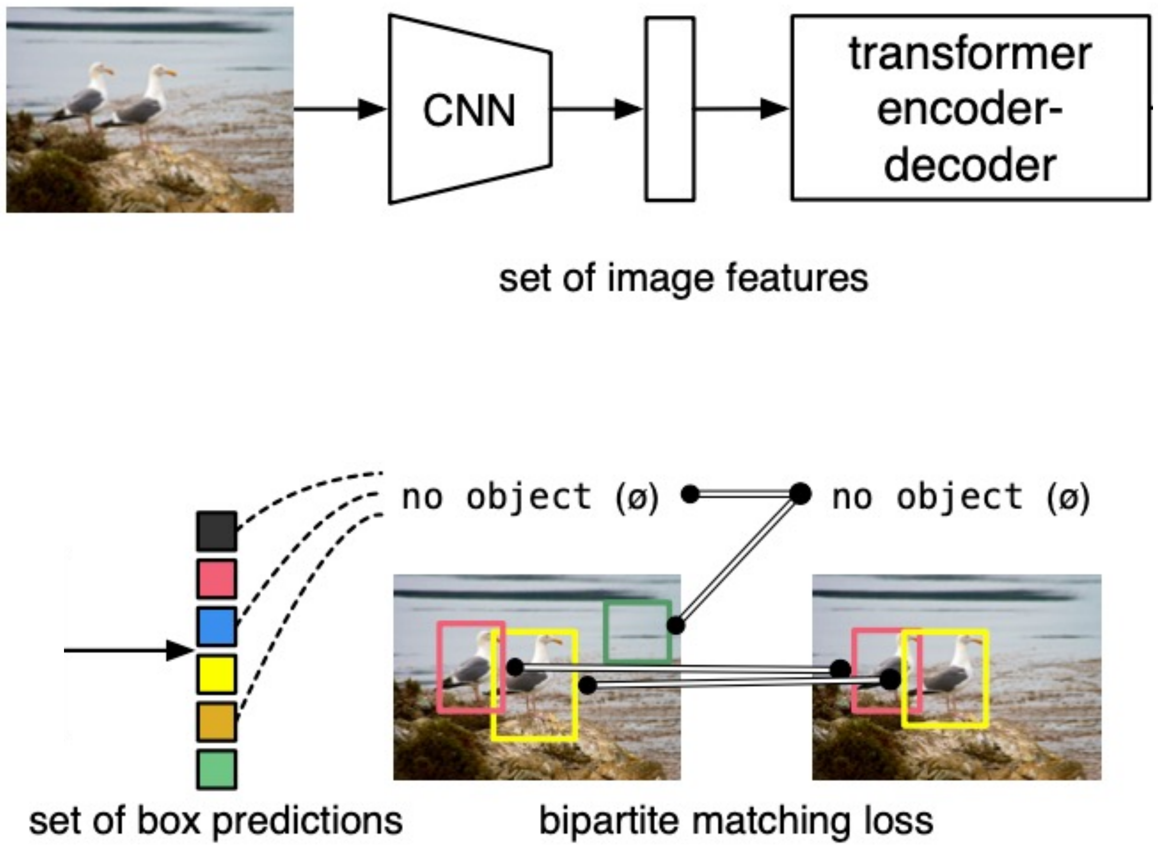


TL;DR

- A new detection model based on Transformer using bipartite-matching
- Approaching object detection as a direct set-prediction problem
- It does not use hand-designed components (RPN, NMS)

