Efficient AI: Maximizing Throughput Without Diminishing Accuracy in Real-time In-Situ Computer Vision

Andre Colon, Khushi Pai, Aash Jatin Shah [AAI 595B]

Problem Statement:

How can throughput be maximized without diminishing accuracy in real-time computer vision? This project addresses one of the most pressing challenges in Efficient AI: How to design systems that deliver reliable predictions at high frame rates without sacrificing precision?

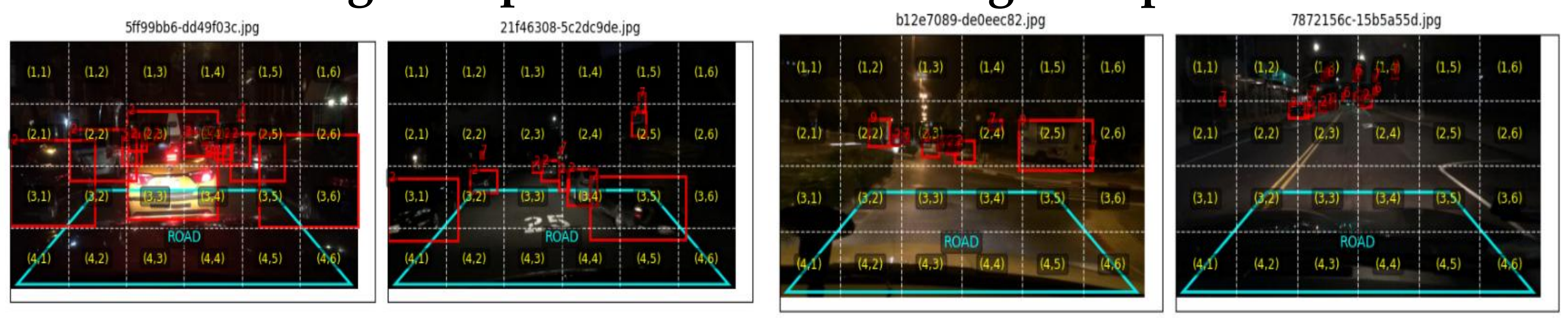
Why is this important?

Even small latency or throughput drops can compromise safety and reliability, leading to hazards, costly errors,or potential injury.

Dataset Generation

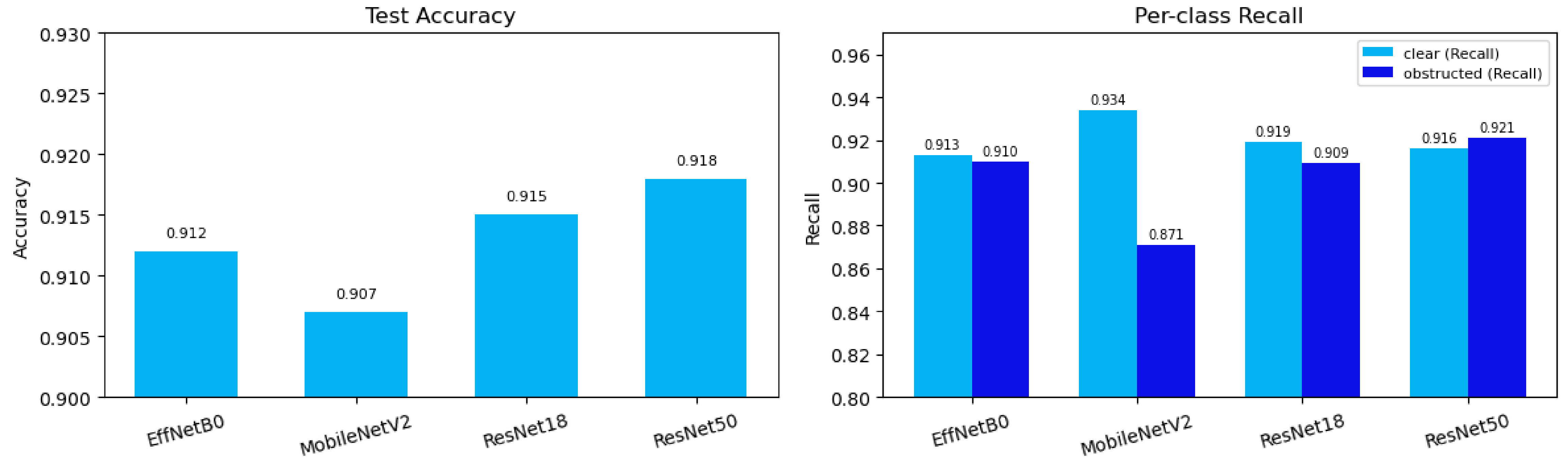
Obstructed image samples:

Clear image samples:



Defined a Road Region on a subset(20K Samples) of the BDD100K Dataset with preset YOLO Labels to create our “clear” and “obstructed” images -> easier for binary classification and training.

