# Inference Time

Time spent on a single forward pass. FPS = 1/tinference

### Yolov7-Tiny

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Device | Hardware | Input Size | Model Format | Avg. Inference Time |
| SDW4100 | CPU | 320x320 | .torchscript.ptl | 400 ms |
| SDW4100 | CPU | 640x640 | .torchscript.ptl | 1600 ms |
| Samsung Galaxy S22 | CPU | 320x320 | .torchscript.ptl | 80 ms |
| Samsung Galaxy S22 | CPU | 640x640 | .torchscript.ptl | 200 ms |
| PC | CPU\* | 320x320 | .pt | 90 ms |
| PC | CPU\* | 640x640 | .pt | 150 ms |
| PC | GPU\*\* | 320x320 | .pt | 8 ms |
| PC | GPU\*\* | 640x640 | .pt | 10 ms |

\*AMD Ryzen 5 2600, \*\*Nvidia GTX 1070

### Yolov5s

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Device | Hardware | Input Size | Model Format | Precision | Avg. Inference Time |
| SDW4100 | CPU | 224x224 | .tflite | fp32 | 198 ms |
| SDW4100 | CPU | 224x224 | .tflite | int8 | 132 ms |
| SDW4100 | CPU | 320x320 | .tflite | fp32 | 380 ms |
| SDW4100 | CPU | 320x320 | .tflite | int8 | 245 ms |
| SDW4100 | GPU | 224x224 | .tflite | fp32 | 410 ms |
| SDW4100 | GPU | 224x224 | .tflite | int8 | 218 ms |
| SDW4100 | GPU | 320x320 | .tflite | fp32 | 821 ms |
| SDW4100 | GPU | 320x320 | .tflite | int8 | 438 ms |
| Samsung Galaxy S22 | CPU | 224x224 | .tflite | fp32 | ms |
| Samsung Galaxy S22 | CPU | 320x320 | .tflite | fp32 | ms |
| Samsung Galaxy S22 | CPU | 320x320 | .tflite | int8 | ms |
| Samsung Galaxy S22 | CPU |  | .tflite | fp32 | ms |
| Samsung Galaxy S22 | CPU |  | .tflite | int8 | ms |
| Samsung Galaxy S22 | GPU | 320x320 | .tflite | fp32 | ms |
| Samsung Galaxy S22 | GPU |  | .tflite | fp32 | ms |

### Yolov5n

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Device | Hardware | Input Size | Model Format | Precision | Avg. Inference Time |
| SDW4100 | CPU | 224x224 | .tflite | fp32 | 82 ms |
| SDW4100 | CPU | 224x224 | .tflite | int8 | 80 ms |
| SDW4100 | CPU | 320x320 | .tflite | fp32 | 159 ms |
| SDW4100 | CPU | 320x320 | .tflite | int8 | 130 ms |
| SDW4100 | CPU | 640x640 | .tflite | fp32 | 580 ms |
| SDW4100 | CPU | 640x640 | .tflite | int8 | 490 ms |
| SDW4100 | GPU | 224x224 | .tflite | fp32 | 177 ms |
| SDW4100 | GPU | 224x224 | .tflite | int8 | 124 ms |
| SDW4100 | GPU | 320x320 | .tflite | fp32 | 320 ms |
| SDW4100 | GPU | 320x320 | .tflite | int8 | 218 ms |
| SDW4100 | GPU | 640x640 | .tflite | fp32 | 1250 ms |
| SDW4100 | GPU | 640x640 | .tflite | int8 | 880 ms |
| Samsung Galaxy S22 | CPU | 224x224 | .tflite | fp32 | 16 ms |
| Samsung Galaxy S22 | CPU | 224x224 | .tflite | int8 | 12 ms |
| Samsung Galaxy S22 | CPU | 320x320 | .tflite | fp32 | 31 ms |
| Samsung Galaxy S22 | CPU | 320x320 | .tflite | int8 | 20 ms |
| Samsung Galaxy S22 | CPU | 640x640 | .tflite | fp32 | 98 ms |
| Samsung Galaxy S22 | CPU | 640x640 | .tflite | int8 | 57 ms |
| Samsung Galaxy S22 | GPU | 320x320 | .tflite | fp32 | 65 ms |
| Samsung Galaxy S22 | GPU | 640x640 | .tflite | fp32 | 230 ms |

### Yolov5p

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Device | Hardware | Input Size | Model Format | Precision | Avg. Inference Time |
| SDW4100 | CPU | 320x320 | .tflite | fp32 | 78 ms |
| SDW4100 | CPU | 320x320 | .tflite | int8 | 85 ms |
| SDW4100 | CPU | 640x640 | .tflite | fp32 | 310 ms |
| SDW4100 | CPU | 640x640 | .tflite | int8 | 320 ms |
| SDW4100 | GPU | 320x320 | .tflite | fp32 | 170 ms |
| SDW4100 | GPU | 320x320 | .tflite | int8 | 131 ms |
| SDW4100 | GPU | 640x640 | .tflite | fp32 | 660 ms |
| SDW4100 | GPU | 640x640 | .tflite | int8 | 510 ms |
| Samsung Galaxy S22 | CPU | 320x320 | .tflite | fp32 | 16 ms |
| Samsung Galaxy S22 | CPU | 320x320 | .tflite | int8 | 11 ms |
| Samsung Galaxy S22 | CPU | 640x640 | .tflite | fp32 | 54 ms |
| Samsung Galaxy S22 | CPU | 640x640 | .tflite | int8 | 40 ms |
| Samsung Galaxy S22 | GPU | 320x320 | .tflite | fp32 | 30 ms |
| Samsung Galaxy S22 | GPU | 640x640 | .tflite | fp32 | 115 ms |

# Performance

Precision, recall and mean average precision at IoU threshold 95%/50% on the test set with approx. 100k images. Note that in this case mAP is the same as AP since there is only a single class.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Precision | Recall | mAP@50 | mAP@95 |
| yolov7-tiny\_640\_fp32 | 0.9920 | 0.9873 | 0.9962 | 0.8438 |
| yolov7-tiny\_320\_fp32 | 0.9793 | 0.9714 | 0.9931 | 0.7758 |
| yolov5s\_320\_fp32 | 0.9884 | 0.9817 | 0.9941 | 0.7808 |
| yolov5s\_320\_int8 | 0.965 | 0.967 | 0.991 | 0.710 |
| yolov5s\_224\_fp32 | 0.9814 | 0.9734 | 0.9927 | 0.7639 |
| yolov5s\_224\_int8 | 0.939 | 0.944 | 0.982 | 0.661 |
| yolov5n\_640\_fp32 | 0.9894 | 0.9823 | 0.9940 | 0.7977 |
| yolov5n\_640\_int8 | 0.980 | 0.970 | 0.992 | 0.729 |
| yolov5n\_320\_fp32 | 0.9780 | 0.9584 | 0.9890 | 0.7247 |
| yolov5n\_320\_int8 | 0.947 | 0.934 | 0.980 | 0.649 |
| yolov5n\_224\_fp32 | 0.9662 | 0.9336 | 0.9770 | 0.6859 |
| yolov5n\_224\_int8 | 0.916 | 0.897 | 0.956 | 0.589 |
| yolov5p\_640\_fp32 | 0.9492 | 0.8323 | 0.9195 | 0.6314 |
| yolov5p\_640\_int8 | 0.944 | 0.819 | 0.910 | 0.587 |
| yolov5p\_320\_fp32 | 0.9110 | 0.7808 | 0.8734 | 0.5472 |
| yolov5p\_320\_int8 |  |  |  |  |