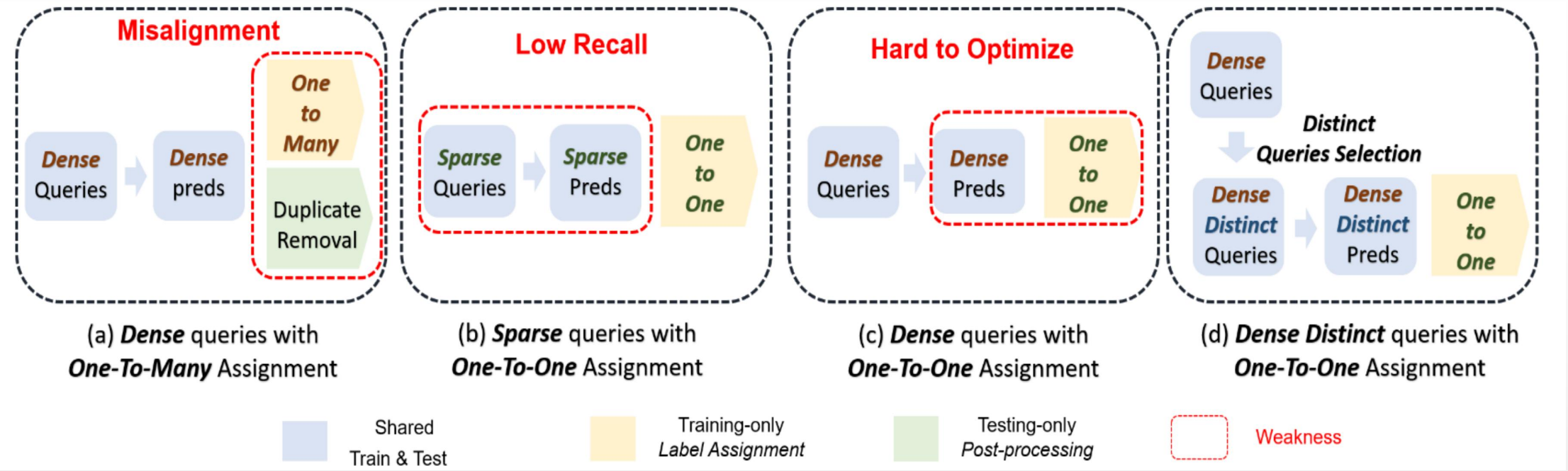
香 港 大 學
THE UNIVERSITY OF HONG KONG

OpenMMLab

Dense Distinct Query for End-to-End Object Detection

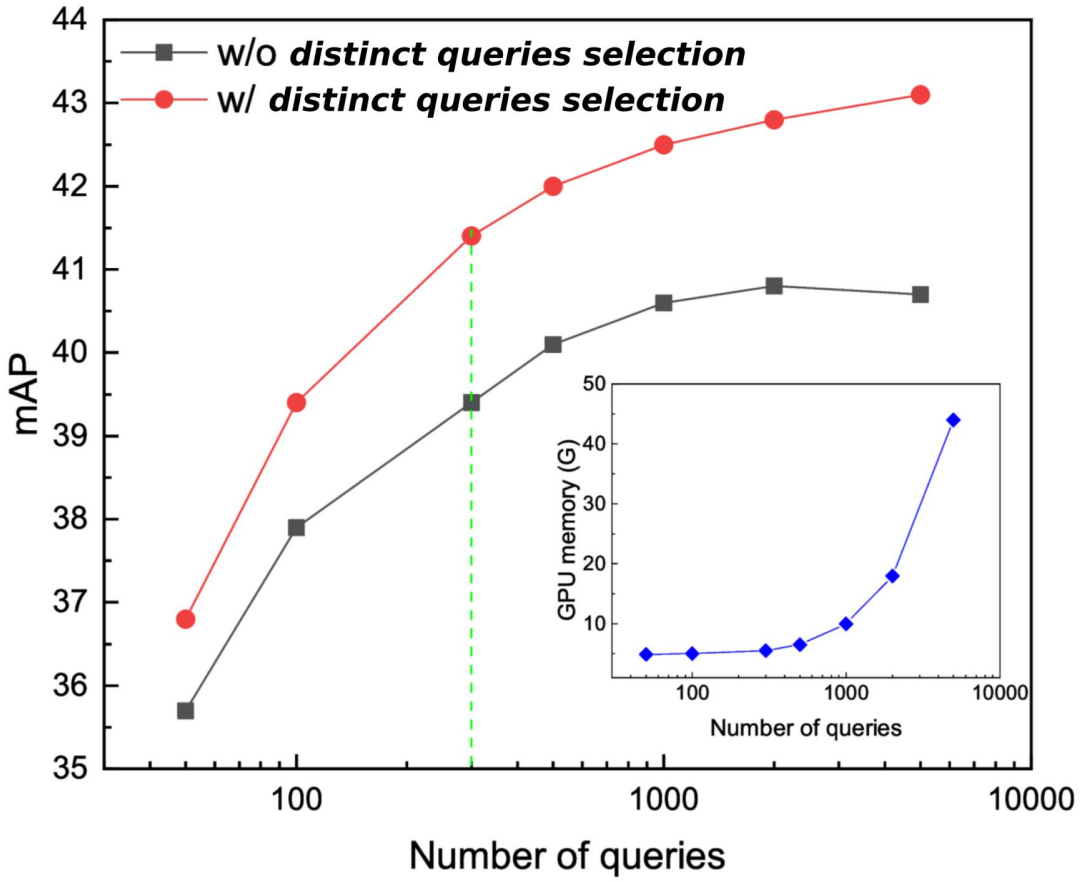
Presenter： Shilong Zhang

Background: Pros and Cons of existing detection methods



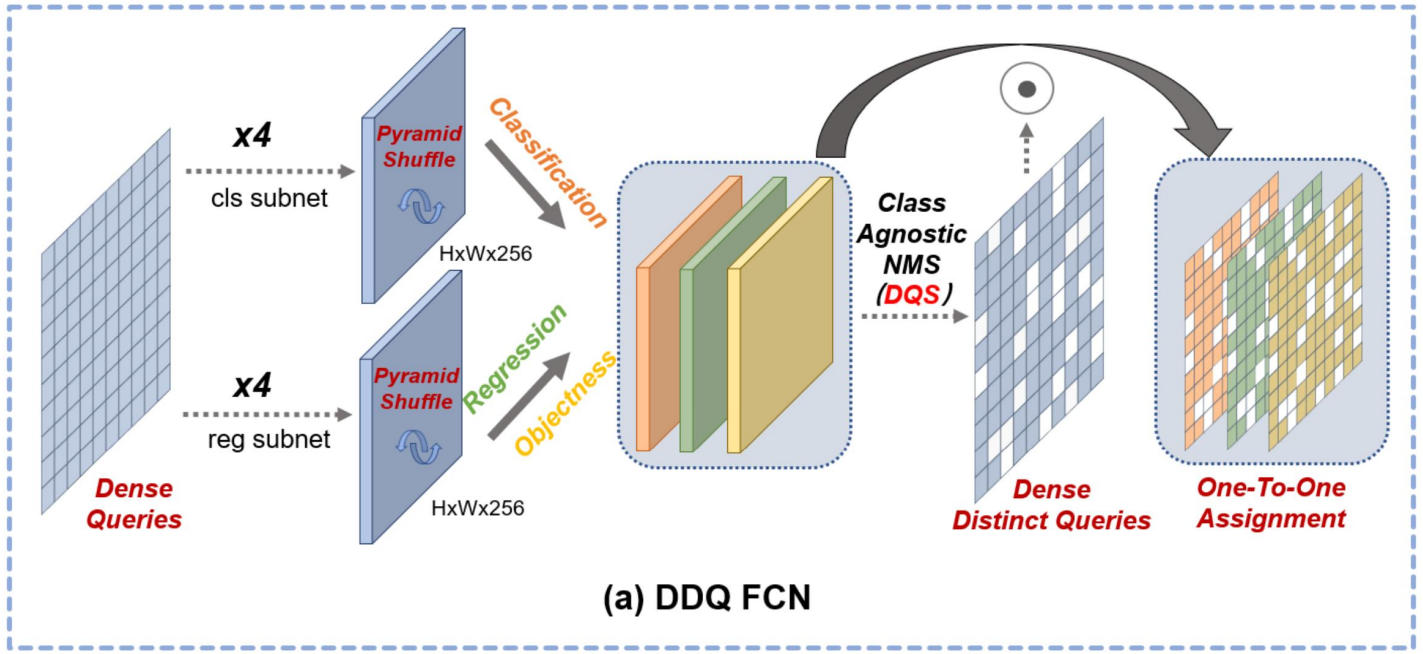
- ***(a) Misalignment & Not applicable for crowded scenes***
  RetinaNet, FCOS, Faster R-CNN, Cascade R-CNN, etc.
- ***(b) Low recall caused by sparse queries***
  DETR, Sparse R-CNN, Deformable DETR, etc.
- ***(c) Hard to optimize due to opposite labels for similar queries.***
  Two-Stage Deformable DETR, Efficient DETR, etc.

