JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Siamese DETR

Zeren Chen[1,2], Gengshi Huang[2], Wei Li[3], Jianing Teng[2], KunWang[2],
Jing Shao[2], Chen Change Loy[3], Lu Sheng[1]
[1]Beihang University, [2]SenseTime Research, [3]Nanyang Technological University

Tag: WED-PM-321

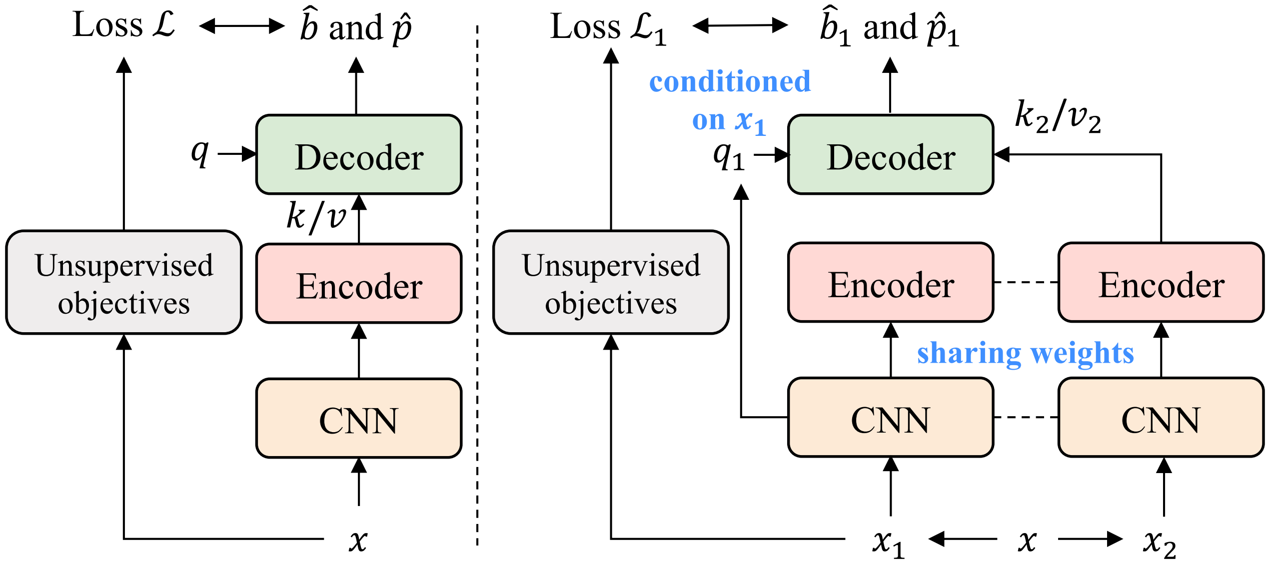
Code is available at https://github.com/Zx55/SiameseDETR

Overview

- Combine Siamese networks with cross-attention mechanism in DETR.
- Two newly-designed self-supervised pretext tasks.
  - **Multi-View Region Detection**
  - **Multi-View Semantic Discrimination**
- Siamese DETR outperforms its counterpart with multiple DETR variants on the COCO and PASCAL VOC benchmark.

