Deformable Detr

Paper Details

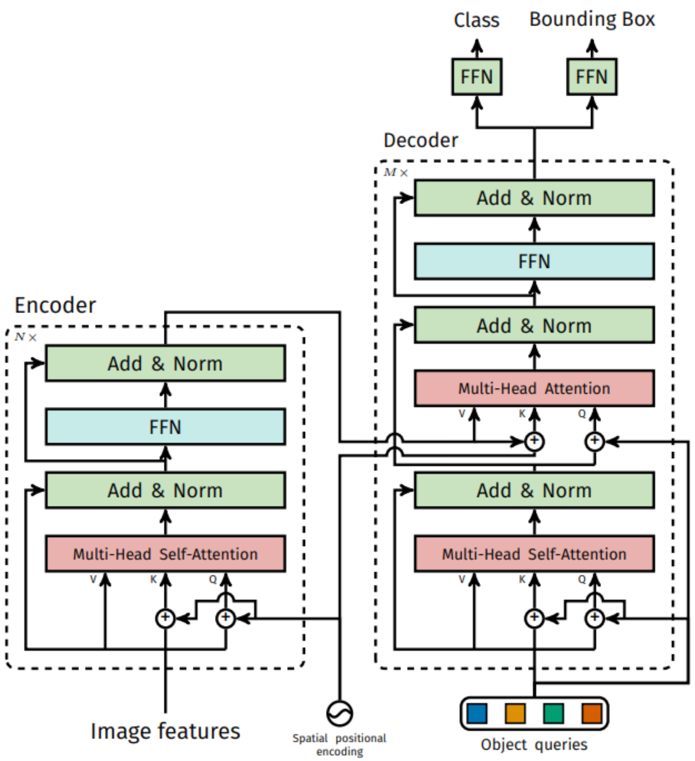
- Paper Title: DEFORMABLE DETR: DEFORMABLE TRANSFORMERS FOR END-TO-END OBJECT DETECTION
- Publication Date: 18 Mar 2021
- Publisher: Xizhou Zhu, Weijie Su
- Affiliation: SenseTime Research, University of Science and Technology of China, The Chinese University of Hong Kong
- Conderence: ICLR

Motivation

- Drawback DETR
  - slow convergence
  - high complexity
  - low performance in detecting small objects
- Effiecient Attention Mechanism
  - use pre-defined sparse attention
  - learn data-dependent sparse attention
  - explore the low-rank property in self-attention
- Multi-scale Feature Representation for Object Detection

