# Weed Detection with Yolov8n Optimization

## 1) Performance Metrics Overview

This document provides a performance analysis of YOLOv8n using three different optimization techniques: PyTorch, FP16 (TensorRT), and INT8 (TensorRT). The experiments were conducted on an NVIDIA RTX 3070 Ti GPU

| | **yolov8n** | | |
|---|---|---|---|
| | **pytorch** | **fp16(tensorrt)** | **int 8(tensorrt)** |
| **mAP50** | 0.965 | 0.96 | 0.948 |
| **mAP50-95** | 0.7 | 0.63 | 0.57 |
| **Inference time per image(ms)** | 12.6 | 6.1 | 5.3 |
| **FPS** | 75 | 131 | 148 |
| **size** | 11MiB | 8MiB | 6MiB |

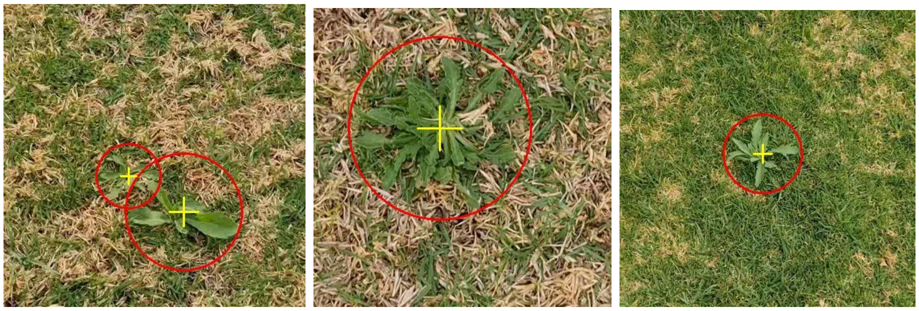
**Note**: After a bit of research it has been found out that rtx 3070ti has been optimized for fp16 operations, hence not much significant improvement from fp16 to int8

## 2) Observations

- **Accuracy (mAP50):** The PyTorch implementation of YOLOv8n achieves the highest accuracy with an mAP50 of 0.965. The accuracy slightly decreases with FP16 and INT8 TensorRT optimizations.
- **Accuracy (mAP50-95):** The PyTorch version also leads in mAP50-95, followed by FP16 and INT8, which show a gradual decline.
- **Inference Speed:** The INT8 TensorRT optimized model delivers the fastest inference speed at 148 FPS, making it highly efficient for real-time applications.
- **Model Size:** The INT8 TensorRT model is the most compact at 6 MiB, which is beneficial for deployment in environments with limited storage.

## 3) Inference

**FP16**



## INT8