01 DETR

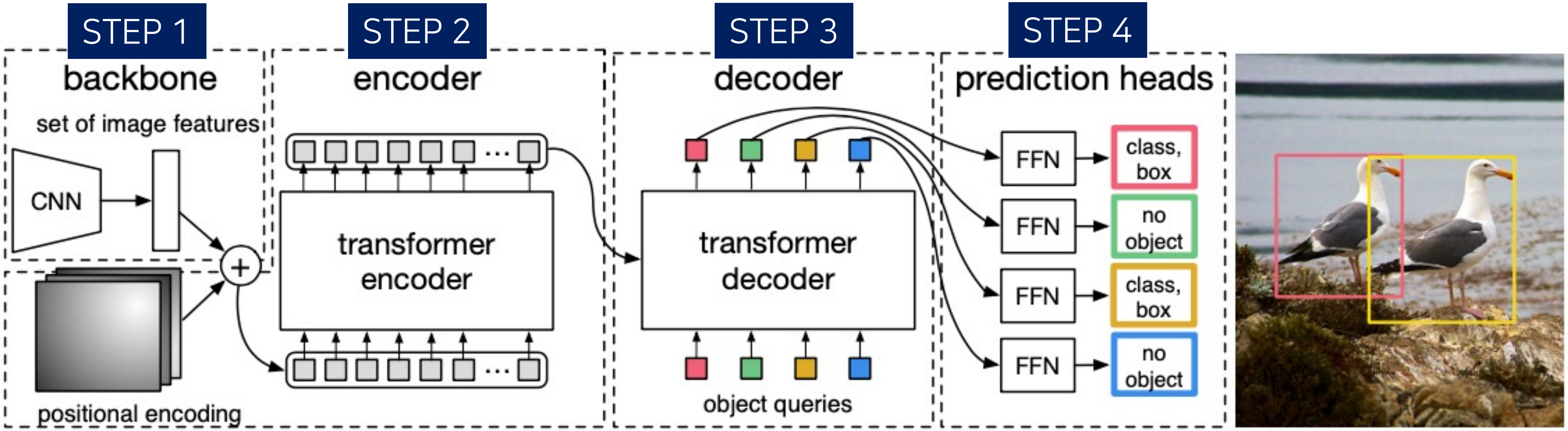
Overview



- **Main Contribution**
  - Unlike CNN-based models, global information can be learned using a Transformer.
  - Input: Image features are extracted using backbone CNN without using the image directly to the transformer
  - Model: Effective model without complex preprocessing & post processing
  - Performance is significantly better than Faster R-CNN (especially Large object detection)
  - The implementation code is very simple