Motivation

- Drawback DETR
  - slow convergence



Motivation

- Drawback DETR
  - slow convergence
  - high complexity

$$O(N_q C^2 + N_k C^2 + N_q N_k C)$$

Complexity of Multihead attention

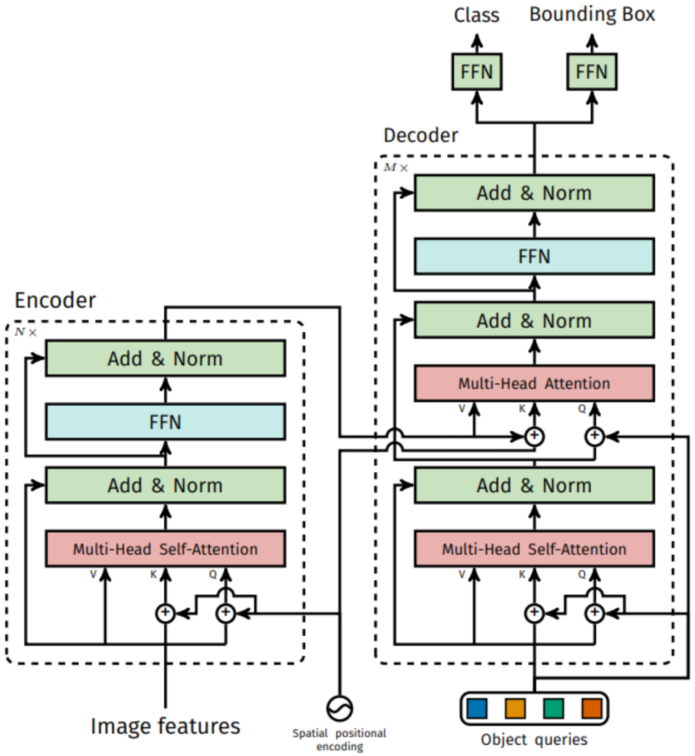
$$O(N_q N_k C)$$

Motivation

- Drawback DETR
  - slow convergence
  - high complexity

$$O(N_qC^2 + N_kC^2 + N_qN_kC)$$ Complexity of Multihead attention

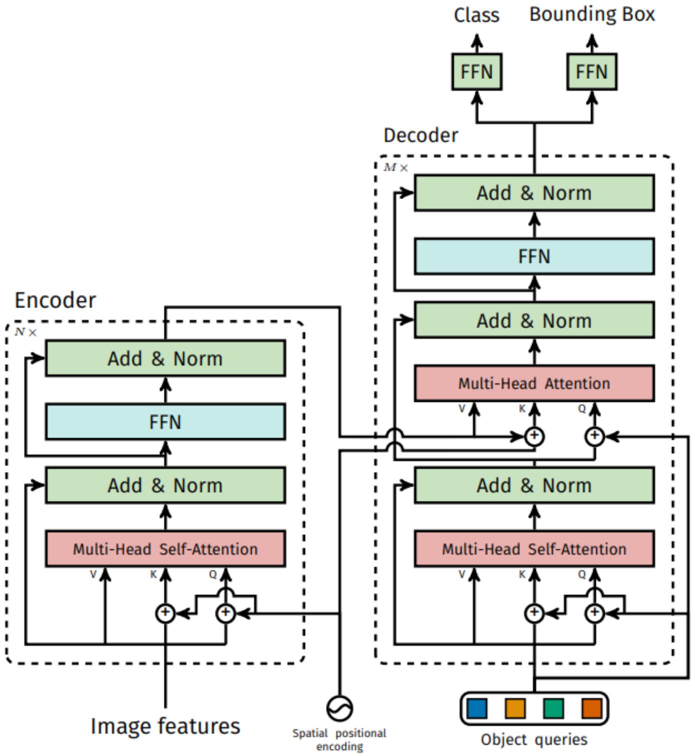
$$O(N_qN_kC)$$



$$O(H^2W^2C)$$ Complexity of self-attention in encoder

Motivation

- Drawback DETR
  - slow convergence
  - high complexity



$$O(N_qC^2 + N_kC^2 + N_qN_kC)$$ Complexity of Multihead attention

$$O(N_qN_kC)$$

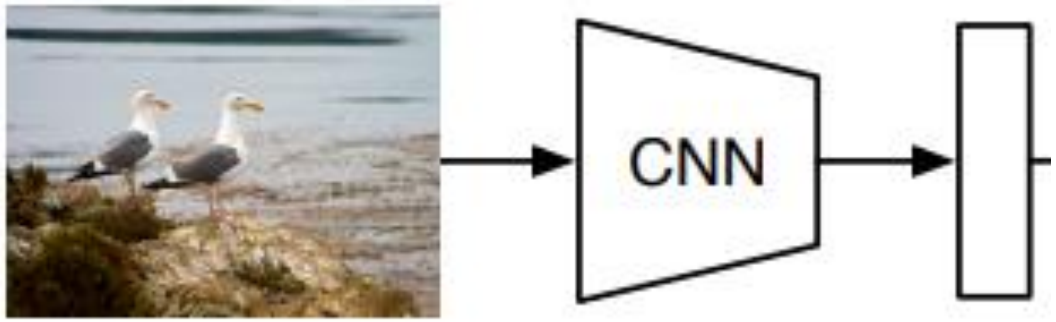
$$O(H^2W^2C)$$ Complexity of self-attention in encoder

$$O(HWC^2 + NHWC)$$ Complexity of cross-attention in decoder

$$O(2NC^2 + N^2C)$$ Complexity of self-attention in decoder

Motivation

- Drawback DETR
  - slow convergence
  - high complexity
  - low performance in detecting small objects



$x_{\text{img}} \in \mathbb{R}^{3 \times H_0 \times W_0}$

$\mathbb{R}^{C \times H \times W}$

$$C = 2048 \text{ and } H, W = \frac{H_0}{32}, \frac{W_0}{32}.$$

Motivation

- Drawback DETR
  - slow convergence
  - high complexity
  - low performance in detecting small objects
- Effiecient Attention Mechanism
  - pre-defined sparse attention

Motivation

- Drawback DETR
  - slow convergence
  - high complexity
  - low performance in detecting small objects
- Effiecient Attention Mechanism
  - pre-defined sparse attention
  - learn data-dependent sparse attention

Motivation

- Drawback DETR
  - slow convergence
  - high complexity
  - low performance in detecting small objects
- Effiecient Attention Mechanism
  - pre-defined sparse attention
  - learn data-dependent sparse attention
  - explore the low-rank property in self-attention

Motivation

- Drawback DETR
  - slow convergence
  - high complexity
  - low performance in detecting small objects
- Effiecient Attention Mechanism
  - pre-defined sparse attention
  - learn data-dependent sparse attention
  - explore the low-rank property in self-attention

Motivation

- Drawback DETR
  - slow convergence
  - high complexity
  - low performance in detecting small objects
- Effiecient Attention Mechanism
  - use pre-defined sparse attention
  - learn data-dependent sparse attention
  - explore the low-rank property in self-attention
- Multi-scale Feature Representation for Object Detection

Deformable Transformer

$$A_{mqk} \propto \exp\{\frac{\boldsymbol{z}_q^T \boldsymbol{U}_m^T \boldsymbol{V}_m \boldsymbol{x}_k}{\sqrt{C_v}}\}$$

- Deformable attetion module

$$\text{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \right]$$

m	index for attention head	$\boldsymbol{W}_m$	output projection matrix at m^{th} head
q	index for query element	$\boldsymbol{U}_m$	input query projection matrix at m^{th} head
k	index for key element	$\boldsymbol{V}_m$	input key projection matrix at m^{th} head
$\boldsymbol{z}_q$	input feature of q^{th} query	$\boldsymbol{W}'_m$	input value projection matrix at m^{th} head

A_{mqk} attention weight of q^{th} query to k^{th} key at m^{th} head

$\boldsymbol{x}$ input feature map (input feature of key elements)

$\boldsymbol{x}_k$ input feature of k^{th} key

Deformable Transformer

- Deformable attetion module

$K \ll HW$

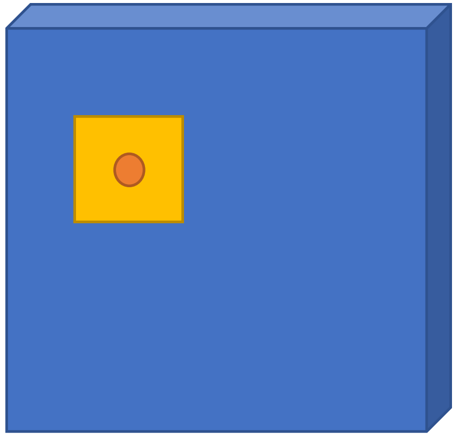
$$\text{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \Big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \Big]$$

m	index for attention head	$\boldsymbol{W}_m$	output projection matrix at m^{th} head
q	index for query element	$\boldsymbol{U}_m$	input query projection matrix at m^{th} head
k	index for key element	$\boldsymbol{V}_m$	input key projection matrix at m^{th} head
$\boldsymbol{z}_q$	input feature of q^{th} query	$\boldsymbol{W}'_m$	input value projection matrix at m^{th} head
A_{mqk}	attention weight of q^{th} query to k^{th} key at m^{th} head		
$\boldsymbol{x}$	input feature map (input feature of key elements)		
$\boldsymbol{x}_k$	input feature of k^{th} key		