Introduction & Motivation

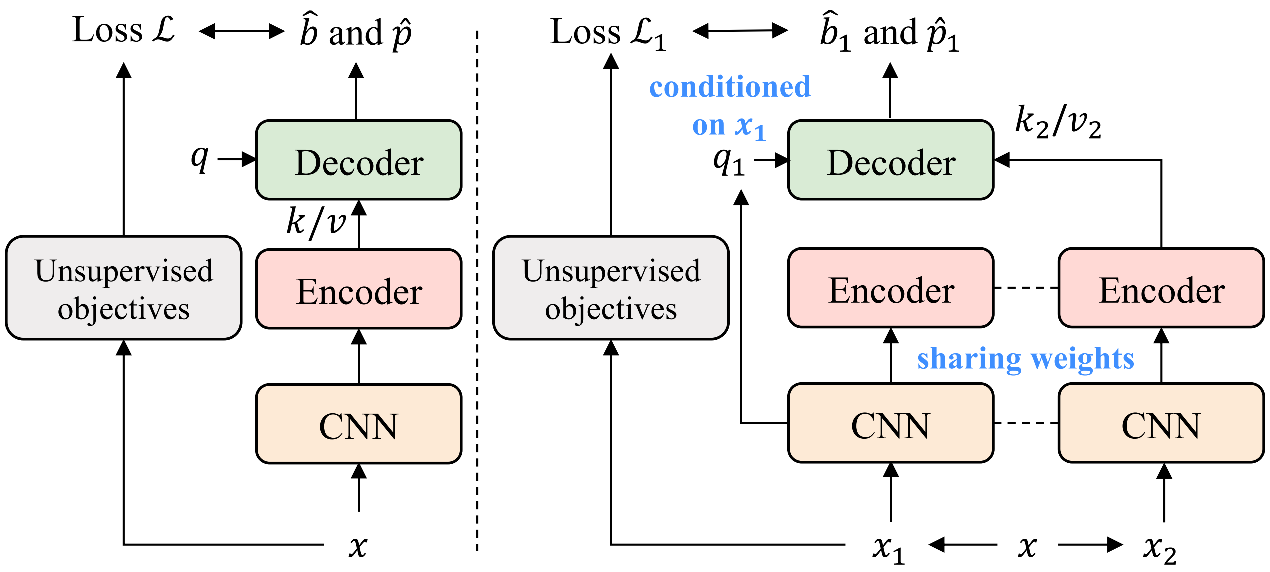
- Objective: Design a self-supervised learning approach for DETR pretraining to alleviate the massive appetite for labeled data in training DETR.
- Existing self-supervised learning approaches cannot be extended to DETR effectively (e.g., SimCLR, MoCo).



(a) (b)

Introduction & Motivation

- Several recent attempts (e.g., UP-DETR, DETReg) follows a single-view paradigm (see a), ignoring the ability of learning view-invariant representation.
- Siamese DETR learn view-invariant and detection-oriented representation through two pretext tasks (see b).

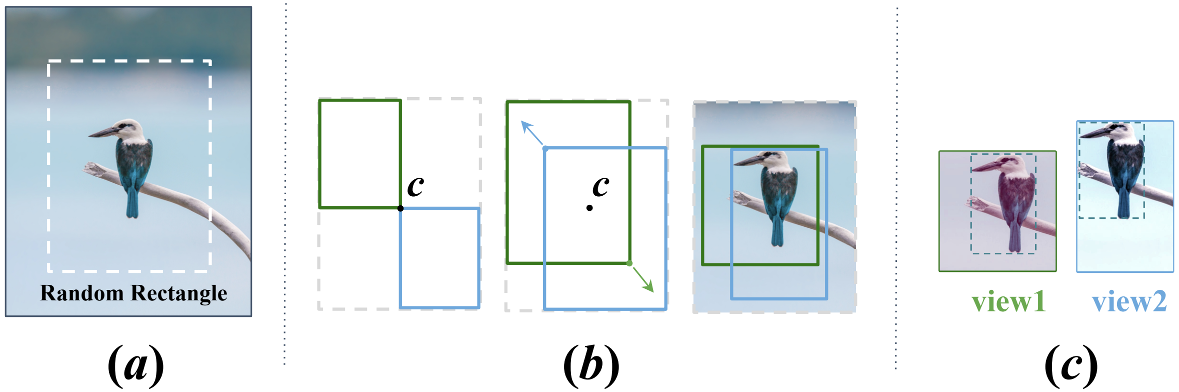


(a)

(b)

View Construction

- Generate two views based on an IoU-constrained policy.
  - Generate two rectangles and keep their IoU larger than a threshold.
  - Crop two rectangles and apply augmentations (following SimSiam).
  - Generate offline object proposals with Edgeboxes in the overlapping area.



