ABHISHEK BALASUBRAMANIAM

Austin, TX | abhishekbalu.com
970.227.4428 | balasubramaniamabhishek@gmail.com

Research Scientist specializing in quantized and hardware-aware deep learning for autonomous systems with expertise in model compression (pruning, quantization, knowledge distillation), transformers and LLM optimization, and real-time deployment on embedded platforms (Qualcomm SoCs, NVIDIA Jetson, RTX GPUs). Strong research track record with publications at DATE, DAC, IEEE TCAD, IEEE ICPCSI, and a book chapter on efficient autonomous systems.

EXPERIENCE

SENIOR AI SOLUTIONS ENGINEER
Electronic Arts | Austin, TX
MAY 2026 – PRESENT

- Optimized AI solutions for gaming applications, improving response quality, latency, and inference cost through prompt engineering and model selection.
- Developed RAG pipelines to deliver context-aware responses grounded in game documentation, mechanics, and player support knowledge.
- Built agentic workflows integrating LLMs, APIs, and tools to automate gaming operations and support interactive player experiences.

ADAS/AI ENGINEER
Qualcomm Technologies | San Diego, CA
JULY 2022 – MAY2026

- Quantized and deployed CNNs, Vision Transformers, LLMs (Llama3, Phi-3.5, Qwen) and Stable Diffusion on Qualcomm SoCs.
- Built mixed-precision pipelines (INT8/FP8/FP16) with operator fusion, reducing latency by 40% and energy by 35%.
- Collaborated with hardware and runtime teams to optimize model-to-accelerator mappings and meet ADAS real-time constraints.
- Benchmarked deployments using Qualcomm AI Engine, TensorRT, and ONNX Runtime.
- Integrated object detection, lane segmentation, and driver monitoring models into ADAS pipelines, ensuring compliance with ISO 26262 safety standards.

ADAS/AI INTERN
Qualcomm Technologies | San Diego, CA
MAY 2021 – AUGUST 2021

- Integrated deep learning perception models into embedded ADAS pipelines with focus on quantization-aware optimizations.
- Enhanced model execution speed on Qualcomm heterogeneous hardware by implementing pruning + PTQ strategies.

GRADUATE RESEARCH ASSISTANT
EPIC Lab (Colorado State University) | Fort Collins, CO

SEP 2019 – PRESENT

- Designed a Compression framework with Convolutional Kernel based Optimization for real-time, energy-efficient 3D object detection, achieving 5.6× compression, ~2× faster latency, and ~2× energy savings on Jetson Orin Nano.
- Designed a semi-structured pruning Framework for 2D object detection, enabling sparse execution on embedded hardware with minimal accuracy loss.
- Co-authored an optical PIM accelerator with Optically Programmed Phase-Change Memory (OPCM), delivering 2.98× higher throughput and 137× energy efficiency, with up to 83× lower EPB and 27.5× higher FPS/W vs GPUs and edge accelerators.
- Broader research in pattern-based pruning, mixed-precision quantization, NAS, and sensor-fusion models.

EDUCATION

PH.D. COMPUTER ENGINEERING
Colorado State University - Present

- Researching sensor fusion and 3D perception model optimization for autonomous systems.
- Focus on embedded hardware–software co-design for real-time, energy-efficient deployment.

M.S. COMPUTER ENGINEERING
Colorado State University - 2022

- I specialized in hardware-aware optimization of perception models (quantization, pruning, operator fusion).
- Developed and deployed compressed models on embedded AI platforms (Jetson TX2, Jetson Orin).
- Developed ROS Based Stereo Vision System for Autonomous Robot Navigation.

B. TECH MECHATRONICS
Hindustan University – 2017

- Invited as Track Speaker (Mars Rover Design and Automation) 20th and 21st Mars Society Convention.
- Developed Semi-Automated UAV with position and altitude lock using Raspberry pi and Robot operating System (ROS).