Deformable Transformer

- Deformable attetion module

$$\text{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \big]$$



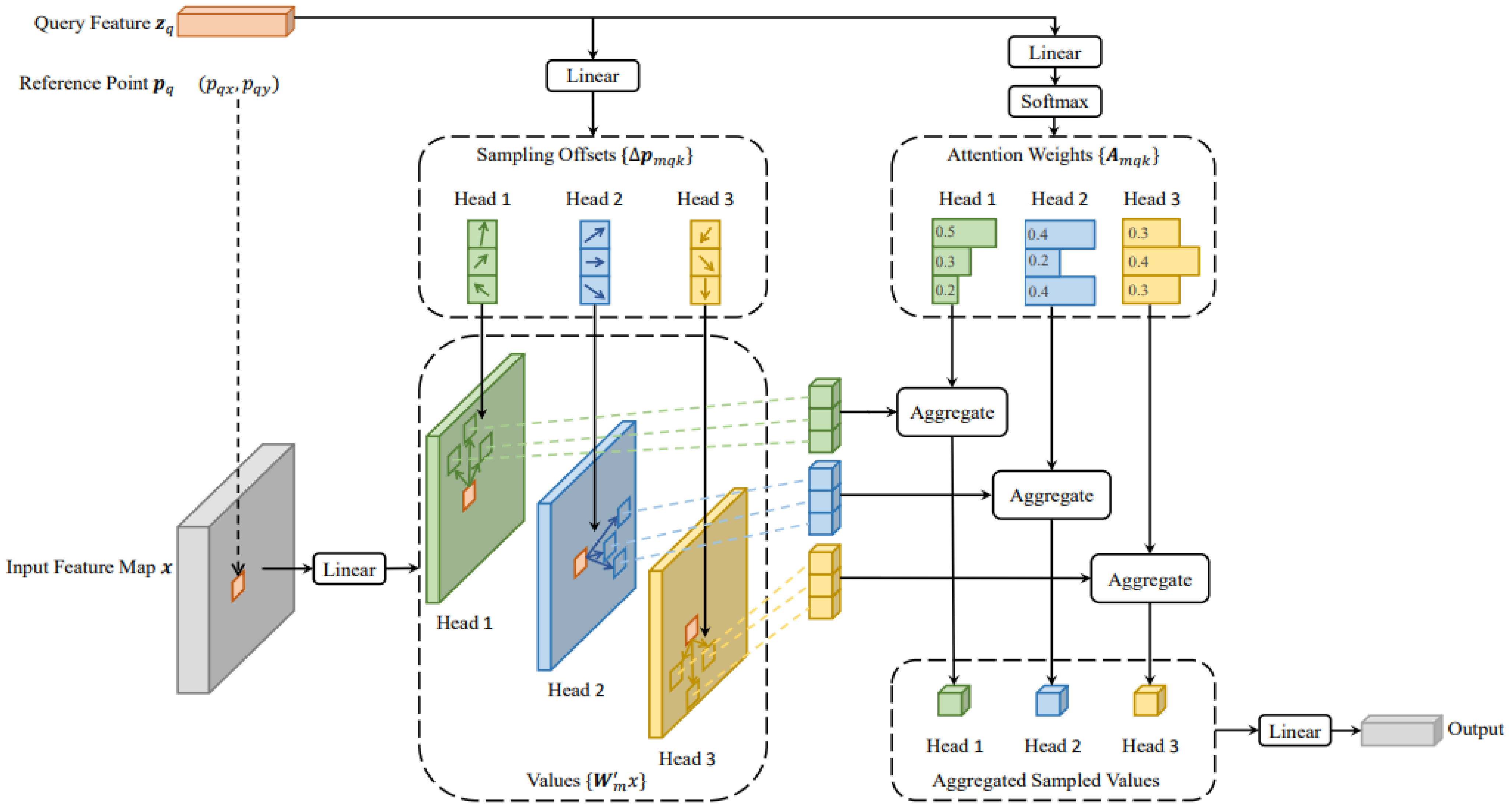
Softmax from Zq (1MK channels)

$$\text{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \big]$$

Linear projection from Zq (2MK channels)

$$\mathrm{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \big]$$

Deforr



Deformable Transformer

- Deformable attetion module

$$K \ll HW$$

$$\text{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \left[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \right]$$