Cross-View Learning

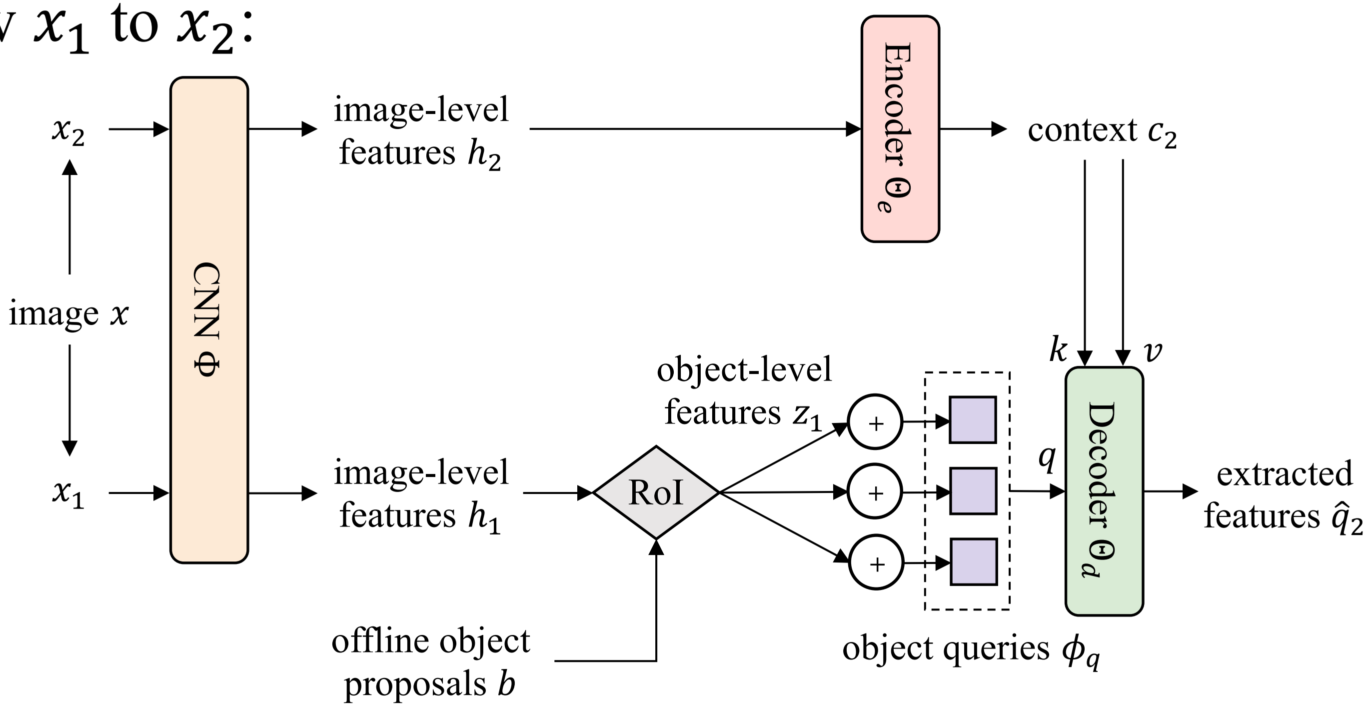
- We propose Multi-View Cross-Attention for multi-view representation learning.
- E.g., cross-attention from view x_1 to x_2:

$$\boldsymbol{q}_1 = f_q(\boldsymbol{z}_1 + \boldsymbol{\phi}_q)$$

$$\boldsymbol{k}_2 = f_k(\boldsymbol{c}_2)$$

$$\boldsymbol{v}_2 = f_v(\boldsymbol{c}_2)$$

$$\widehat{\boldsymbol{q}}_2 = \mathtt{MVCA}(\boldsymbol{c}_2, \boldsymbol{\phi}_q, \boldsymbol{z}_1) = \mathtt{MHA}(\boldsymbol{q}_1, \boldsymbol{k}_2, \boldsymbol{v}_2)$$