DETR

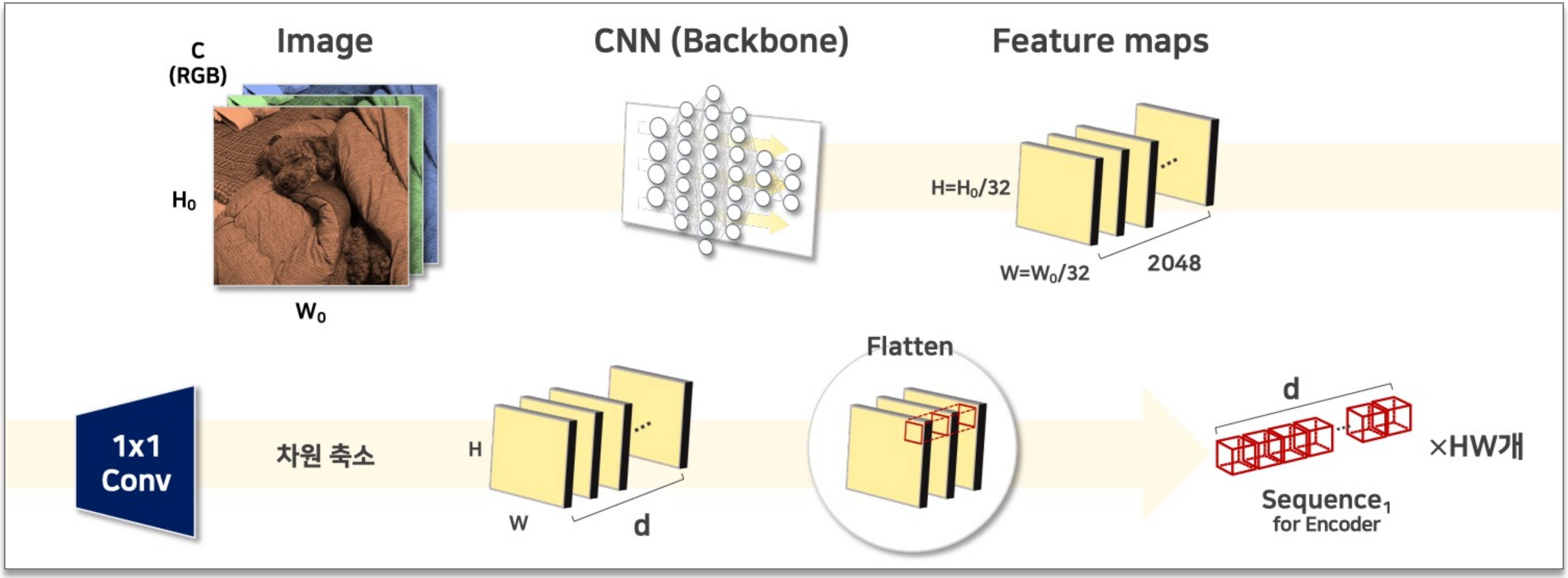
Method



STEP 1	STEP 2	STEP 3	STEP 4
CNN (ResNet)	Transformer encoder	Transformer decoder	Prediction heads
:Get image feature maps	:Learn pixel level relation	:Image to object query	:Predict class, bbox (bipartite matching)

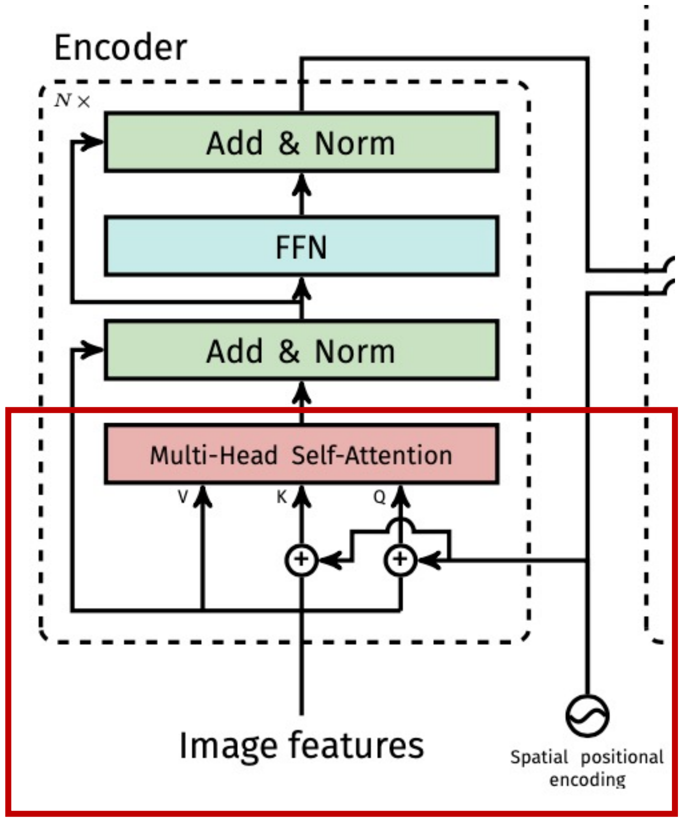
image → CNN (ResNet) → Transformer encoder → Transformer decoder → Prediction heads → output

Step 1: CNN (Get image feature maps)



Reference : http://dsba.korea.ac.kr/seminar/?mod=document&uid=1784

Step 2: Transformer Encoder (Learn pixel level relation)

