Daniel Dubinsky

daniel@dubinsky.dev
dubinsky.dev
linkedin.com/in/daniel-dubinsky

Edge AI Optimization · M.Sc. Computer Science, Tel Aviv University · Matzov Alumnus

When FPS falls short on target hardware, or your quantized model drops too much accuracy, I'm the person who closes that gap - whether the fix is in the model or on the hardware.

12 years spanning the full applied ML stack - from training and fine-tuning models to hit business KPIs, through quantization, to production C++ inference on Hailo, Ambarella, NXP, and ARM platforms.

- PROVEN RESULTS

85 FPS

YOLO26n on Raspberry Pi 5 + Hailo-8L

Ported an unsupported model to edge NPU. With C++ post-processor, achieved 12x speedup over CPU baseline - enabling real-time multi-stream detection on a $100 SBC.

5x Speedup

SRNN Inference Optimization (CUDA)

Optimized a recurrent neural network via custom CUDA kernels. Fused roll-sum-relu into single memory access, eliminated Python overhead. Achieved 5x faster execution while maintaining bit-exact output.

- SERVICES

- **Model Porting & Deployment** - Production-ready inference on target hardware. From training and fine-tuning for your accuracy targets, through quantization, to optimized C++ deployment.
- **Performance Optimization** - More FPS from existing pipelines. Systematic profiling, quantization tuning, and hardware-specific optimization.
- **Architecture & Integration** - Multi-model pipelines that fit your hardware budget. Hybrid CPU/NPU architectures, detection + tracking + classification in a single real-time loop.

- HARDWARE & PLATFORM EXPERTISE

NPU / Accelerators	SoCs	GPU	CPU	Edge Boards
Hailo-8L, Hailo-8	Ambarella CV-series, NXP, Renesas	NVIDIA (TensorRT, CUDA)	ARM Cortex-A, x86 (AVX/VNNI)	RPi 5, Jetson, Custom SBCs

- BACKGROUND

Edge AI Consultant 2026–Present
Specializing in CV model optimization for Hailo, Ambarella, and ARM platforms.

Quris AI - Sr. Data Scientist & Team Lead 2021–2025
Built CV/ML infrastructure for drug discovery. Peer-reviewed research (NViR).

Brodmann17 - Software Engineer 2018–2021
Core developer of proprietary C++ inference engine deployed on Ambarella, NXP, and Renesas platforms for automotive clients.

IDF (Matzov) - Kernel & Embedded Developer 2014–2018
Linux kernel security modules. Real-time C/C++ communication systems.

- ENGAGEMENT MODEL

Engagements start with a **1-2 week diagnostic sprint** - profiling your pipeline, identifying bottlenecks, and delivering a concrete optimization roadmap. From there, typical optimization work runs 4-8 weeks with weekly progress reports. On-site or remote. **Based in Tel Aviv, Israel.** | **Live demo:** dubinsky.dev

Let's talk about your pipeline. daniel@dubinsky.dev