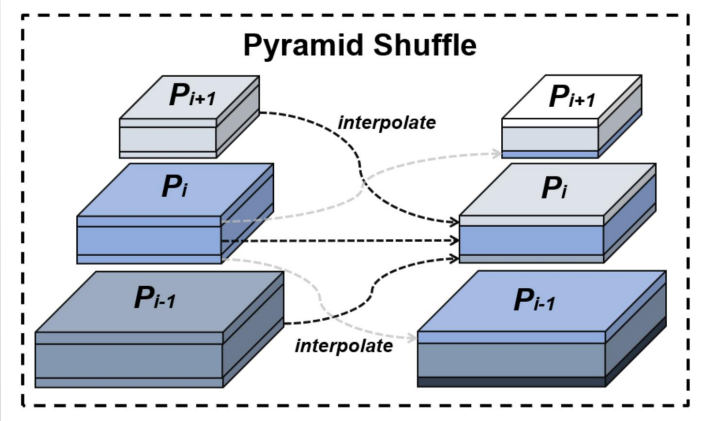
Analysis： Sparse and Dense Queries in end-to-end detectors



- Sparse R-CNN's performance initially improves with increasing queries (higher recall), but eventually drops due to similar queries hindering optimization.
- When ensuring that queries are distinct from each other, the performance of Sparse R-CNN can consistently increase with the number of queries

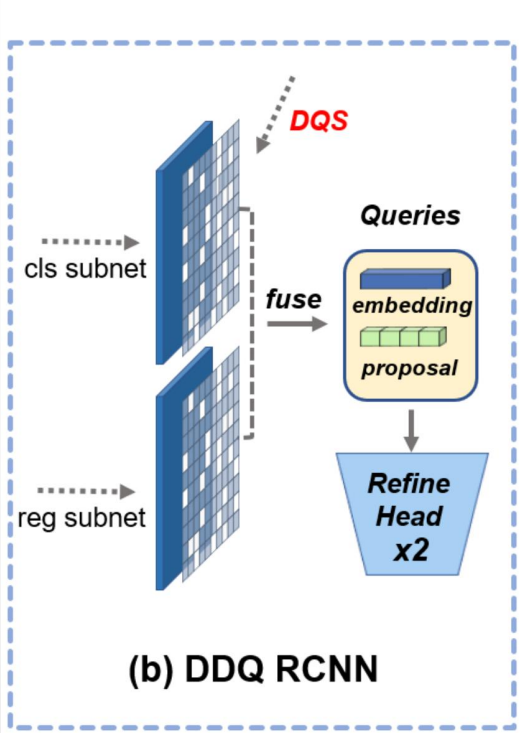
Improving Your Detector with DDQ : DDQ FCN