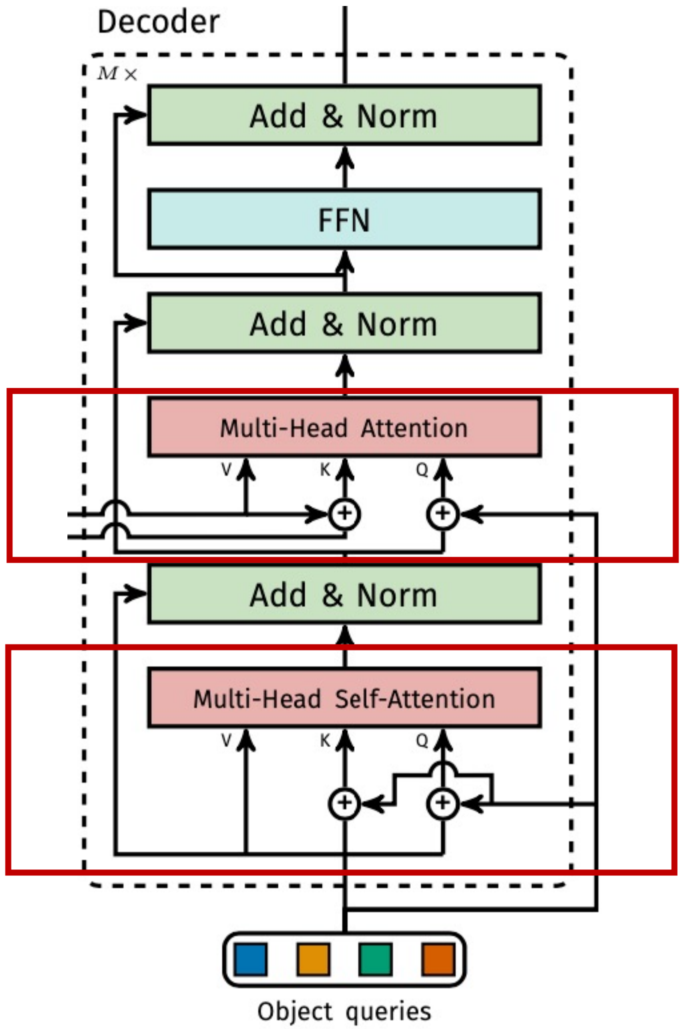


raw image → image feature → input sequence (# of pixels)

+ spatial positional encoding
(query, key / all encoder layer)

02 DETR
Method

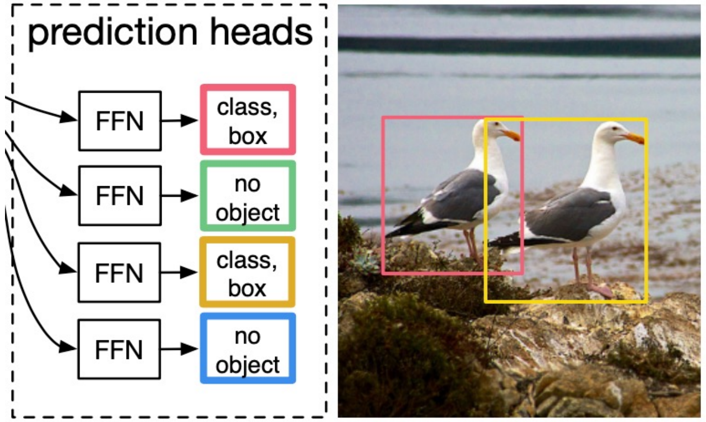
Step 3: Transformer Decoder (Image to object query)



encoder-decoder cross attention : Relation between images and object queries

decoder self attention : Relation between object queries

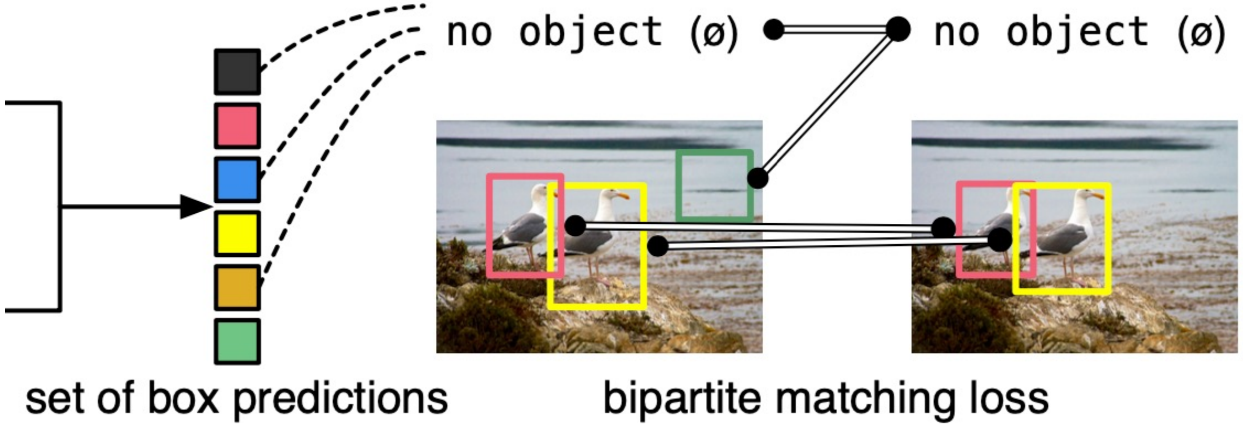
Step 4: Prediction heads (Predict class, bbox)



$$\mathcal{L}_{match}(y_i, \hat{y}_{\sigma(i)}) = \underbrace{-\mathbb{1}_{\{c_i \neq \varnothing\}} \hat{p}_{\sigma(i)}(c_i)}_{\text{predict class}} + \underbrace{\mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\sigma(i)})}_{\text{predict bbox}}$$

$$\Downarrow$$

$$\mathcal{L}_{\text{Hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{box}}(b_i, \hat{b}_{\hat{\sigma}}(i)) \right]$$