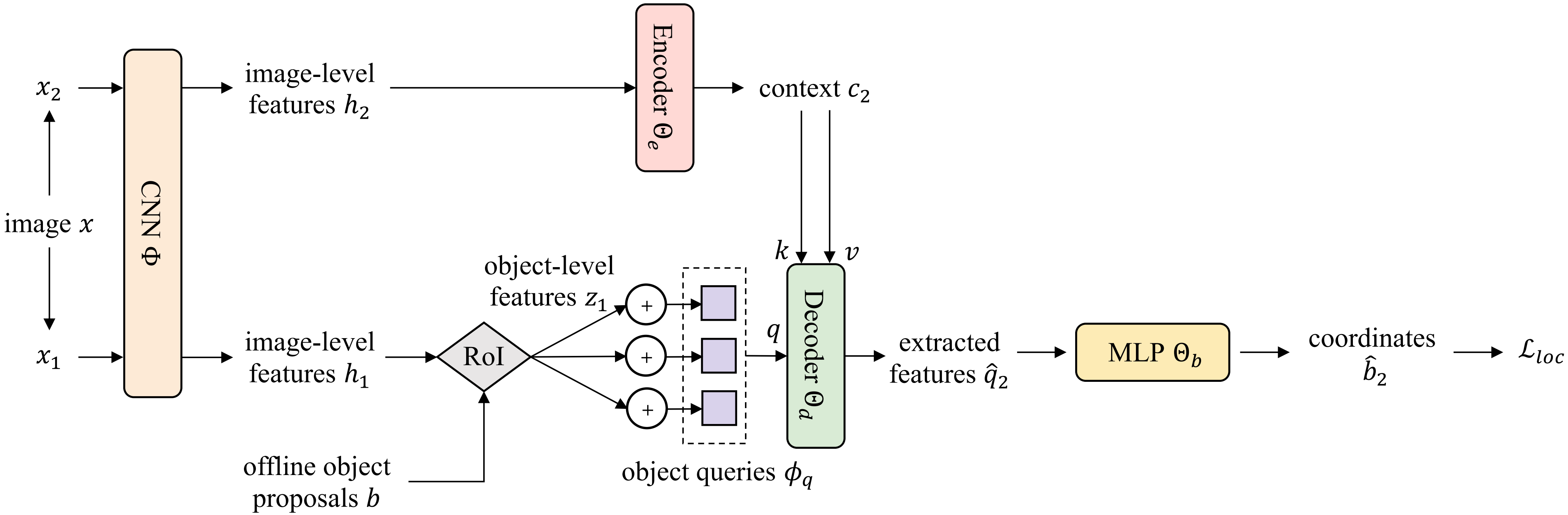
Here, $\hat{q}_2$ are supposed to be semantically consistent with the corresponding region features z_1 on view x_1.

Learning to Locate

- Locate the region in view x_2 that is relative to the region feature z_1.

$$\widehat{\boldsymbol{b}}_2 = f_{box}(\widehat{\boldsymbol{q}}_2) \in \mathbb{R}^{N\times 4}$$

$$\mathcal{L}_{loc} = \ell_{box}(\widehat{\boldsymbol{b}}_2, \boldsymbol{b}_2) + \ell_{box}(\widehat{\boldsymbol{b}}_1, \boldsymbol{b}_1)$$