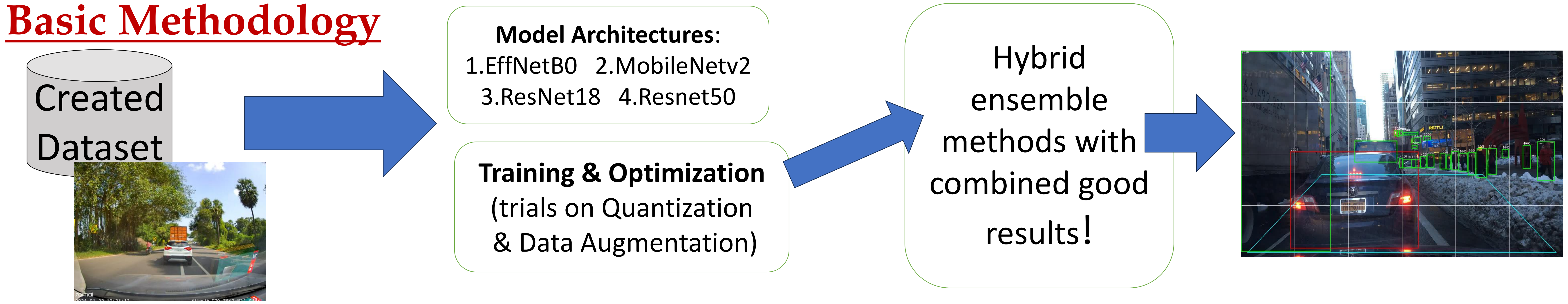
Baseline Model Comparison



Key Findings:

- ❖ EffNetB0 and MobileNetv2 didn’t benefit considerably from optimization techniques.
- ❖ RandAugment showed improved obstructed recall at the expense of overall test accuracy on all models with certain parameters.
- ❖ Ensemble showed improved accuracy (1%) and throughput (120%

