SOMIK DHAR

(+91) 810 769 3095 · somikdhar327@gmail.com · somikdhar327· somikdhar729 · https://dharsomik.tech/

AI / ML Engineer with experience building, optimizing, and deploying ML and computer vision systems, focusing on latency, throughput, and real-time performance.

PROFESSIONAL EXPERIENCE

AI & Perception Engineer **Feb'25 - Present**
NextLeap Aeronautics *Bengaluru, IN*

- Designed and deployed real-time UAV vision pipelines for object detection, tracking, and precision landing, operating under strict latency and compute constraints using C++ and Python.
- Profiled end-to-end perception pipelines to identify CPU/GPU bottlenecks and achieved 30% reduction in perception-to-actuation latency through multi-threading and system-level optimizations.
- Built and maintained real-time inference and streaming pipelines (GStreamer + RTSP/UDP), sustaining 27–28 FPS from a 30 FPS input on resource-constrained hardware.
- Collaborated across perception, control, and systems layers to co-optimize model execution, data flow, and hardware utilization.

RESEARCH EXPERIENCE

Graduate Assistant **Jan'23 - Oct'24**
Ai4CE Lab, New York University *New York, NY*

- Developed deep learning–based visual localization pipelines using CNN and Transformer-based (ViT) encoders, focusing on robustness under viewpoint, illumination, and environmental changes.
- Built end-to-end evaluation pipelines (KD-Trees, sliding-window matching) to benchmark localization accuracy, latency, and robustness under varying conditions
- Improved Recall@5 through feature weighting and temporal aggregation strategies, emphasizing robustness to viewpoint and appearance changes.

Research Intern **Feb'22 - Jun'22**
Indian Institute of Science(IISc.), Bangalore *Bangalore, IN*

- Integrated camera and motion capture data within ROS pipelines for real-time state estimation and control.
- Deployed and validated multi-robot systems under real-time constraints, working with differential and mecanum drive platforms.
- Developed a Python-based CLF motion controller with CBF-based collision avoidance for multi-robot systems
- Achieved a 30cm safety radius in a 6x5 m arena, enabling real-time collision avoidance for multi-robot systems

PROJECTS

GPU Kernel Optimization & Workload Benchmarking

- Implemented and benchmarked performance-critical deep learning primitives (GEMM, reductions, numerically stable softmax) using CUDA.
- Analyzed throughput, latency, and numerical stability trade-offs by comparing custom kernels vs cuBLAS/cuDNN implementations.
- Applied warp-level primitives, shared memory tiling, and block-level reductions to improve kernel efficiency — building strong foundations for hardware-aware model optimization.

TECHNICAL SKILLS

Programming: Python, C++; **ML/ Model Optimization:** PyTorch, CUDA, TensorRT, ONNX; **Edge & Deployment:** Jetson (Xavier/Orin), Raspberry Pi, GStreamer(RTSP/UDP) **Libraries/Tools:** OpenCV, Git

EDUCATION

New York University **Sep'22 - May'24**
MS, Electrical Engineering New York, NY

IIEST, Shibpur **Jul'17 - Jun'21**
B.Tech, Electrical Engineering