Calculate offsets and attention weights $O(3N_qCMK)$

Calculate the equation $O(N_qC^2 + N_qKC^2 + 5N_qKC)$

Calculate whole $O(N_qC^2 + \min(HWC^2, N_qKC^2) + 5N_qKC + 3N_qCMK)$

$$O(2N_qC^2 + \min(HWC^2, N_qKC^2))$$

Deformable Transformer

- Deformable attetion module

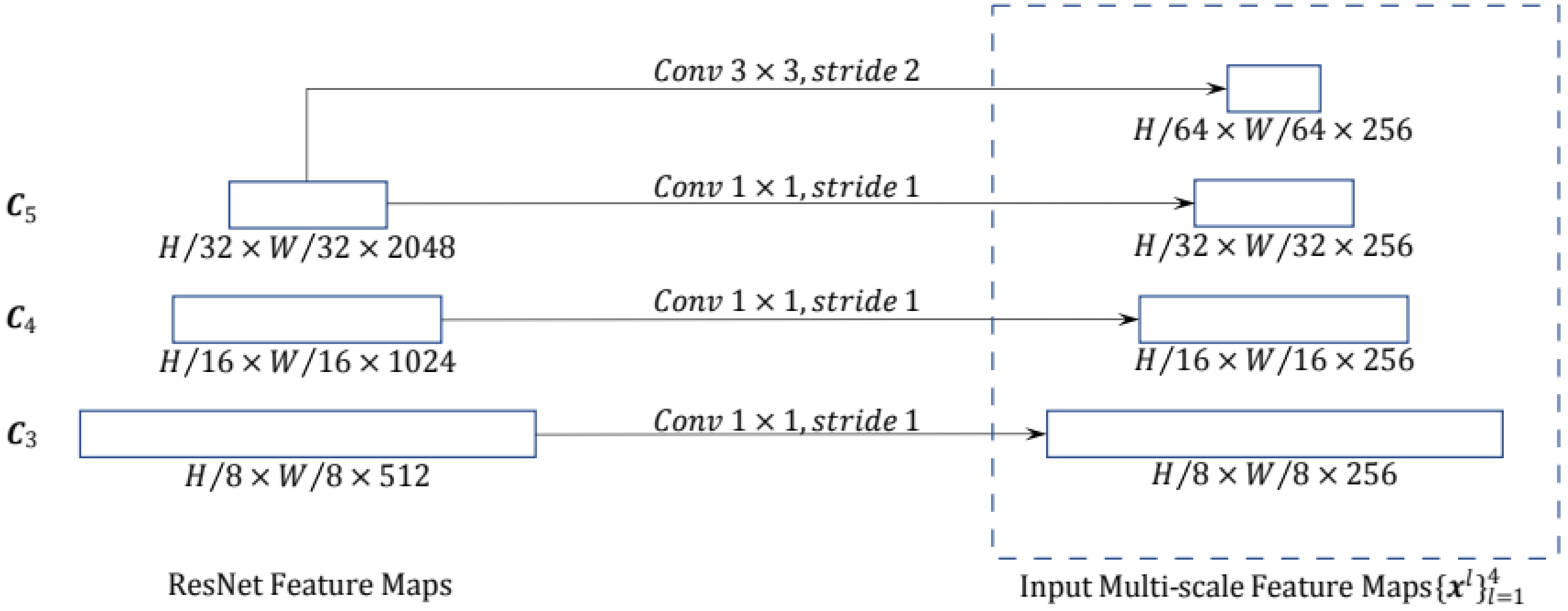
$$K \ll HW$$

$$\text{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \Big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \Big]$$

$$O(2N_qC^2 + \min(HWC^2, N_qKC^2)) \left\{ \begin{array}{lll} \text{encoder} & N_q = HW & O(HWC^2) \\ \text{cross-attention} & N_q = N & O(NKC^2) \end{array} \right.$$

Deformable Transformer

- Deformable attetion module
- Multi-scale Deformable Attention Module



Deformable Transformer

- Deformable attetion module
- Multi-scale Deformable Attention Module

$$\text{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \big]$$

$$\text{MSDeformAttn}(\boldsymbol{z}_q, \hat{\boldsymbol{p}}_q, \{\boldsymbol{x}^l\}_{l=1}^{L}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{l=1}^{L} \sum_{k=1}^{K} A_{mlqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}^l (\phi_l(\hat{\boldsymbol{p}}_q) + \Delta \boldsymbol{p}_{mlqk}) \big]$$

Deformable Transformer

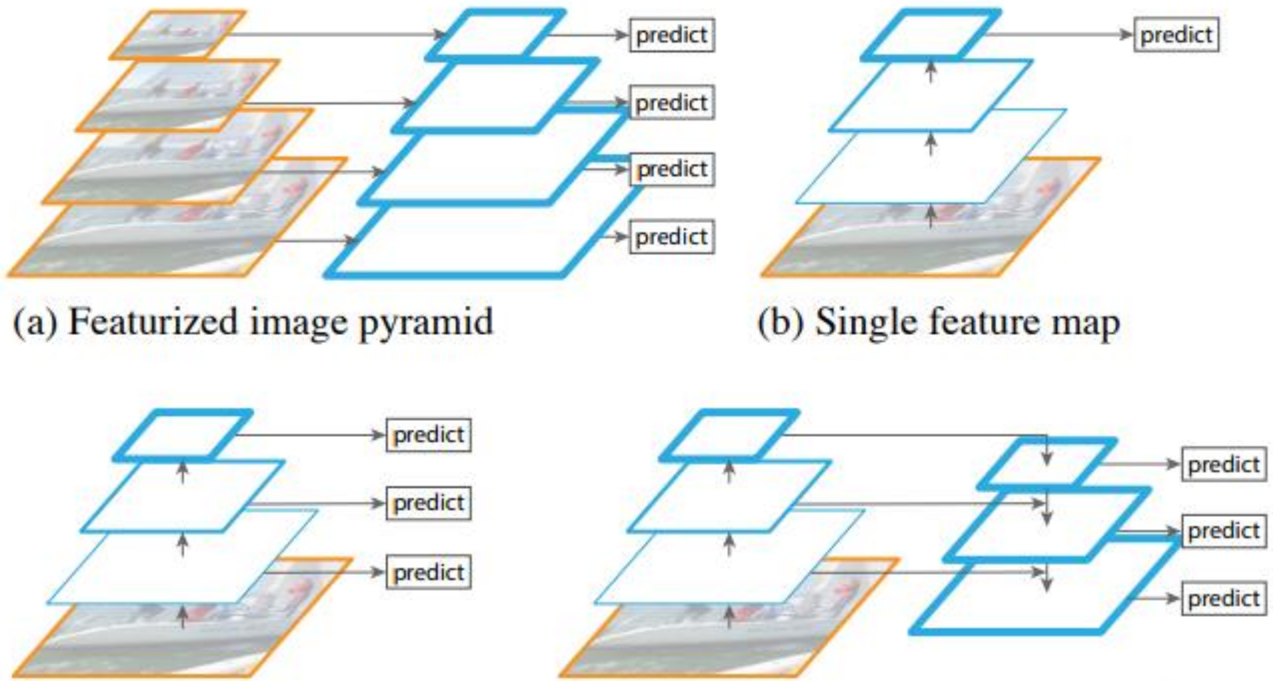
- Deformable attetion module
- Multi-scale Deformable Attention Module

$$\text{MSDeformAttn}(\boldsymbol{z}_q, \hat{\boldsymbol{p}}_q, \{\boldsymbol{x}^l\}_{l=1}^{L}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{l=1}^{L} \sum_{k=1}^{K} A_{mlqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}^l (\phi_l(\hat{\boldsymbol{p}}_q) + \Delta \boldsymbol{p}_{mlqk}) \big]$$