Basic Methodology

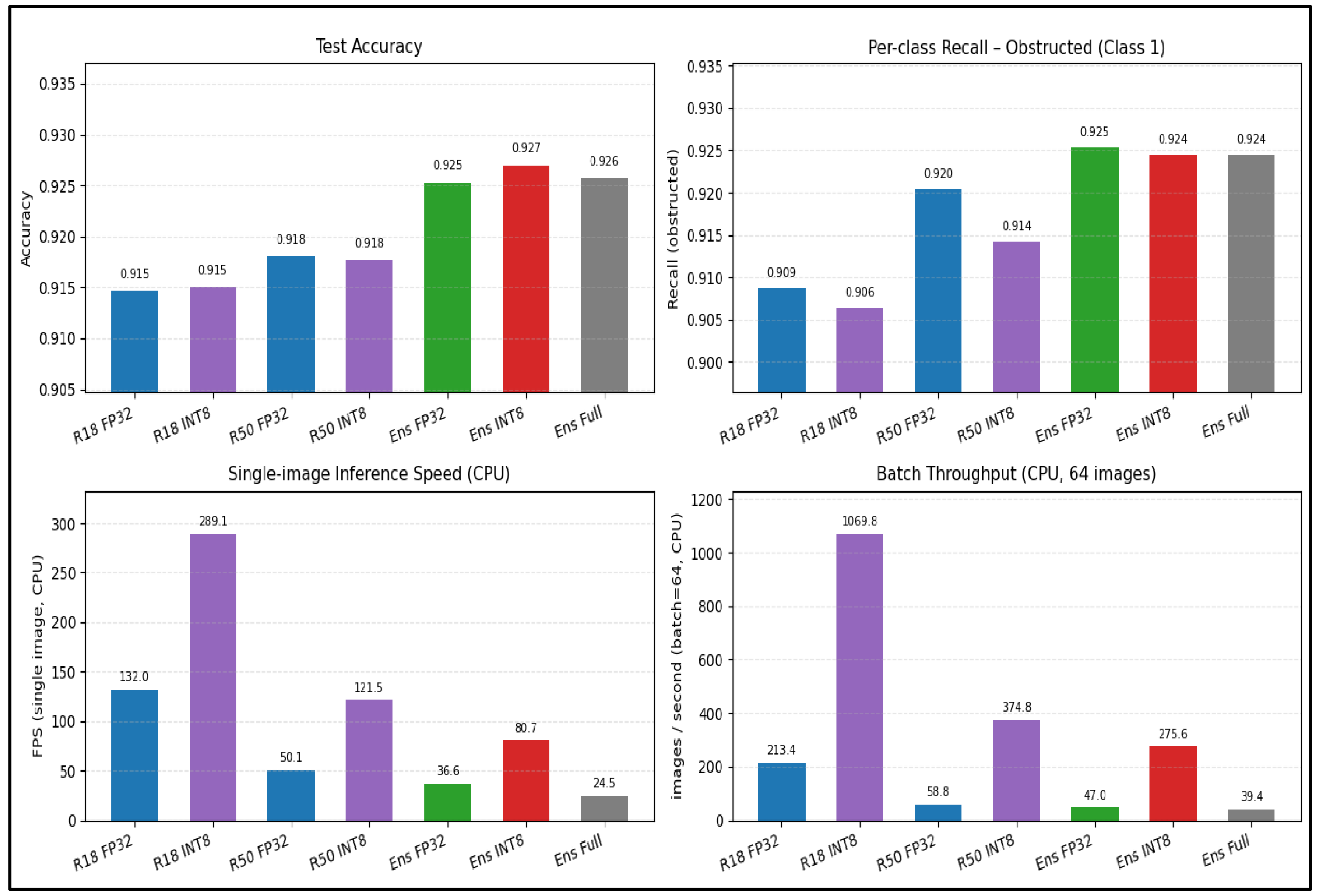


Model Architectures:
1.EffNetB0 2.MobileNetv2
3.ResNet18 4.Resnet50

