A Additional experiment

A.1 More detailed accuracy and speed data for GD-YOLO

In this section, we report the test performace of our GD-YOLO with or without LAF module and pre-training. FPS and latency are measured in FP16-precision on a Tesla T4 in the same environment with TensorRT 7. Both the accuracy and the speed performance of our models are evaluated with the input resolution of 640x640. The result shown in Table 1 .

Table 1: Test results of GD-YOLO series model on COCO 2017 val. '†' represents that the self-distillation method is utilized, '$\diamond$' represents that the model don't have LAF module, and '$\star$' represents that the MIM pre-training method is utilized.

Method	$\mathbf{AP}^{val}$	$\mathbf{AP}^{val}_{50}$	$\mathbf{AP}^{val}_{small}$	$\mathbf{AP}^{val}_{medium}$	$\mathbf{AP}^{val}_{large}$	**FPS** $_{(bs=32)}$	**Params**	**FLOPs**
GD-YOLO-N$^{\diamond}$	38.37% / 38.82%†	54.43% / 54.96%†	18.37% / 18.40%†	42.62% / 43.45%†	55.13% / 56.00%†	1087	5.6 M	12.0 G
GD-YOLO-S$^{\diamond}$	44.47% / 45.57%†	61.55% / 62.92%†	24.04% / 24.90%†	49.34% / 50.38%†	61.95% / 63.50%†	462	21.5 M	45.8 G
GD-YOLO-M$^{\diamond}$	49.41% / 50.26%†	66.64% / 67.58%†	31.19% / 31.59%†	54.30% / 55.26%†	65.91% / 67.62%†	229	41.3 M	86.8 G
GD-YOLO-L$^{\diamond}$	51.68% / 52.65%†	69.07% / 70.25%†	34.86% / 34.20%†	56.92% / 57.70%†	69.00% / 69.71%†	119	75.0 M	150.6 G
GD-YOLO-N	39.57% / 39.92%†	55.70% / 55.94%†	19.67% / 19.15%†	44.08% / 44.32%†	56.98% / 57.75%†	1030	5.6 M	12.1 G
GD-YOLO-S	45.36% / 46.11%†	62.48% / 63.33%†	25.32% / 25.22%†	50.21% / 51.23%†	62.63% / 63.42%†	446	21.5 M	46.0 G
GD-YOLO-M	49.77% / 50.86%†	67.01% / 68.23%†	32.32% / 31.01%†	55.29% / 56.24%†	66.27% / 67.83%†	220	41.3 M	87.5 G
GD-YOLO-L	51.84% / 53.16%†	68.94% / 70.49%†	34.12% / 34.53%†	57.36% / 58.60%†	68.17% / 70.07%†	116	75.1 M	151.7 G
GD-YOLO-S*	45.52% / 46.36%†	62.20% / 63.36%†	24.66% / 25.26%†	50.76% / 51.30%†	63.24% / 63.64%†	446	21.5 M	46.0 G
GD-YOLO-M*	50.16% / 51.14%†	67.52% / 68.53%†	30.52% / 32.33%†	55.54% / 56.10%†	67.64% / 68.55%†	220	41.3 M	87.5 G
GD-YOLO-L*	52.25% / 53.28%†	69.61% / 70.93%†	33.09% / 33.83%†	57.77% / 58.92%†	69.01% / 69.92%†	116	75.1 M	151.7 G

A.2 MIM pre-training ablation experiment

We also compared the GD-YOLO-S on COCO 2017 validation results for different MIM pre-training epochs without self-distillation. The result shown in Table 2 .

Table 2: Test results on COCO 2017 val for different pre-training epoch setting.