Learning to Discriminate

- Global discrimination

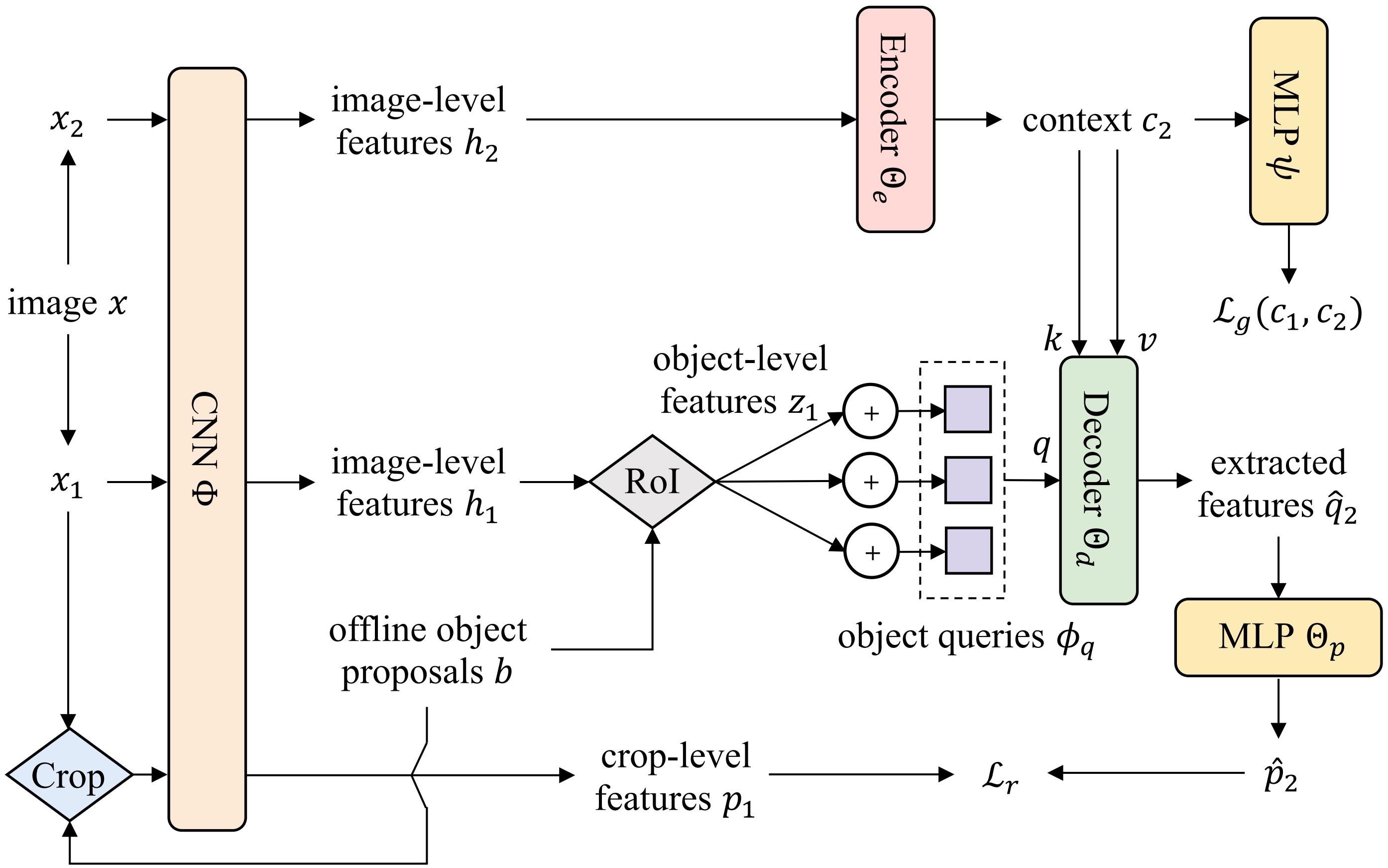
$$\mathcal{L}_g = \mathcal{C}\big[\texttt{MLP}(\boldsymbol{c}_1),\texttt{detach}(\boldsymbol{c}_2)\big] + \mathcal{C}\big[\texttt{MLP}(\boldsymbol{c}_2),\texttt{detach}(\boldsymbol{c}_1)\big]$$

- Regional discrimination

$$\boldsymbol{p}_1 = \texttt{Backbone}(\texttt{Crop}(\boldsymbol{x}_1, \boldsymbol{b}_1))$$