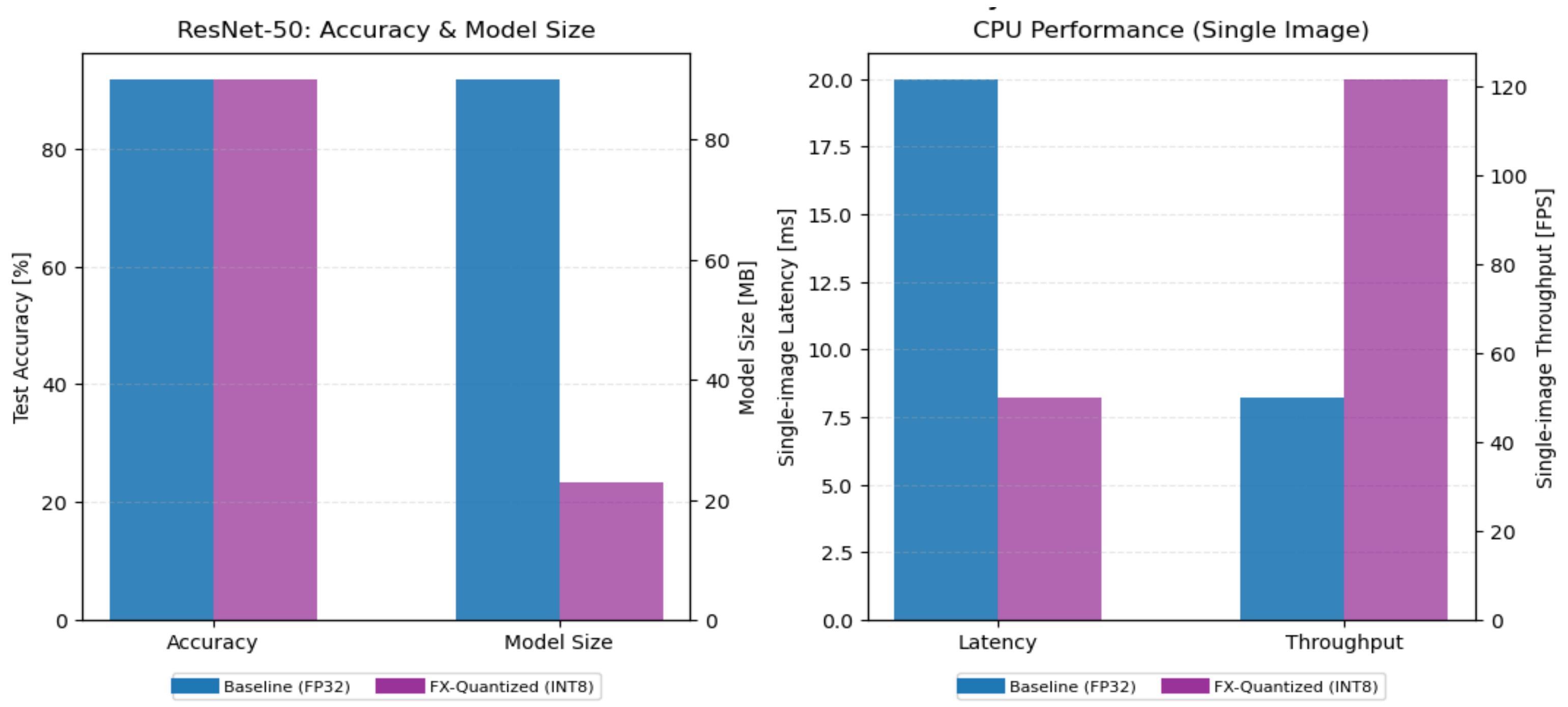
Training & Optimization
(trials on Quantization & Data Augmentation)

