提示

- 利用区域入侵检测算法测试yolov5和yolov8各版本模型在视频流推理时占用资源情况
- yolov5和yolov8官方指标对比，可以看出yolov5的mAP指标与yolov8相近

Model	size (pixels)	mAP^{val} 50-95
YOLOv8n	640	37.3
YOLOv8s	640	44.9
YOLOv8m	640	50.2
YOLOv8l	640	52.9
YOLOv8x	640	53.9

Model	size (pixels)	mAP^{val} 0.5:0.95
YOLOv5n	640	28.0
YOLOv5s	640	37.4
YOLOv5m	640	45.4
YOLOv5l	640	49.0
YOLOv5x	640	50.7
YOLOv5n6	1280	36.0
YOLOv5s6	1280	44.8
YOLOv5m6	1280	51.3
YOLOv5l6	1280	53.7

- **下面的测试结果中，未标明路数时即表示1路**
- **随着路数的增加通过观察推理耗时的变化可判断推理延时情况**
- **所有的模型均用同一个视频测试，信息如下,共65*30=1950帧**

类型	VLC media file (.mp4)
大小	29.6 MB
文件位置	C:\用户\wmingdru\桌面\work...
修改日期	2024/7/4 14:36
时长	00:01:05
帧宽度	1920
帧高度	1080
帧速率	30.00 帧/秒
数据速率	3631kbps
总比特率	3759kbps

- **测试的显卡信息如下**

```
+-----------------------------------------+
|   1  NVIDIA GeForce RTX 3090        Off |
| 50%   43C    P2             113W / 350W |
|                                         |
```

1.过程记录

1.1 yolov5

1.1.1 yolov5n.pt 模型大小3.8Mb 推理耗时55s 即35.5帧/s

```
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 50%   38C    P2             112W / 350W |    344MiB / 24576MiB |      5%      Default |
|                                         |                      |                  N/A |
```

50路，平均推理耗时250s,即8帧/s

```
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 49%   45C    P2             171W / 350W |  17059MiB / 24576MiB |     78%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
```

1.1.2 yolov5n6.pt 模型大小6.8Mb(后面带上6的模型不不带的优化了对高分辨率图像的处理效果，但模型更大) 推理耗时68

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 50%   38C    P2             112W / 350W |    348MiB / 24576MiB |      7%      Default |
|                                         |                      |                  N/A |
```

1.1.3 yolov5s.pt 模型大小14.1Mb 推理耗时56s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 50%   38C    P2             113W / 350W |    370MiB / 24576MiB |      6%      Default |
|                                         |                      |                  N/A |
```

1.1.4 yolov5m.pt 模型大小 40.8Mb 推理耗时63s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 49%   37C    P2             115W / 350W |    432MiB / 24576MiB |      9%      Default |
|                                         |                      |                  N/A |
```

1.1.5 yolov5l.pt 模型大小89.2 推理耗时76s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A
| 50%   40C    P2             120W / 350W |    522MiB / 24576MiB |     12%      Default
|                                         |                      |                  N/A
```

1.1.6 yolov5l6.pt 模型大小147.3 推理耗时76s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A
| 50%   41C    P2             122W / 350W |    650MiB / 24576MiB |     16%      Default
|                                         |                      |                  N/A
```

1.1.7 以上面记录的的模型yolov5l.pt做多路压力测试

5路，平均推理耗时89s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 49%   60C    P2             225W / 350W |   2597MiB / 24576MiB |     63%      Default |
|                                         |                      |                  N/A |
```

10路，平均推理耗时140s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A
| 49%   41C    P2             227W / 350W |   5194MiB / 24576MiB |     74%      Default
|                                         |                      |                  N/A
```

15路，平均推理耗时193s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A
| 49%   54C    P2             245W / 350W |   7789MiB / 24576MiB |     80%      Default
|                                         |                      |                  N/A
```

20路，平均推理耗时244s，即8帧/s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 67%   63C    P2             261W / 350W |  10385MiB / 24576MiB |     91%      Default |
|                                         |                      |                  N/A |
```

30路，平均推理耗时350s,即5.5帧/s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 61%   63C    P2             262W / 350W |  15577MiB / 24576MiB |     90%      Default |
|                                         |                      |                  N/A |
```

45路，平均推理耗时517s，即3.7帧/s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 60%   62C    P2             272W / 350W |  23363MiB / 24576MiB |    100%      Default |
|                                         |                      |                  N/A |
```

46路，平均推理耗时529s，即3.7帧/s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A
| 49%   61C    P2             270W / 350W |  23882MiB / 24576MiB |     89%      Default
|                                         |                      |                  N/A
```

1.1.8 yolov5结论

随着路数的增加，推理时间成比例增加，即延迟增加。但路数增加到即便显存即将占满时，算力并没有时刻被占满，而是在80%-100%之间跳动。所以，虽然算力没有被占满，但路数的增加还是会导致推理速度成比例地下降。

1.2 yolov8

1.2.1 yolov8n 模型大小6.2Mb 推理耗时51s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A
| 50%   42C    P2             116W / 350W |    432MiB / 24576MiB |      7%      Default
|                                         |                      |                  N/A
```

50路,平均推理耗时255s，即8帧/s

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 49%   59C    P2             211W / 350W |  21459MiB / 24576MiB |     86%      Default |
|                                         |                      |                  N/A |
```

1.2.2 yolov8m.pt 模型大小49.7Mb

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 51%   52C    P2             134W / 350W |    568MiB / 24576MiB |     21%      Default |
|                                         |                      |                  N/A |
```

1.2.3 yolov8s.pt 模型大小21.5Mb

```
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 51%   53C    P2             123W / 350W |    476MiB / 24576MiB |     11%      Default |
|                                         |                      |                  N/A |
```

1.2.4 yolov8l.pt 模型大小83.7Mb 推理耗时58s

```
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 50%   48C    P2             143W / 350W |    684MiB / 24576MiB |     26%      Default |
|                                         |                      |                  N/A |
```

35路，平均推理耗时555s，即3.5帧/s

```
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA GeForce RTX 3090        Off | 00000000:38:00.0 Off |                  N/A |
| 52%   62C    P2             319W / 350W |  23841MiB / 24576MiB |     98%      Default |
|                                         |                      |                  N/A |
+-----------------------------------------+----------------------+----------------------+
```

2.结论

- **不应关注GPU算力占比，而应该关注实际测得的随着路数的增加推理耗时的情况**
- **同等大小的模型，yolov5比yolov8占用更少的显存**