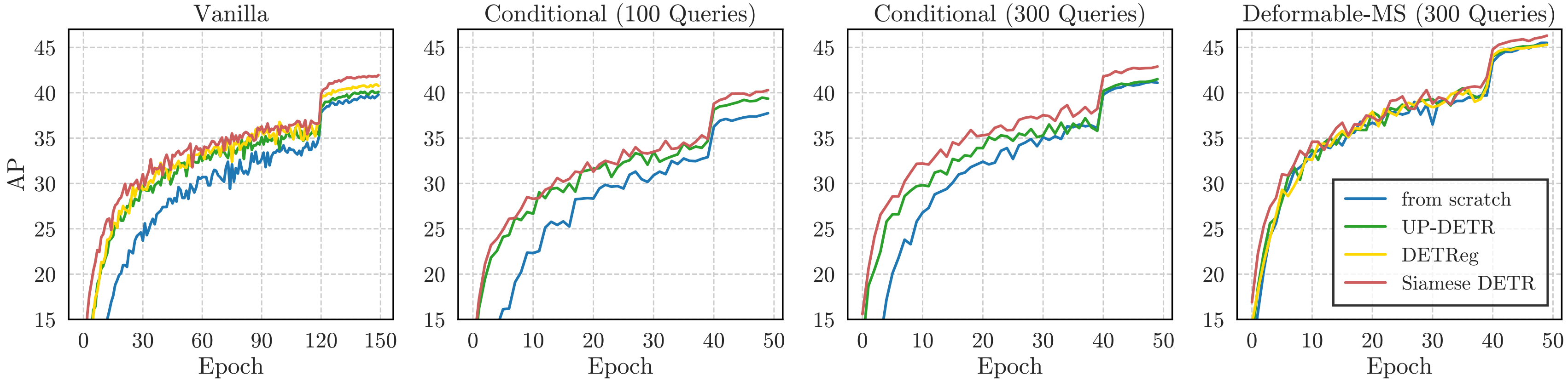
$$\widehat{\boldsymbol{p}}_2 = f_{sem}(\widehat{\boldsymbol{q}}_2) \in \mathbb{R}^{N \times C'}$$

$$\mathcal{L}_r = \mathcal{D}(\widehat{\boldsymbol{p}}_2, \boldsymbol{p}_1) + \mathcal{D}(\widehat{\boldsymbol{p}}_1, \boldsymbol{p}_2)$$

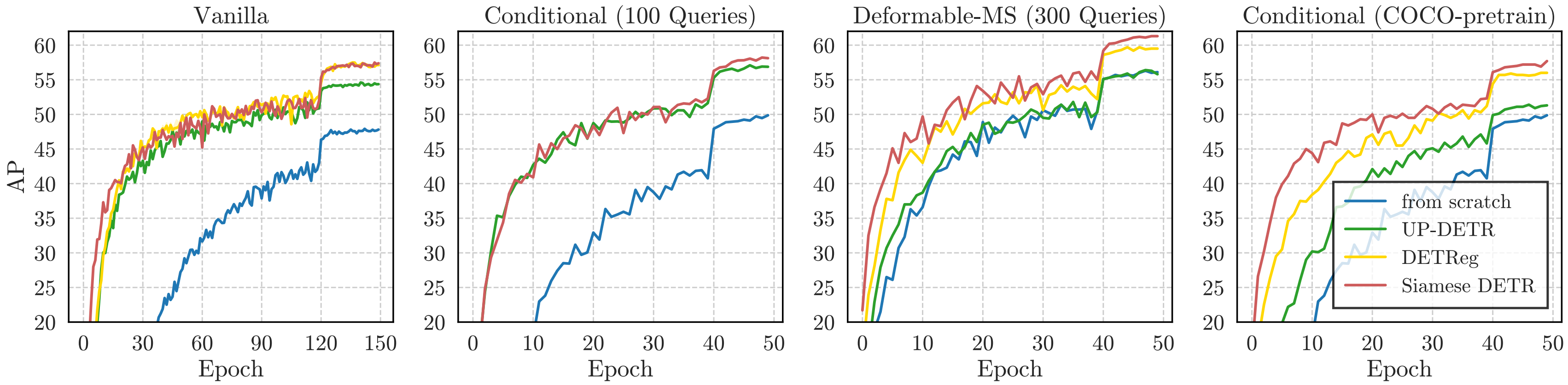


Experiments

- MS-COCO benchmark (ImageNet -> COCO)