Hybrid ensemble methods with combined good results!

ResNet 18/50 – Singles vs Ensembles (CPU)

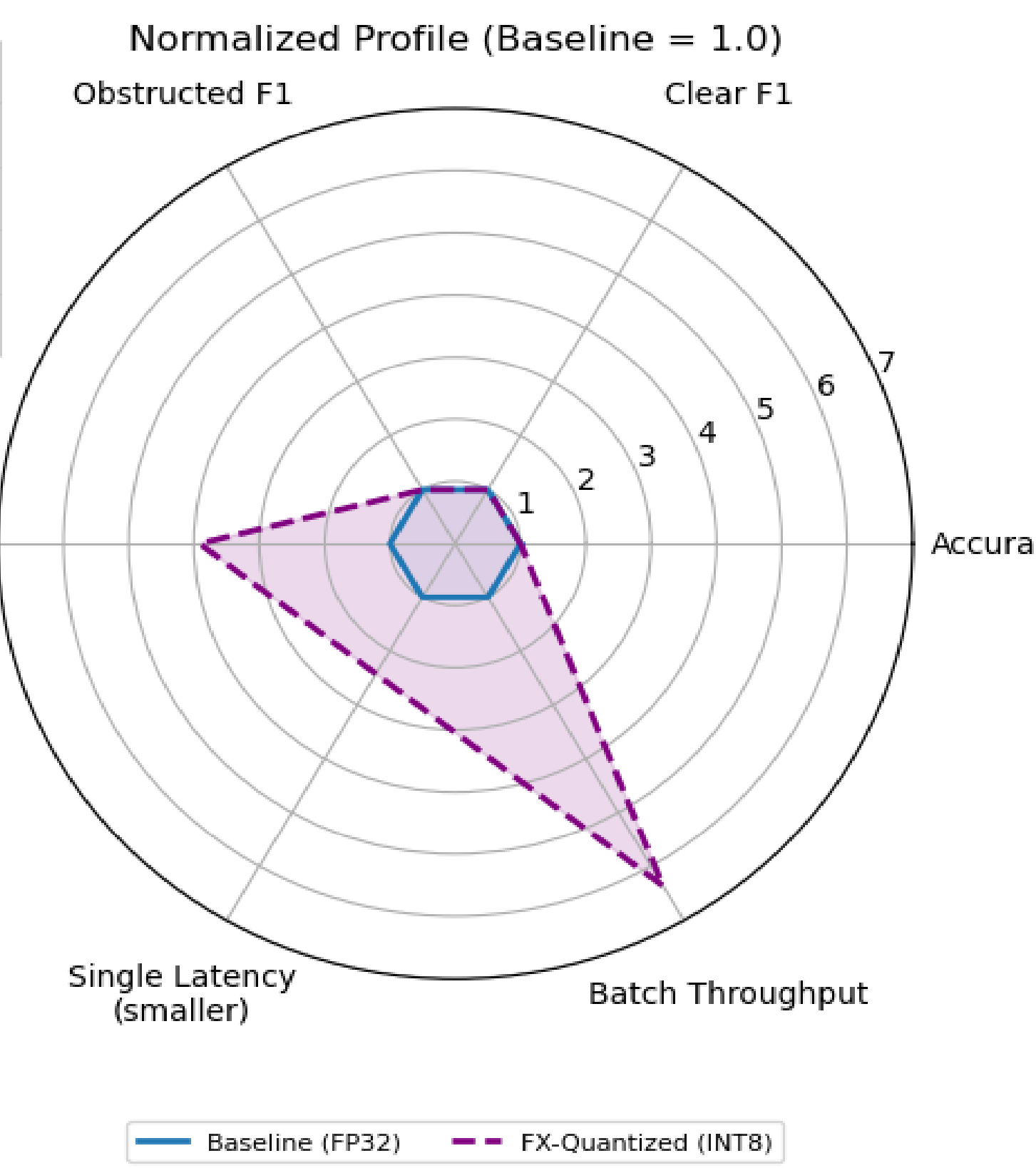


Resnet 50 Quantization Results

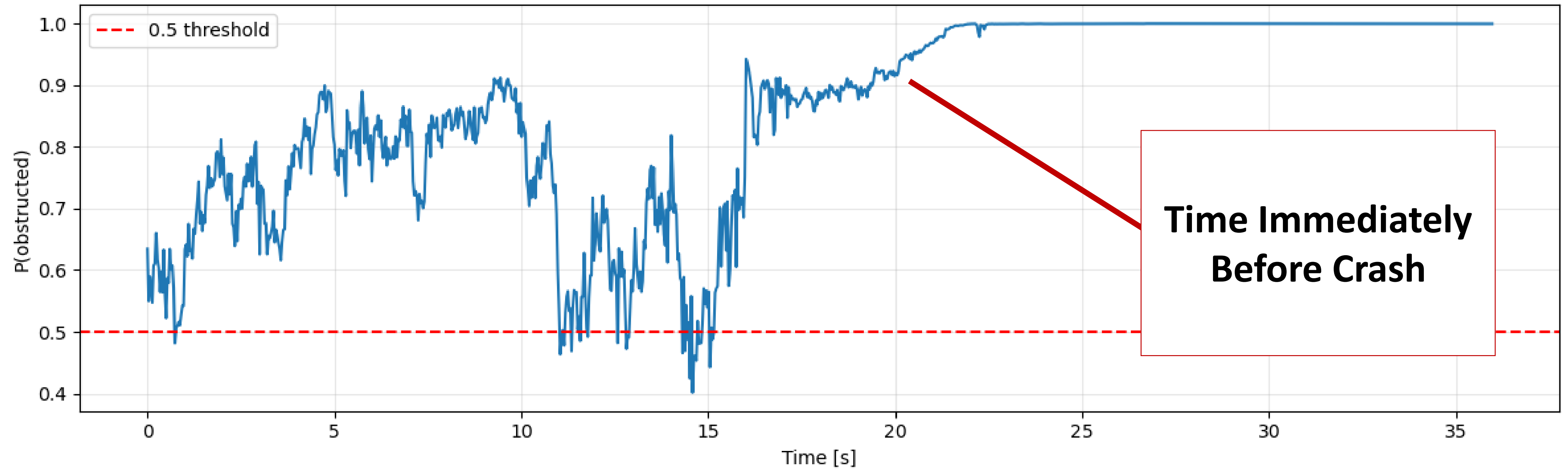


Metric	Baseline	FX-Quantized	Δ% (Quant − Base)
Accuracy	91.80	91.77	-0.04
Model Size [MB]	90.00	22.99	-74.46
Single Latency [ms]	19.97	8.23	-58.80
Batch Throughput [img/s]	58.82	374.78	+537.19

Quantized Resnets (50/18) showed latency boost (59%/ 54%) and batch throughput increase (535% / 400%) without degrading accuracy, yielding CPU deployable version.



ResNet18+50 FP32 Ensemble Real-Time Monitoring: P(obstructed) vs Time



QUANTIZED ENSEMBLE ACHIEVED THE BEST TRADEOFF!

Highest accuracy: 92.70%