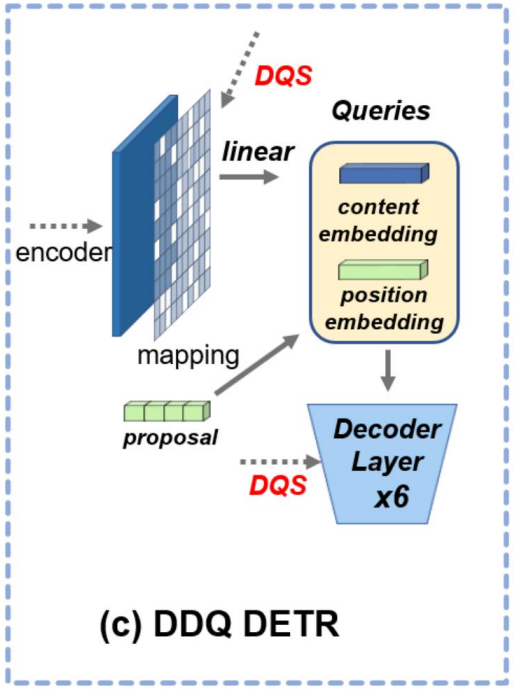




Method	AP	AP_{50}	AP_{75}
FCOS*	36.5	54.4	40.3
+ **PS**	37.6	56.3	41.3
+**DQS**	40.6	60.3	44.5
DDQ FCN	41.5	60.9	45.4

Improving Your Detector with DDQ : DDQ R-CNN & DDQ DETR





Method	AP	AP_{50}	AP_{75}
Sparse R-CNN	39.4	57.7	42.5
+7000Q	40.6	58.7	44.0
+DQS	43.1	62.6	47.1
DDQ R-CNN	44.6	63.0	48.8

Method	AP	AP_{50}	AP_{75}
D-DETR*	45.4	63.0	49.1
TS D-DETR	46.7	64.5	50.8
+Dense	48.5	66.2	52.7
+AUX-Decoder	50.0	67.4	54.8
+DQS	50.7	68.1	55.7
DDQ $DETR_{5scale}$	52.1	68.9	57.3

Results on COCO & CrowdHuman

Method	Backbone	Val/Test	Epochs	AP	AP_{50}	AP_{75}
Aug:DETR						
Cascade R-CNN [1]	ResNet-50	val	36	44.3	62.4	48
DAB DETR [19]	ResNet-50	val	50	42.6	63.2	45.6
DN-DETR [14]	ResNet-50	val	50	44.1	64.4	46.7
Deformable DETR [38]	ResNet-50	val	50	46.2	65.2	50.0
Efficient DETR [32]	ResNet-50	val	36	44.2	62.2	48.0
Sparse R-CNN [26]	ResNet-50	val	36	45.0	63.4	48.2
$DINO_{4scales}$ [33]	ResNet-50	val	36	50.9	69.0	55.3
$DINO_{5scales}$ [33]	ResNet-50	val	36	51.2	69.0	55.8
DDQ FCN	ResNet-50	val	36	**44.8**	64.1	49.4
DDQ R-CNN	ResNet-50	val	36	**48.1**	66.6	53.0
DDQ R-CNN$_{with\_encoder}$	ResNet-50	val	36	**51.0**	69.0	56.0
DDQ DETR$_{4scales}$	ResNet-50	val	24	**52.0**	69.5	57.2
DDQ DETR$_{5scales}$	ResNet-50	val	24	**52.8**	69.9	58.1

Table 4. Performance on CrowdHuman

Method	Epochs	AP_{50}	mMR	Recall
ATSS	36	89.6	44.4	95.9
DW	36	89.0	57.6	97.4
Cascade R-CNN	36	86.0	44.1	89.2
Sparse R-CNN	50	89.2	48.3	95.9
Deform DETR	50	89.1	50.0	95.3
DeFCN	36	91.0	46.5	97.9
DDQ FCN	36	**92.7**	**41.0**	**98.2**
DDQ R-CNN	36	**93.5**	**40.4**	**98.6**
DDQ DETR	36	**93.8**	**39.7**	**98.7**

Ablation Study

The Recall Improvement of Dense Queries

Method	AR_{100}	AR_{200}	AR_{300}	L(ms)
Sparse R-CNN	78.4	83.4	85.5	31.0
7000 Q & DQS	88.6	92.3	93.6	135.0
DDQ R-CNN	88.5	91.8	93.2	31.3

DQS with Different IoU Threshold

COCO	0.5	0.6	0.7	0.8	0.9	None
DDQ FCN	40.8	41.4	**41.5**	41.4	40.5	39.5*
DDQ R-CNN	44.0	44.5	**44.6**	44.4	43.8	42.7*
DDQ DETR	50.1	50.7	50.9	**51.3**	51.0	50.7*
ATSS	39.3	**39.5**	39.3	38.7	36.7	19.6

Thanks