03 DETR

Result

- DETR vs Faster R-CNN

COCO validation dataset

Model	GFLOPS/FPS	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN-DC5	320/16	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster RCNN-FPN	180/26	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-R101-FPN	246/20	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster RCNN-DC5+	320/16	166M	41.1	61.4	44.3	22.9	45.9	55.0
Faster RCNN-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-R101-FPN+	246/20	60M	44.0	63.9	**47.8**	**27.2**	48.1	56.0
DETR	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1
DETR-R101	152/20	60M	43.5	63.8	46.4	21.9	48.0	61.8
DETR-DC5-R101	253/10	60M	**44.9**	**64.7**	47.7	23.7	**49.5**	**62.3**

- Large object : DETR > Faster R-CNN
- Small object : DETR < Faster R-CNN

03 DETR

Result

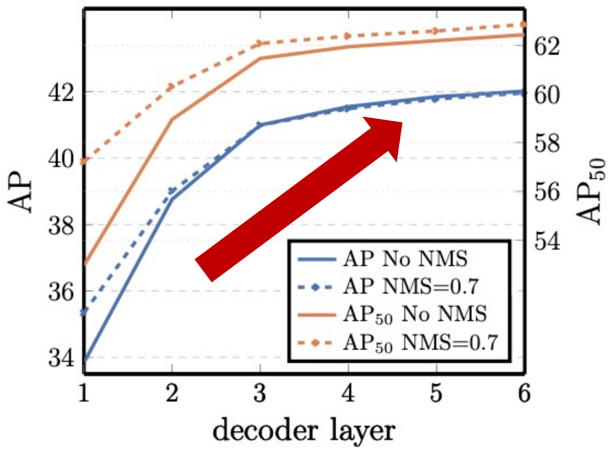
- Ablation1: Encoder

Table 2: Effect of encoder size. Each row corresponds to a model with varied number of encoder layers and fixed number of decoder layers. Performance gradually improves with more encoder layers.

#layers	GFLOPS/FPS	#params	AP	AP_{50}	AP_S	AP_M	AP_L
0	76/28	33.4M	36.7	57.4	16.8	39.6	54.2
3	81/25	37.4M	40.1	60.6	18.5	43.8	58.6
6	86/23	41.3M	40.6	61.6	19.9	44.3	60.2
12	95/20	49.2M	41.6	62.1	19.8	44.9	61.9

Performance gradually improves with more encoder layer

- Ablation2: Decoder



- Performance gradually improves with more decoder layer
- No effect of Non-maximum suppression

03 DETR

Result

- Panoptic segmentation

Model	Backbone	PQ	SQ	RQ	PQ^{th}	SQ^{th}	RQ^{th}	PQ^{st}	SQ^{st}	RQ^{st}	AP
PanopticFPN++	R50	42.4	79.3	51.6	49.2	82.4	58.8	32.3	74.8	40.6	37.7
UPSnet	R50	42.5	78.0	52.5	48.6	79.4	59.6	33.4	75.9	41.7	34.3
UPSnet-M	R50	43.0	79.1	52.8	48.9	79.7	59.7	34.1	78.2	42.3	34.3
PanopticFPN++	R101	44.1	79.5	53.3	**51.0**	**83.2**	60.6	33.6	74.0	42.1	**39.7**
DETR	R50	43.4	79.3	53.8	48.2	79.8	59.5	36.3	78.5	45.3	31.1
DETR-DC5	R50	44.6	79.8	55.0	49.4	80.5	60.6	**37.3**	**78.7**	**46.5**	31.9
DETR-R101	R101	**45.1**	**79.9**	**55.5**	50.5	80.9	**61.7**	37.0	78.5	46.0	33.0

- DETR achieved good performance not only in object detection but also in panoptic segmentation

EOD