Epoch	$\mathbf{AP}^{val}$	$\mathbf{AP}^{val}_{50}$	$\mathbf{AP}^{val}_{small}$	$\mathbf{AP}^{val}_{medium}$	$\mathbf{AP}^{val}_{large}$
400	45.39%	62.17%	25.01%	50.28%	62.74%
600	45.48%	62.18%	25.56%	50.63%	62.85%
800	45.52%	62.20%	24.66%	50.76%	63.24%

B Comprehensive Latency and Throughput Benchmark

B.1 Model Latency and Throughput on T4 GPU with TensorRT 8

Comparisons with other YOLO-series detectors on COCO 2017 val. FPS and latency are measured in FP16-precision on Tesla T4 in the same environment with TensorRT 8.2. The result shown in Table 3.

Table 3: Comparison of Latency and Throughput in YOLO series model on a T4 GPU using TensorRT 8.2.

Method	**Input Size**	**FPS** $(bs{=}1)$	**FPS** $(bs{=}32)$	**Latency** $(bs{=}1)$
YOLOv5-N	640	702	843	1.4 ms
YOLOv5-S	640	433	515	2.3 ms
YOLOv5-M	640	202	235	4.9 ms
YOLOv5-L	640	126	137	7.9 ms
YOLOX-Tiny	416	766	1393	1.3 ms
YOLOX-S	640	313	489	2.6 ms
YOLOX-M	640	159	204	5.3 ms
YOLOX-L	640	104	117	9.0 ms
PPYOLOE-S	640	357	493	2.8 ms
PPYOLOE-M	640	163	210	6.1 ms
PPYOLOE-L	640	110	145	9.1 ms
YOLOv7-Tiny	640	464	568	2.1 ms
YOLOv7	640	128	135	7.6 ms
YOLOv6-3.0-N	640	785	1215	1.3 m s
YOLOv6-3.0-S	640	345	498	2.9 ms
YOLOv6-3.0-M	640	178	238	5.6 ms
YOLOv6-3.0-L	640	105	125	9.5 ms
GD-YOLO-N	640	657	1191	1.4 ms
GD-YOLO-S	640	308	492	3.1 ms
GD-YOLO-M	640	157	241	6.1 ms
GD-YOLO-L	640	94	137	10.3 ms

B.2 Model Latency and Throughput on V100 GPU with TensorRT 7

Comparisons with other YOLO-series detectors on COCO 2017 val. FPS and latency are measured in FP16-precision on Tesla V100 in the same environment with TensorRT 7.2. The result shown in Table 4 .

Table 4: Comparison of Latency and Throughput in YOLO series model on a V100 GPU using TensorRT 7.2.

Method	**Input Size**	**FPS** $(bs{=}1)$	**FPS** $(bs{=}32)$	**Latency** $(bs{=}1)$
YOLOv5-N	640	577	1727	1.4 ms
YOLOv5-S	640	449	1249	1.7 ms
YOLOv5-M	640	271	698	3.0 ms
YOLOv5-L	640	178	440	4.7 ms
YOLOX-Tiny	416	569	2883	1.4 ms
YOLOX-S	640	386	1206	2.0 ms
YOLOX-M	640	245	600	3.4 ms
YOLOX-L	640	149	361	5.6 ms
PPYOLOE-S	640	322	1050	2.4 ms
PPYOLOE-M	640	222	566	4.0 ms
PPYOLOE-L	640	153	406	5.5 ms
YOLOv7-Tiny	640	453	1565	1.7 ms
YOLOv7	640	182	412	4.6 ms
YOLOv6-3.0-N	640	646	2660	1.2 m s
YOLOv6-3.0-S	640	399	1330	2.0 ms
YOLOv6-3.0-M	640	203	676	4.4 ms
YOLOv6-3.0-L	640	125	385	6.8 ms
GD-YOLO-N	640	574	2457	1.7 ms
GD-YOLO-S	640	391	1205	2.5 ms
GD-YOLO-M	640	238	633	4.0 ms
GD-YOLO-L	640	146	365	6.6 ms

C Broader impacts and limitations

Broader impacts. The YOLO model can be widely applied in fields such as healthcare and intelligent transportation. In the healthcare domain, the YOLO series models can improve the early diagnosis rates of certain diseases and reduce the cost of initial diagnosis, thereby saving more lives. In the field of intelligent transportation, the YOLO model can assist in autonomous driving of vehicles, enhancing traffic safety and efficiency. However, there are also risks associated with the military application of the YOLO model, such as target recognition for drones and assisting military reconnaissance. We will make every effort to prevent the use of our model for military purposes.