$\hat{\boldsymbol{p}}_q \in [0, 1]^2$ normalized coordinates of the reference point for each query element q

Deformable Transformer

- Deformable attetion module
- Multi-scale Deformable Attention Module

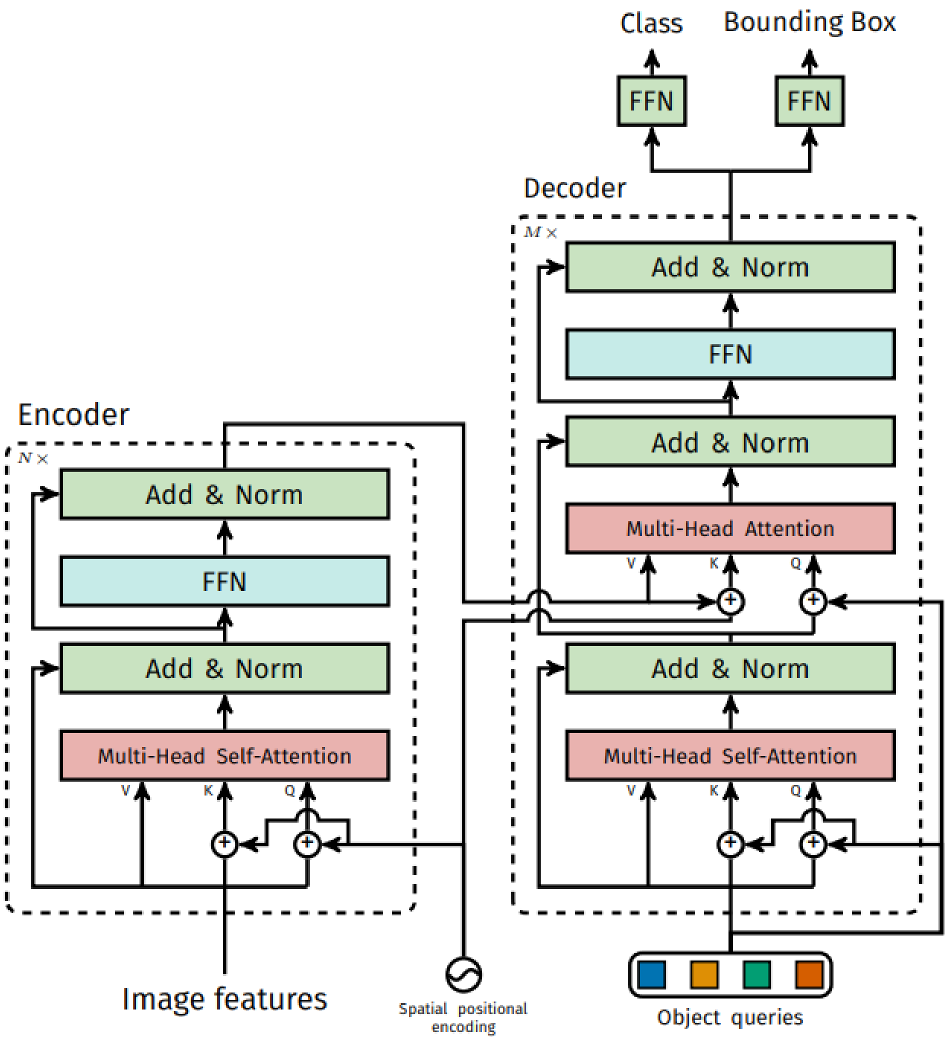


(a) Featurized image pyramid (b) Single feature map

(c) Pyramidal feature hierarchy (d) Feature Pyramid Network

Deformable Transformer

- Deformable attetion module
- Multi-scale Deformable Attention Module
- Deformable Transformer Decoder
  - Cross-Attention and Self-attention



Deformable Transformer

- Deformable attetion module
- Multi-scale Deformable Attention Module
- Deformable Transformer Decoder
  - Bounding box -> relative offsets w.r.t the reference point

$\hat{\boldsymbol{p}}_q$	2-d normalized coordinate	Predicted from its object query embedding via a learnable linear projection followed by a sigmoid function

Deformable Transformer

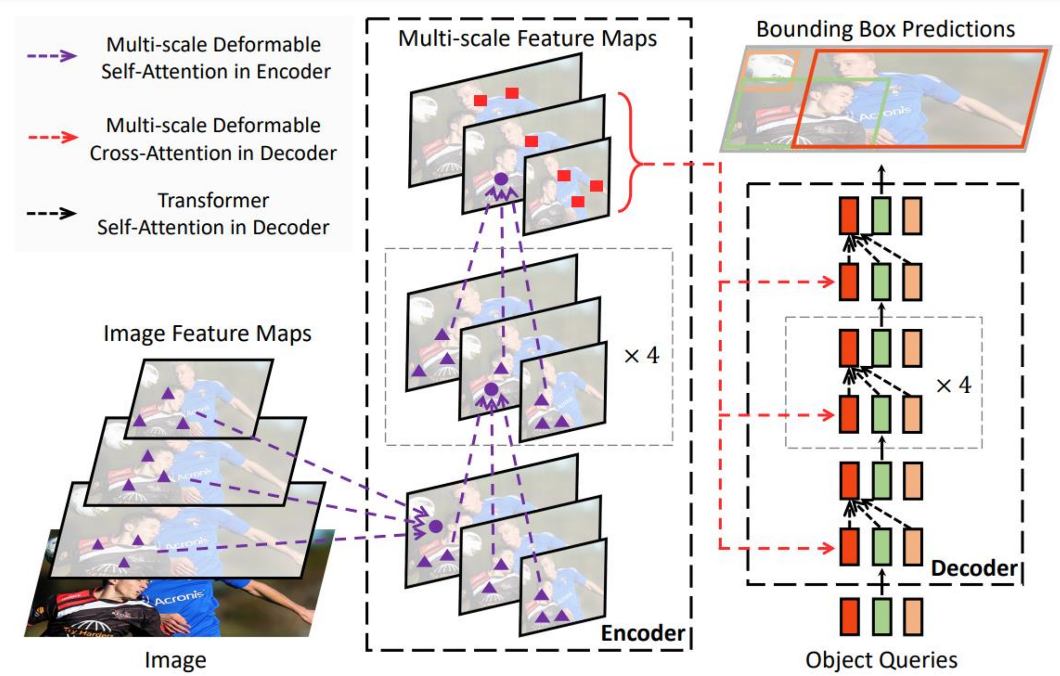
- Deformable attetion module
- Multi-scale Deformable Attention Module
- Deformable Transformer Decoder
  - Bounding box -> relative offsets w.r.t the reference point