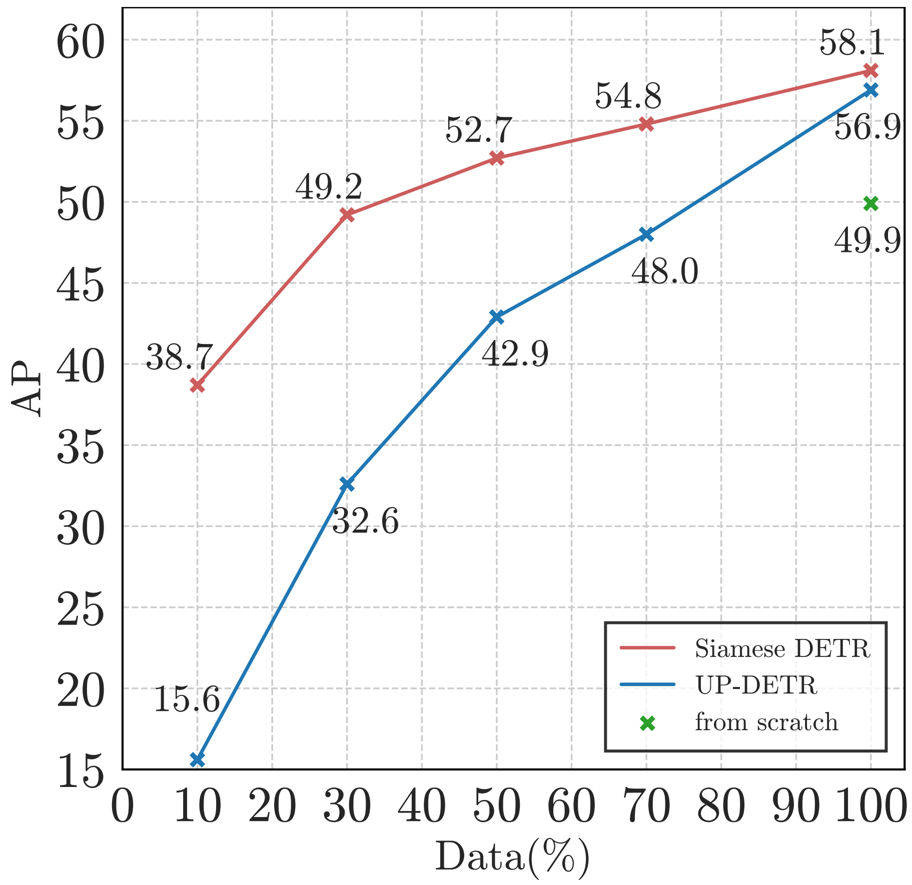
- PASCAL VOC benchmark (ImageNet -> VOC / COCO -> VOC)



