



DETRs with Hybrid Matching

Ding Jia
PKU

Yuhui Yuan
MSRA

Haodi He
Stanford

Xiaopei Wu
ZJU

Haojun Yu
PKU

Weihong Lin
MSRA

Lei Sun
MSRA

Chao Zhang
PKU

Han Hu
MSRA

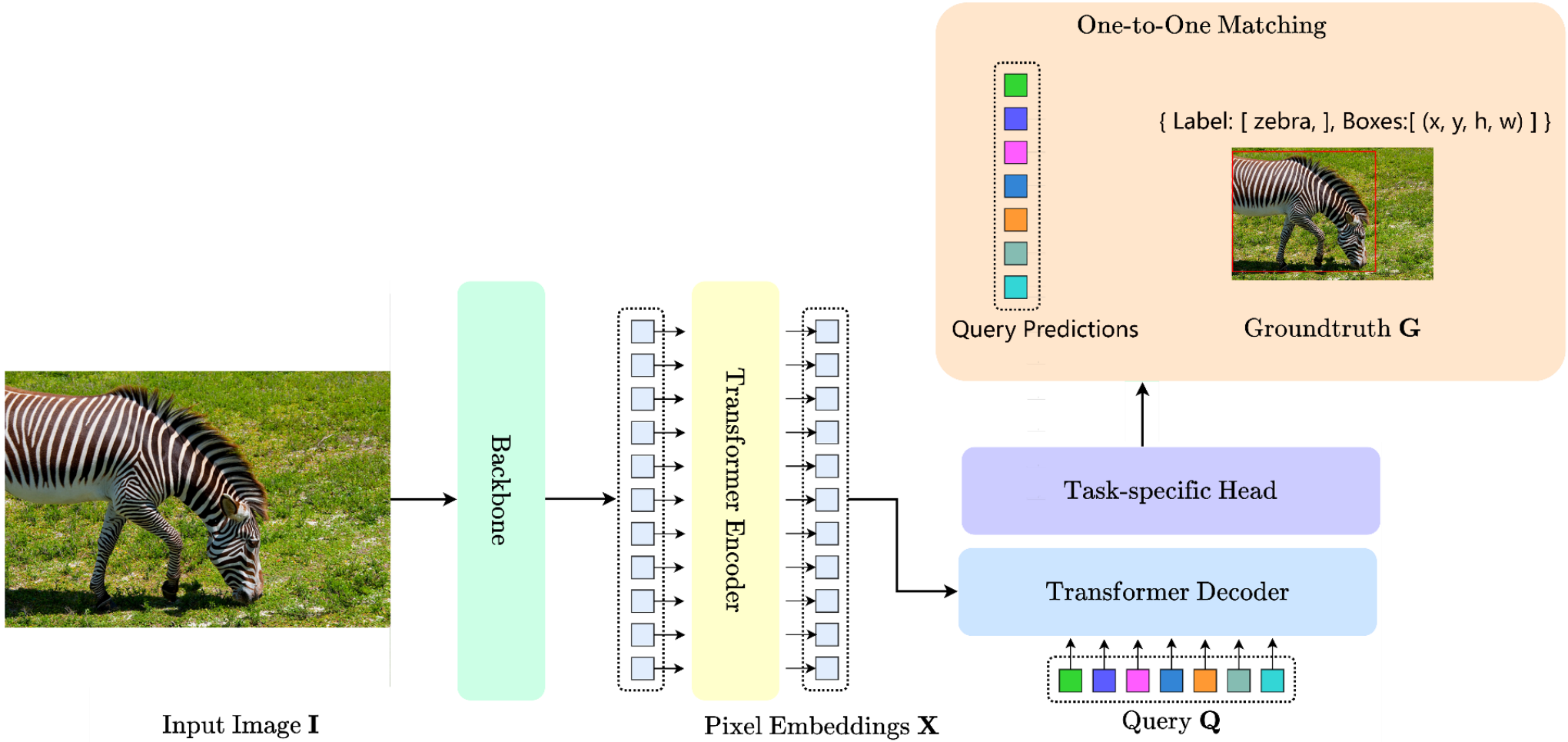
THU-AM-306

Code

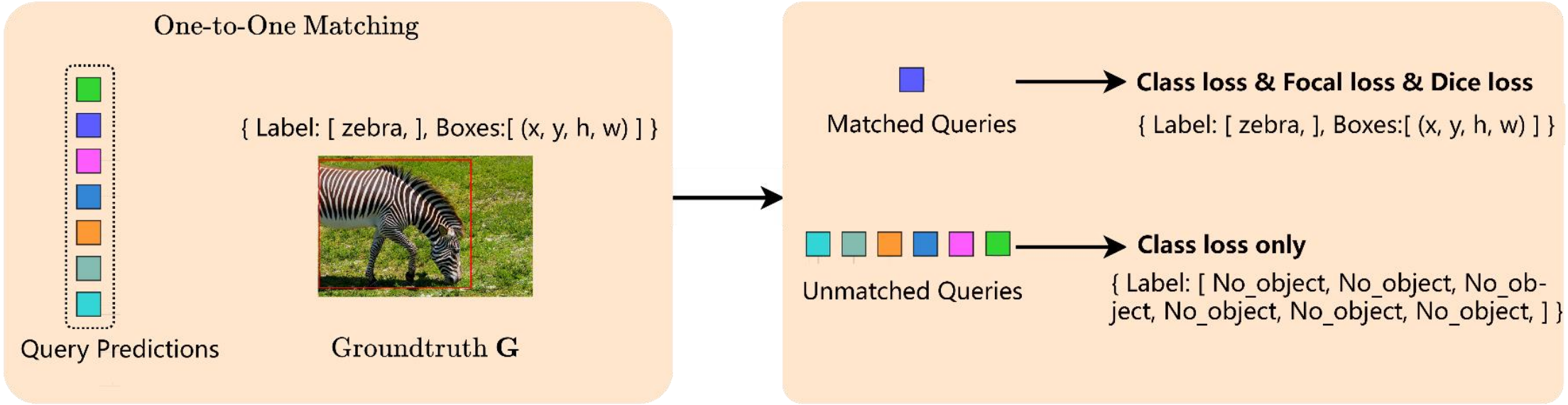
Paper

Contact: yuyua@microsoft.com
Slide is prepared by Ding Jia

The conventional DETR framework

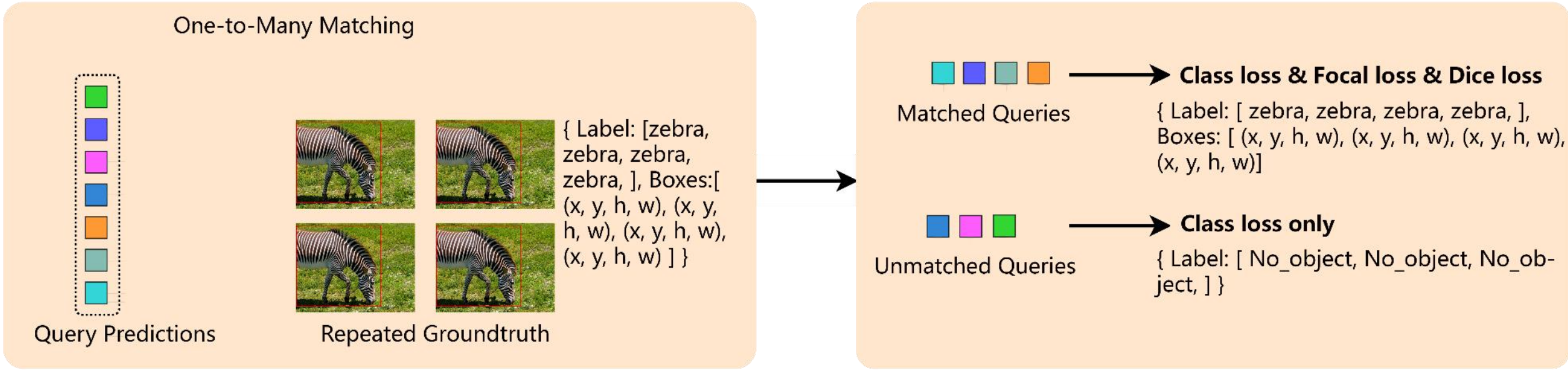


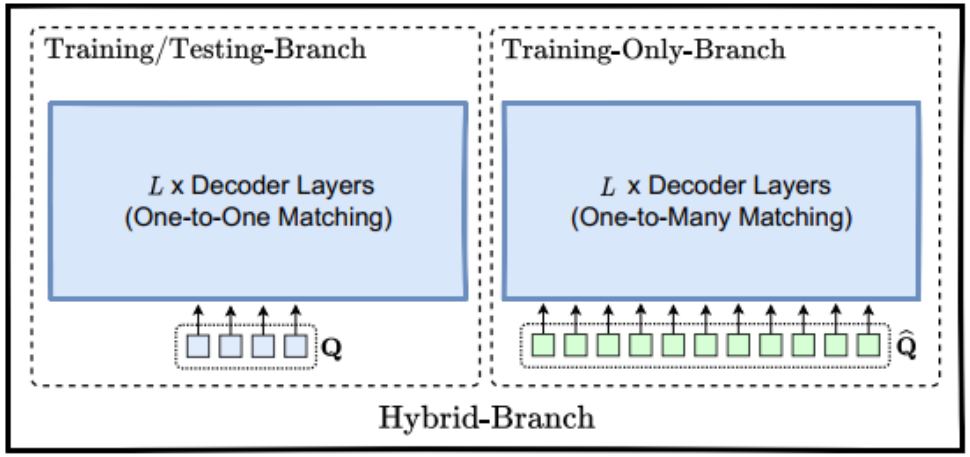
Hybrid Matching Framework

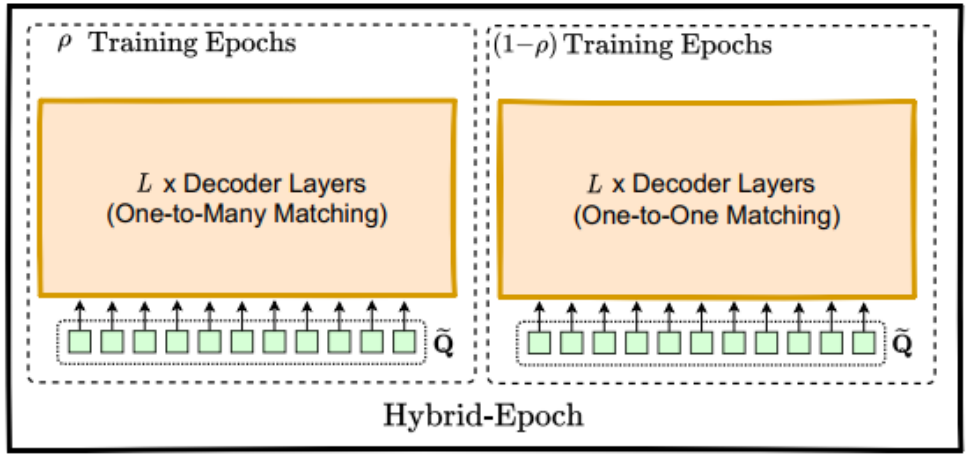


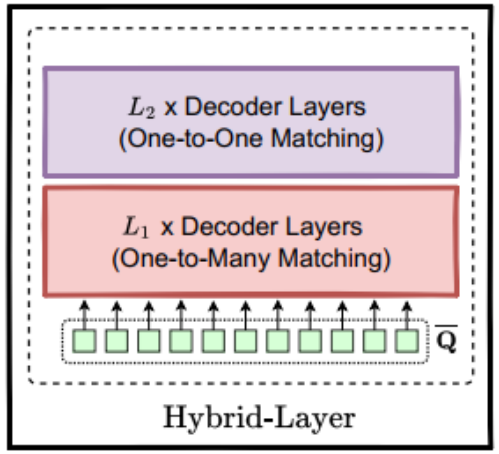