$$\hat{\boldsymbol{p}}_q \qquad (\hat{p}_{qx}, \hat{p}_{qy})$$

$$\hat{\boldsymbol{b}}_q = \{\sigma\big(b_{qx} + \sigma^{-1}(\hat{p}_{qx})\big), \sigma\big(b_{qy} + \sigma^{-1}(\hat{p}_{qy})\big), \sigma(b_{qw}), \sigma(b_{qh})\}$$

Deformable Transformer

- Deformable attetion module
- Multi-scale Deformable Attention Module
- Deformable Transformer Decoder



Addtional Improvements

- Iterative Bounding Box Refinement.

$$\hat{\boldsymbol{b}}_q^d = \{\sigma(\Delta b_{qx}^d + \sigma^{-1}(\hat{b}_{qx}^{d-1})), \sigma(\Delta b_{qy}^d + \sigma^{-1}(\hat{b}_{qy}^{d-1})), \sigma(\Delta b_{qw}^d + \sigma^{-1}(\hat{b}_{qw}^{d-1})), \sigma(\Delta b_{qh}^d + \sigma^{-1}(\hat{b}_{qh}^{d-1}))\}$$

where $d \in \{1, 2, ..., D\}$, $\Delta b_{q\{x,y,w,h\}}^d \in \mathbb{R}$

$\hat{b}_{qx}^0 = \hat{p}_{qx}$, $\hat{b}_{qy}^0 = \hat{p}_{qy}$, $\hat{b}_{qw}^0 = 0.1$, and $\hat{b}_{qh}^0 = 0.1$

d-th decoder layer, $(\hat{b}_{qx}^{d-1}, \hat{b}_{qy}^{d-1})$ serves as the new reference point.

$\Delta \boldsymbol{p}_{mlqk}$ is also modulated by the box size, as $(\Delta p_{mlqkx}\, \hat{b}_{qw}^{d-1}, \Delta p_{mlqky}\, \hat{b}_{qh}^{d-1})$

Addtional Improvements

- Iterative Bounding Box Refinement.
- Two-Stage Deformable DETR (Region proposal/Encoder/Top scoring)

$$\hat{\boldsymbol{b}}_i = \{\sigma(\Delta b_{ix} + \sigma^{-1}(\hat{p}_{ix})), \sigma(\Delta b_{iy} + \sigma^{-1}(\hat{p}_{iy})), \sigma(\Delta b_{iw} + \sigma^{-1}(2^{l_i - 1}s)), \sigma(\Delta b_{ih} + \sigma^{-1}(2^{l_i - 1}s))\}$$

Addtional Improvements

- Iterative Bounding Box Refinement.
- Two-Stage Deformable DETR (Region proposal/Encoder/Top scoring)

Bias parameters of the linear projection are initialized to make $A_{mlqk} = \frac{1}{LK}$ and $\{\Delta \boldsymbol{p}_{1lqk} = (-k, -k), \Delta \boldsymbol{p}_{2lqk} = (-k, 0), \Delta \boldsymbol{p}_{3lqk} = (-k, k), \Delta \boldsymbol{p}_{4lqk} = (0, -k), \Delta \boldsymbol{p}_{5lqk} = (0, k), \Delta \boldsymbol{p}_{6lqk} = (k, -k), \Delta \boldsymbol{p}_{7lqk} = (k, 0), \Delta \boldsymbol{p}_{8lqk} = (k, k)\}$ $(k \in \{1, 2, ...K\})$ at initialization.

Experiments

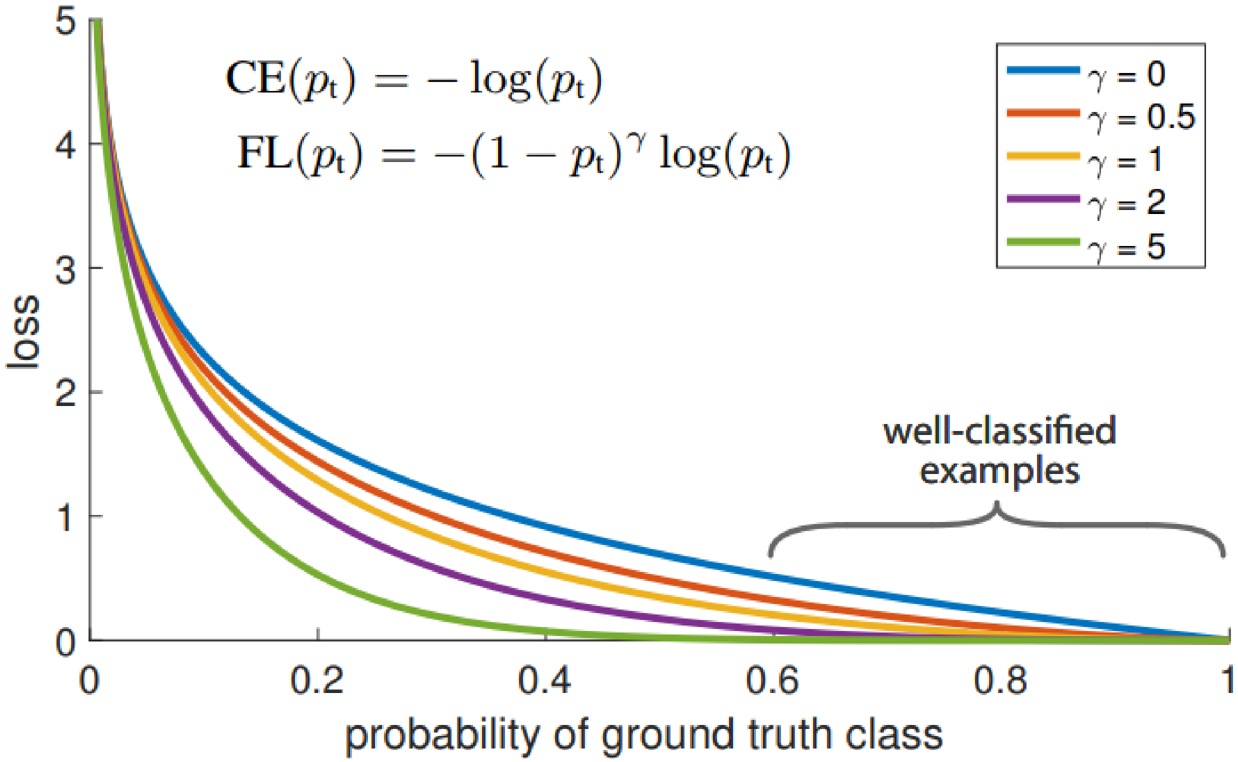
- Parameters of the deformable Transformer encoder are shared among different feature levels
- Focal loss

