

Table 4: Comparisons with concurrent works, including YOLO-MS [12], Gold-YOLO [58], YOLOv10 [57], and RT-DETR [45] on COCO. For YOLO-MS and Gold-YOLO, we measure the total latency in two settings for NMS: official implementation and tuned score threshold. For YOLOv10, we report the results in the official paper [57]. RT-DETR is based on DETR, and the total latency is equal to the model latency. We provide the best latency among the reported inference time in the paper [45] and the measured time in our environment for RT-DETR. LW-DETR consistently gets superior results. ‘pretraining’ means that the results are based on pretraining on Objects365.

Method	pretraining	#Params (M)	FLOPs (G)	Model Latency (ms)	official implementation		tuned score threshold	
					Total Latency (ms)	mAP	Total Latency (ms)	mAP
YOLO-MS-XS		4.5	8.7	3.0	6.9	43.4	3.2	43.3
YOLO-MS-XS	✓	4.5	8.7	3.0	6.9	43.9	3.2	43.8
YOLO-MS-S		8.1	15.6	5.4	9.2	46.2	5.6	46.1
YOLO-MS-S	✓	8.1	15.6	5.4	9.2	46.8	5.6	46.7
YOLO-MS		22.0	40.1	8.6	12.3	51.0	9.0	50.8
Gold-YOLO-S		21.5	23.0	2.9	3.6	45.5	3.4	45.4
Gold-YOLO-S	✓	21.5	23.0	2.9	3.6	46.1	3.4	46.0
Gold-YOLO-M		41.3	43.8	5.8	6.3	50.2	6.1	50.2
Gold-YOLO-M	✓	41.3	43.8	5.8	6.3	50.4	6.1	50.3
Gold-YOLO-L		75.1	75.9	10.2	10.6	52.3	10.5	52.2
YOLOv10-N		2.3	6.7	-	1.84	38.5	-	-
YOLOv10-S		7.2	21.6	-	2.49	46.3	-	-
YOLOv10-M		15.4	59.1	-	4.74	51.1	-	-
YOLOv10-B		19.1	92.0	-	5.74	52.5	-	-
YOLOv10-L		24.4	120.3	-	7.28	53.2	-	-
YOLOv10-X		29.5	160.4	-	10.70	54.4	-	-
RT-DETR-R18		20	30.0	4.6	4.6	46.5	-	-
RT-DETR-R18	✓	20	30.0	4.6	4.6	49.2	-	-
RT-DETR-R50		42	69.4	9.3	9.3	53.1	-	-
RT-DETR-R50	✓	42	69.4	9.3	9.3	55.3	-	-
RT-DETR-R101		76	131.0	13.5	13.5	54.3	-	-
RT-DETR-R101	✓	76	131.0	13.5	13.5	56.2	-	-
LW-DETR-tiny	✓	12.1	11.2	2.0	2.0	42.6	-	-
LW-DETR-small	✓	14.6	16.6	2.9	2.9	48.0	-	-
LW-DETR-medium	✓	28.2	42.8	5.6	5.6	52.5	-	-
LW-DETR-large	✓	46.8	71.6	8.8	8.8	56.1	-	-
LW-DETR-xlarge	✓	118.0	174.2	19.1	19.1	58.3	-	-