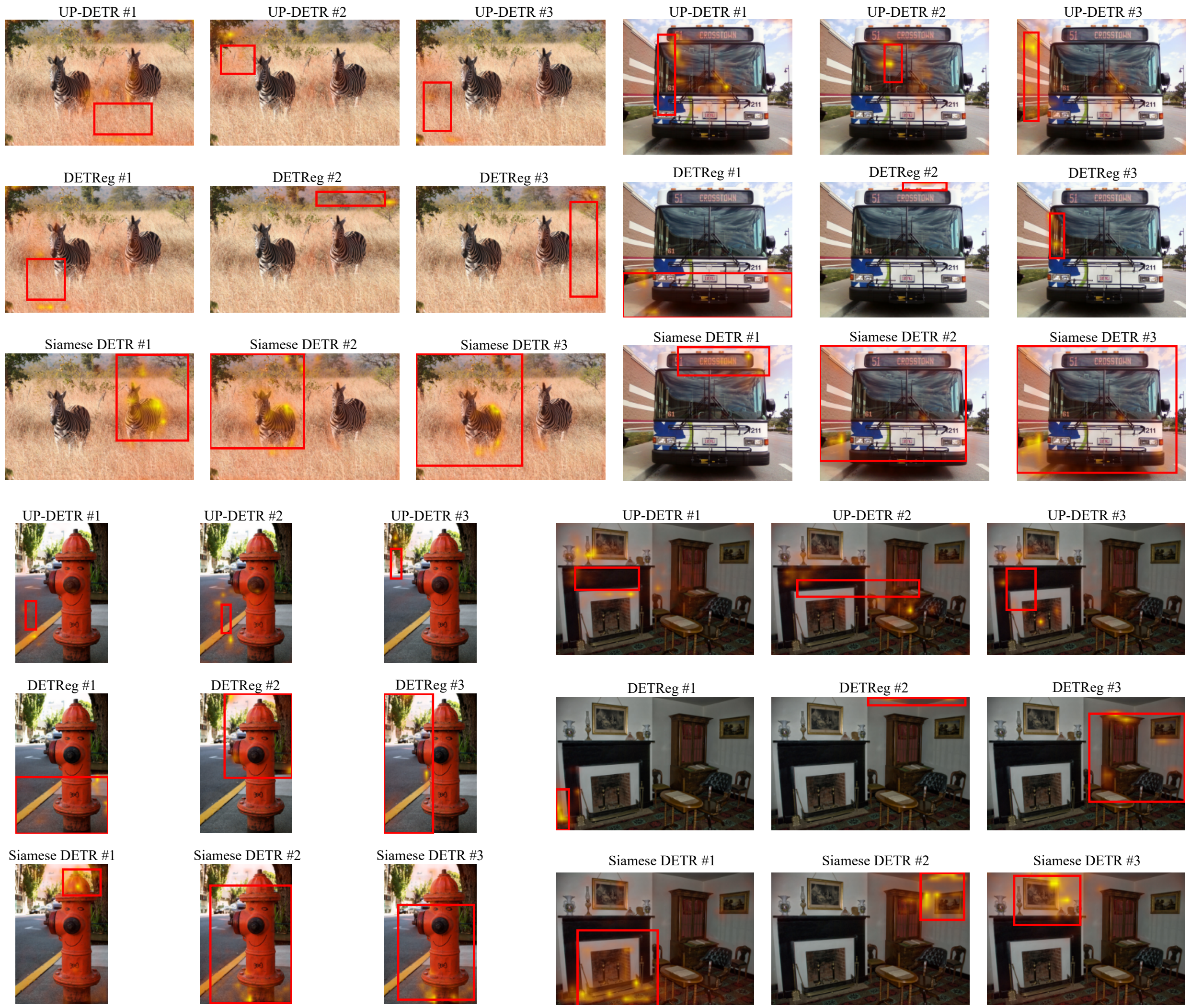
Experiments

Method	Dataset	Proposals	AP
UP-DETR	ImageNet→COCO	Random	39.4
DETReg	ImageNet→COCO	Random	40.3
ours	ImageNet→COCO	Random	**40.4**
UP-DETR	ImageNet→COCO	Edgeboxes	39.3
DETReg	ImageNet→COCO	Edgeboxes	40.3
ours	ImageNet→COCO	Edgeboxes	**40.5**
UP-DETR	COCO→VOC	Random	51.3
DETReg	COCO→VOC	Random	51.9
ours	COCO→VOC	Random	**54.9**
DETReg	COCO→VOC	SelectiveSearch	55.9
ours	COCO→VOC	SelectiveSearch	**56.2**
UP-DETR	COCO→VOC	Edgeboxes	57.0
DETReg	COCO→VOC	Edgeboxes	56.3
ours	COCO→VOC	Edgeboxes	**57.7**

DETR	VOC		VOC 10%	
	DAB-DETR	DN-DETR	DAB-DETR	DN-DETR
from scratch	57.9	58.9	32.2	32.9
Siamese DETR	**62.2 (+4.3)**	**63.4 (+4.5)**	**41.8 (+9.6)**	**43.6 (+10.7)**



Visualization



Thanks for your watching!