Experiments

- Parameters of the deformable Transformer encoder are shared among different feature levels
- Focal loss

$$\mathrm{CE}(p, y) = \begin{cases} -\log(p) & \text{if } y = 1 \\ -\log(1-p) & \text{otherwise.} \end{cases}$$

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma} \log(p_t)$$

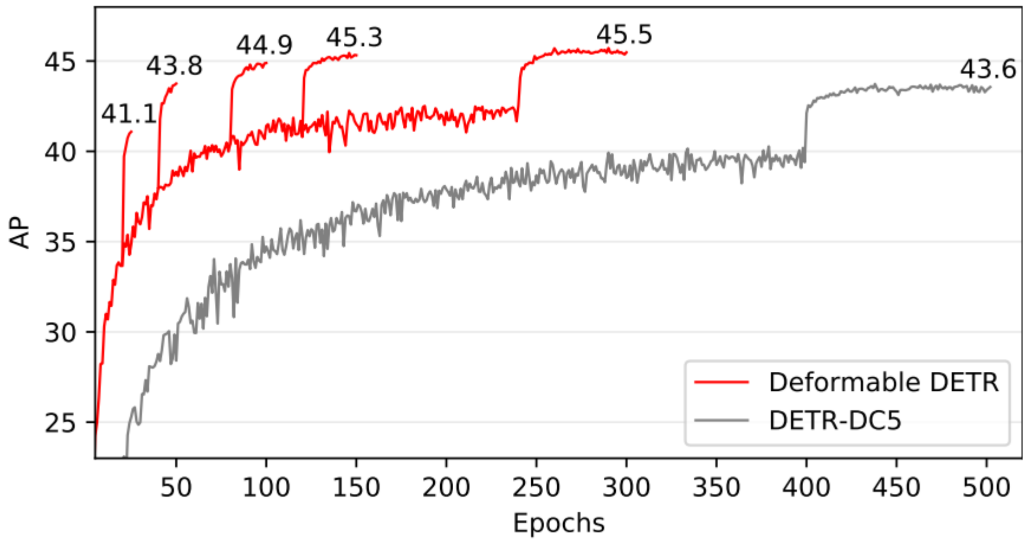


Experiments

- Parameters of the deformable Transformer encoder are shared among different feature levels
- Focal loss
- Adam optimizer

Results

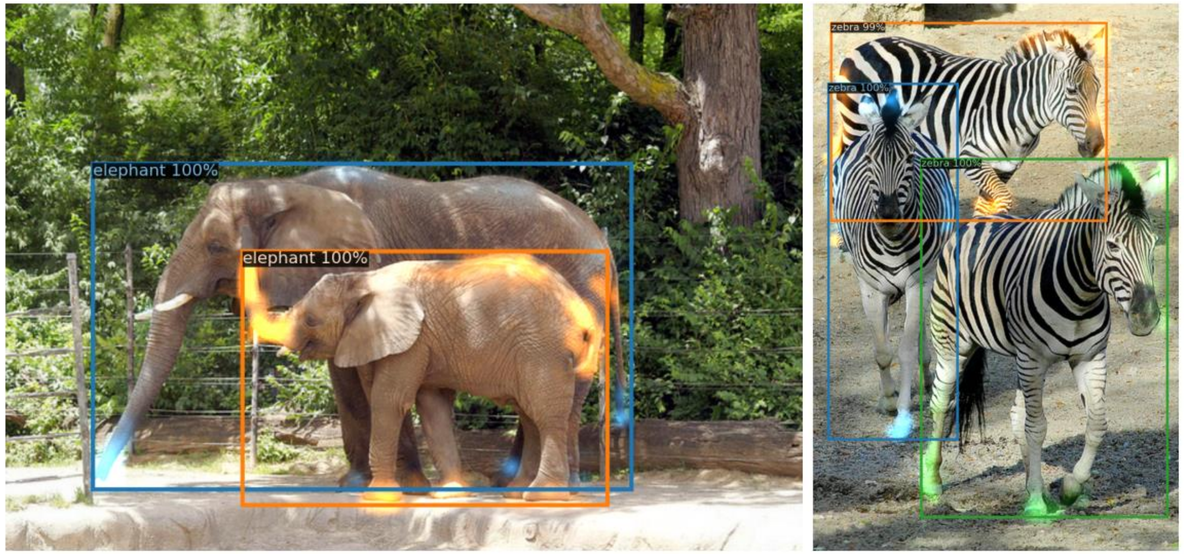
Method	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	params	FLOPs	Training GPU hours	Inference FPS
Faster R-CNN + FPN	109	42.0	62.1	45.5	26.6	45.4	53.4	42M	180G	380	26
DETR	500	42.0	62.4	44.2	20.5	45.8	61.1	41M	86G	2000	28
DETR-DC5	500	43.3	63.1	45.9	22.5	47.3	61.1	41M	187G	7000	12
DETR-DC5	50	35.3	55.7	36.8	15.2	37.5	53.6	41M	187G	700	12
DETR-DC5$^+$	50	36.2	57.0	37.4	16.3	39.2	53.9	41M	187G	700	12
Deformable DETR	50	43.8	62.6	47.7	26.4	47.1	58.0	40M	173G	325	19
+ iterative bounding box refinement	50	45.4	64.7	49.0	26.8	48.3	61.7	40M	173G	325	19
++ two-stage Deformable DETR	50	46.2	65.2	50.0	28.8	49.2	61.7	40M	173G	340	19