- Deformable-DETR typically only selects less than 30 queries from a pool of 300 queries
- Nearly 99% of the COCO images consist of less than 30 bounding boxes annotations
- Make queries suffer from very limited localization supervision

One-to-Many matching

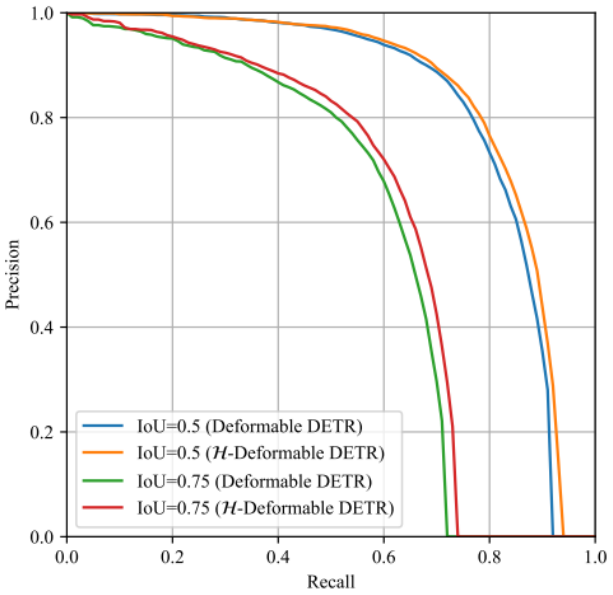








- Hybrid-Branch:
  - Training: One-to-one matching branch & One-to-many matching branch
  - Eval: One-to-one matching branch only
- Hybrid-Epoch:
  - One-to-many matching training epochs first
  - One-to-one matching training epochs then
- Hybrid-Layer:
  - One-to-many matching decoder layers first
  - One-to-one matching decoder layers then



method	inference GLOPs	train time (average)	inference FPS	# epochs		
				12	24	36
Deformable-DETR	268G	65min	6.7	43.7	46.4	46.8
Deformable-DETR[1.3 x]	268G	65min	6.7	44.6	46.3	46.7
Deformable-DETR[†]	282G	75min	6.3	44.1	46.6	47.1
Deformable-DETR[† 1.15 x]	282G	75min	6.3	44.5	46.7	46.9
Deformable-DETR+Hybrid-Branch	268G	80min	6.7	45.9	47.6	48.0
Deformable-DETR+Hybrid-Epoch[†]	282G	95min	6.3	45.5	47.0	47.9
Deformable-DETR+Hybrid-Layer[†]	282G	100min	6.3	45.6	47.9	48.0