Limitations. Generally, making finer adjustments on the structure will help further improve the model's performance, but this requires a significant amount of computational resources. Additionally, due to our algorithm's heavy usage of attention operations, it may not be as friendly to some earlier hardware support.

D CAM visualization

Below are the CAM visualization results of the neck for YOLOv5, YOLOv6, YOLOv7, YOLOv8, and our GD-YOLO, shown in Fig. 1. It can be observed that our model assigns higher weights to the detected regions of the targets.

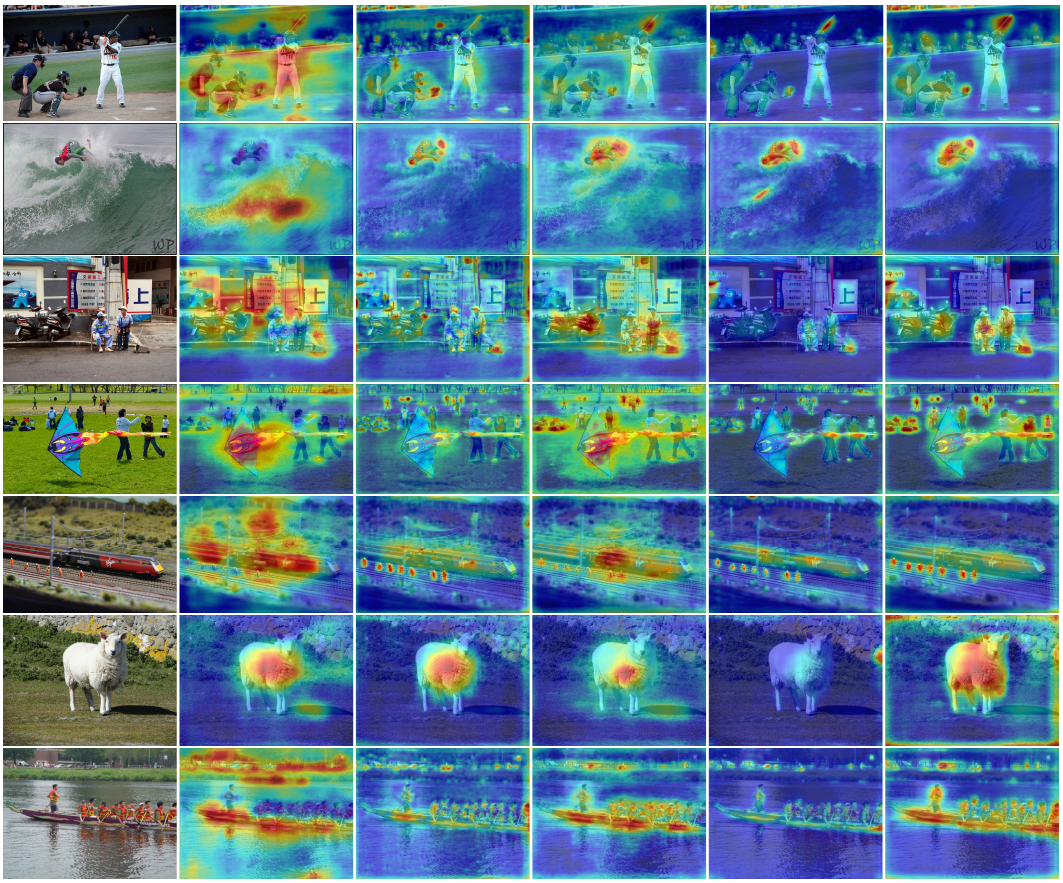

Figure 1: The CAM visualization results of the neck for YOLOv5, YOLOv6, YOLOv7, YOLOv8, and our GD-YOLO.