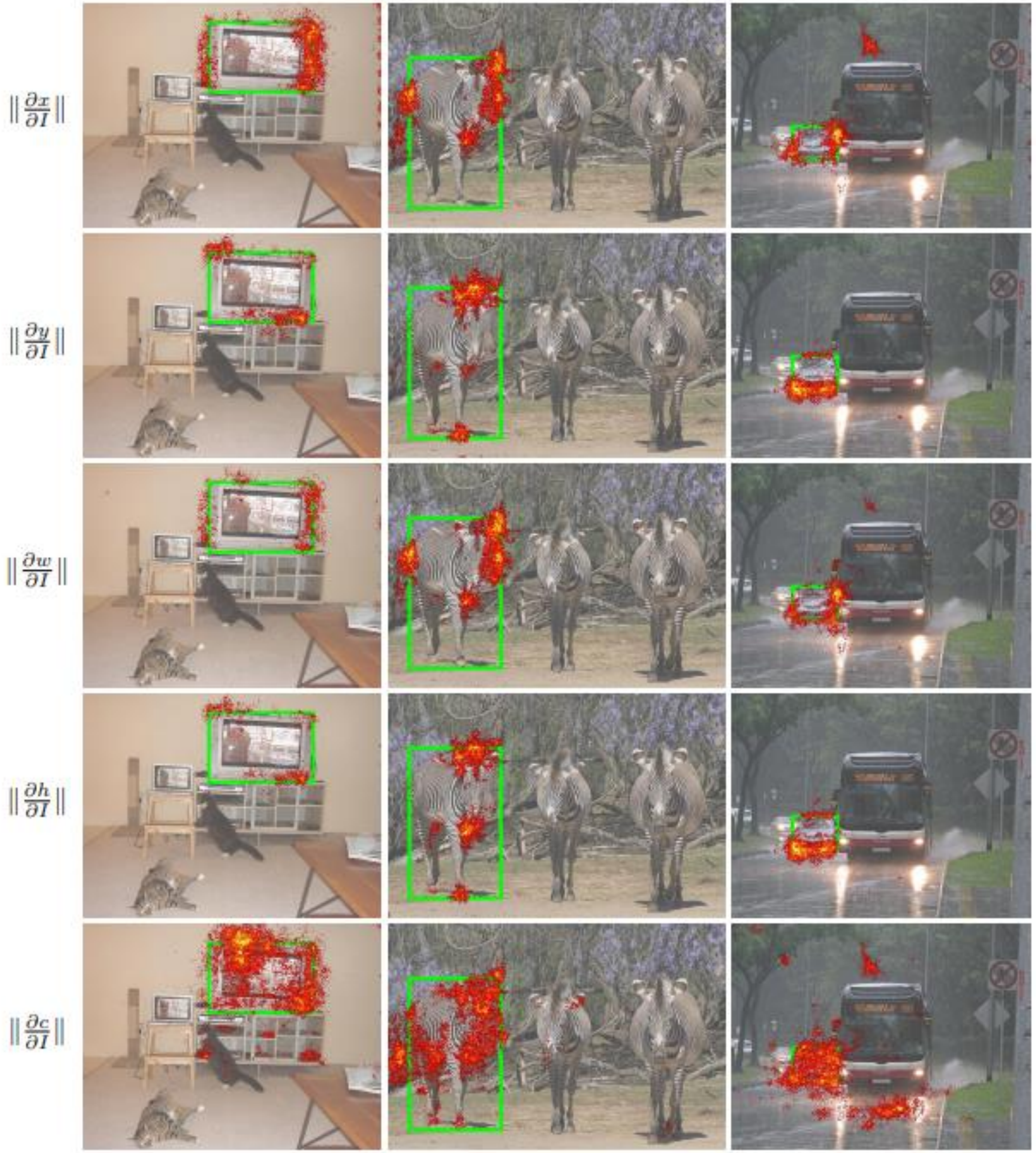


Results

MS inputs	MS attention	K	FPNs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
✓	✓	4	FPN (Lin et al., 2017a)	43.8	62.6	47.8	26.5	47.3	58.1
✓	✓	4	BiFPN (Tan et al., 2020)	43.9	62.5	47.7	25.6	47.4	57.7
		1	w/o	39.7	60.1	42.4	21.2	44.3	56.0
✓		1	w/o	41.4	60.9	44.9	24.1	44.6	56.1
✓		4	w/o	42.3	61.4	46.0	24.8	45.1	56.3
✓	✓	4	w/o	43.8	62.6	47.7	26.4	47.1	58.0

Results


elephant 100%
elephant 100%
zebra 99%
zebra 100%
zebra 100%


$\|\frac{\partial x}{\partial I}\|$
$\|\frac{\partial y}{\partial I}\|$
$\|\frac{\partial w}{\partial I}\|$
$\|\frac{\partial h}{\partial I}\|$
$\|\frac{\partial c}{\partial I}\|$