2D Object detection results

- COCO

method	backbone	# epochs	AP	AP_S	AP_M	AP_L
Deformable-DETR	R50	12	47.0	29.1	50.0	61.6
H-Deformable-DETR	R50	12	$48.7_{+1.7}$	31.2	51.5	63.5
Deformable-DETR	Swin-L	36	56.3	39.2	60.4	71.8
H-Deformable-DETR	Swin-L	36	$57.1_{+0.8}$	39.7	61.4	73.4

- LVIS v1.0

method	backbone	# epochs	AP	AP_S	AP_M	AP_L
Deformable-DETR	R50	24	32.2	23.2	41.6	49.3
H-Deformable-DETR	R50	24	$33.5_{+1.3}$	24.1	42.4	50.2
Deformable-DETR	Swin-L	48	47.0	35.9	57.8	66.9
H-Deformable-DETR	Swin-L	48	$47.9_{+0.9}$	36.3	58.6	67.9

Multi-view 3D object detection results

method	backbone	# epochs	mAP	NDS
PETRv2	VoVNet-99	24	41.04	50.25
PETRv2 (Our repro.)	VoVNet-99	24	40.41	49.69
H-PETRv2	VoVNet-99	24	$41.93_{+1.52}$	51.23
PETRv2 (Our repro.)	VoVNet-99	36	41.07	50.68
H-PETRv2	VoVNet-99	36	$42.59_{+1.52}$	52.38

Multi-person pose estimation results

method	backbone	# epochs	AP	AP_M	AP_L
PETR	R50	100	68.8	62.7	77.7
PETR (Our repro.)	R50	100	69.3	63.3	78.4
H-PETR	R50	100	$70.9_{+1.6}$	64.4	80.3
PETR	R101	100	70.0	63.6	79.4
PETR (Our repro.)	R101	100	69.9	63.4	79.4
H-PETR	R101	100	$71.0_{+1.1}$	64.7	80.2
PETR	Swin-L	100	73.1	67.2	81.7
PETR (Our repro.)	Swin-L	100	73.3	67.7	81.6
H-PETR	Swin-L	100	$74.9_{+1.6}$	69.3	83.3

Multi-object tracking results

method	# epochs	MOTA ↑	IDF1 ↑	IDF1 ↓
MOT17 val				
TransTrack	20	67.1	70.3	15820
TransTrack (Our repro.)	20	67.1	68.1	15680
H-TransTrack	20	$68.7_{+1.6}$	68.3	13657
MOT17 val				
TransTrack	20	74.5	63.9	112137
TransTrack (Our repro.)	20	74.6	63.2	111105
H-TransTrack	20	$75.7_{+1.1}$	64.4	91155

GPU memory & FLOPS

backbone	method	# query [hyper-parameter]	GFLOPs	Training time (min)	GPU memory (M)
ResNet50 (Swin-L)	Baseline	300 [n=300, T=0, K=0]	268.19 (912.29)	65 (202)	5480 (8955)
	Ours	600 [n=300, T=300, K=6]	271.24 (915.34)	71 (205)	5719 (9190)
		900 [n=300, T=600, K=6]	274.03 (918.12)	72 (208)	6045 (9530)
		1200 [n=300, T=900, K=6]	276.82 (920.91)	75 (210)	6528 (10006)
		1500 [n=300, T=1200, K=6]	279.60 (923.69)	78 (213)	7071 (10558)
		1800 [n=300, T=1500, K=6]	282.39 (926.48)	80 (215)	7728 (11203)

Comparison with State-of-the-art

method	framework	backbone	# epochs	AP					
				AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Swin	HTC	Swin-L	36	57.1	75.6	62.5	42.4	60.7	71.1
CBNetV2	HTC	2× Swin-L	12	59.1	-	-	-	-	-
ConvNeXt	Cascade Mask R-CNN	ConvNeXt-XL	36	55.2	74.2	59.9	-	-	-
MViTv2	Cascade Mask R-CNN	MViTv2-L	50	55.8	74.3	64.3	-	-	-
MOAT	Cascade Mask R-CNN	MOAT-3	36	59.2	**77.8**	60.9	-	-	-
Group-DETR	DETR	Swin-L	36	58.4	-	-	41.0	62.5	73.9
DINO-DETR	DETR	Swin-L	36	58.5	77.0	64.1	41.5	62.3	74.0
H-Deformable-DETR	DETR	Swin-L	36	**59.4**	**77.8**	**65.4**	**43.1**	**63.1**	**74.2**

Thanks
Q&A