PUBLICATIONS

- A. Balasubramaniam and S. Pasricha, "MAPLE: Modality-Aware Projection-free LiDARCamera Fusion for 3D Vehicular Object Detection," *2026 Design, Automation & Test in Europe Conference (DATE)*, Verona, Italy, 2026
- A. Balasubramaniam, F. Sunny and S. Pasricha. (2025) "UPAQ: A Framework for Real-Time and Energy-Efficient 3D Object Detection in Autonomous Vehicles", 2025 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2025
- F. Sunny, A. Shafiee, A. Balasubramaniam, M. Nikdast and S. Pasricha, "OPIMA: Optical Processing-in-Memory for Convolutional Neural Network Acceleration," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Nov. 2024
- Balasubramaniam, A., Sunny, F. P. Sunny, and Pasricha, S. "R-TOSS: A Framework for Real-Time Object Detection using Semi-Structured Pruning." arXiv preprint arXiv:2303.02191. 2023.
- Balasubramaniam, A., Pasricha, S. (2023). Object Detection in Autonomous Cyber-Physical Vehicle Platforms: Status and Open Challenges, Kukkala, V.K., Pasricha, S. (eds) Machine Learning and Optimization Techniques for Automotive Cyber-Physical Systems. Springer, Cham.
- Abhishek, K. Keshav, S. Gautham, D. V. R. R. Samuel, S. R. Nair, "Low-cost ROS based semi-autonomous drone with position and altitude lock," 2017 IEEE ICPCSI, Chennai, 2017, pp. 2109-2112.
- Abhishek, S. Gautham, D. Varun Rufus Raj Samuel, K. Keshav, U. P. Vignesh and S. R. Nair, "ROS based stereo vision system for autonomous vehicle," 2017 IEEE ICPCSI, Chennai, 2017, pp. 2269-2273.
- Abhishek Balasubramaniam, Shyam Nair, Allen Frederick I. A. "Arachnidan 6 Wheeled All Terrain Explorer Equipped with a 7 DOF Robotic Arm", 20st International Mars Society Convention, Mars Society, USA.

TECHNICAL EXPERTISE

- **Quantization & Compression:** QAT, PTQ, Mixed-Precision, Pattern-Based Pruning, Knowledge Distillation
- **Frameworks & IRs:** PyTorch, TensorFlow, ONNX, TorchScript, TensorRT
- **Hardware Platforms:** Qualcomm SoCs, NVIDIA Jetson Orin/TX2/PX2, RTX GPUs
- **Programming:** Python, C++, CUDA, PyCUDA, Cublas, Cudnn, OpenCV
- **Profiling Tools:** Nsight Systems, perf, Jetson power monitors, hardware debuggers
- **Domains:** 3D Object Detection, Transformers, LLM optimization, Multimodal Perception, Sensor Fusion

---

SELECTED PROJECTS

- **Object Detection Model Compression for Resource constraint platform:** Recreated SSD in TensorFlow/Keras; applied pruning, quantization, and search algorithms, achieving 30% model size reduction with only ~6% mAP drop (0.70 → 0.66).
- **Stereo Vision for Autonomous Navigation:** Built ROS-based stereo vision system with deep learning object detection/tracking; deployed on NVIDIA Jetson TX2.
- **Unmanned aerial vehicle with position and altitude lock with GPS based navigation system:** Developed ROS-based drone with GPS navigation, altitude/position lock using neural networks and fuzzy logic; implemented on Raspberry Pi + Arduino.
- **University Rover Challenge:** Designed navigation, communication, and perception systems for Martian rover prototype. We secured 17th place out of the 36 finalists, and we secured 2nd place out of 8 finalists from India

---

COMMUNITIES

- SIGMA XI HONOUR SOCIETY
- IEEE
- ROS Community
- Pycon India
- Mars society

ACHIEVEMENTS:

- Elected Member, Sigma Xi Scientific Research Honor Society, 2025
- Top 10 best engineering services start-up in India 2018 by Silicon India Magazine
- Invited as Track Speaker (Mars Rover Design and Automation) 20th and 21st Mars Society Convention
- 17th place in University Rover Challenge 2015 at Mars Society, USA
- Finalist in "Robotryst 2013" competition conducted by Robosapiens at IIT Delhi

---