Keshav Sharan Pachipala

pachipala@wisc.edu linkedin.com/in/keshav-s7 https://github.com/Auc7us (608) 896-2239

Technical Skills

Frameworks & Libraries: Pytorch, JAX, TensorRT, TensorFlow, TFLite, ONNX, OpenCV, Transformers, OpenGL, webRTC, WebTransport
Languages & Platforms: C++, Python, JavaScript, CUDA, Docker, Kubernetes, ROS2, GCP, Linux, Jetson, Chrono, IsaacSim, Fusion360
Interests: Real-Time ML Systems, Embedded AI, 3D Perception, Embodied AI, VLMs, LLMs, NLP, Runtime Optimization, GenAI

Work Experience

Wisconsin Autonomous - UW Madison USA
Lead ML Systems Engineer September 2023 - Present

- Achieved 15 fps BEV on **Jetson AGX Orin 64** by parallelizing lane detect + segment and class-partitioned YOLOv11 detectors in **TensorRT** CUDA streams, fusing outputs with **custom CUDA kernels** toward L4 autonomy on the Chevy Bolt.
- Cut image + depth-map transfer **latency ~7 ms→0.06 ms** by replacing ROS with a lock-free C++ **GPU shared memory middleware**, enabling zero-copy access across all perception modules via **CUDA IPC**.
- Deployed a quantized Whisper ASR + Flan-T5 pipeline on an **Amlogic C305X NPU** for edge inference deployment, on-device speech-to-navigation, eliminating cloud latency and supporting **real-time LLM-driven** AV Taxi Mode.

WCB Robotics Inc. India
Software Engineer July 2021 - July 2023

- Led autonomy on E.L.M.O., world's first ropeless facade-cleaning mobile robot, managing a five-member perception and controls team.
- Delivered a **45% improvement** in gravity estimation accuracy resulting in **27% faster** safety trigger response by optimizing sensor fusion on the **Hercules TMS570**, enhancing real-time fall-prevention and trajectory control on vertical surfaces.
- Boosted visual odometry rate 5x using **NVENC + vision accelerators** on Jetson, enabling tracking unlocking SLAM on glass surfaces.

Asteria Aerospace India
Embedded AI Intern January 2021 - June 2021

- Built a real-time human MOT pipeline on **Jetson Xavier NX**, integrating OSD overlays and benchmarking **YOLO, DeepSORT, ResNet** on custom aerial datasets for surveillance UAV deployment.

Research Experience

Simulations Based Engineering Lab (SBEL) - UW Madison USA
Research Assistant (Advisor: Prof. Dan Negrut) July 2024 - Present

- Developed **Decentralised Semantic Collaboration** (DeSC) method for multi-agent search, fusing multiple **BLIP2 Q-former** outputs into shared embeddings for **low-bandwidth** semantic collaboration via novelty-based frontier optimization.
- Enabled **high-fidelity** AV testing on **deformable lunar terrain** by integrating ray-traced Stereo , ToF LiDAR and RADAR sensors into **Chrono physics simulator**, delivering photorealistic data for autonomous-rover validation.

GPTQ++: Compressing Quantized Large Language Models USA
Capstone Research (Advisors: Dr. Min Xu and Dr. Shiliang Hu) May 2024 - Aug 2024

- Introduced an **online weight deduplication** method for GPTQ, using **FAISS** to replace redundant quantized weights and reduce model size under strict memory budgets of edge devices.
- Achieved **4% additional compression** on 4-bit **LLaMA3-8B** with only **6% perplexity loss**, enabling LLM inference on resource-limited edge devices using just **8GB VRAM** in **35 minutes**.

Select Projects

ShapeShifters: 3D Point Cloud Editing with Language

- Trained a multi-modal network using **latent representation learning** (similar to VAE) based on ChangeIt3D + ShapeTalk framework to enable **natural language–driven editing of 3D point clouds**, reducing manual reconstruction effort and accelerating design iteration.

BounceCast: Real-Time Cloud Inference Prototype

- Built a Python + JS web app that streams client simulated 60 fps H.264 frames to the remote server over **WebRTC** and returns commands via bidirectional **WebTransport** with < 17 ms RTT, containerized and orchestrated with **Kubernetes.**

Education

University of Wisconsin-Madison USA
Master of Science, Electrical and Computer Engineering Dec 2025
Research: Machine Learning, Computer Vision and Simulations | GPA: 3.90/4.00

Birla Institute of Technology and Science, Pilani India
Bachelor of Engineering, Electrical and Electronics Engineering June 2021
Minor: Automation and Robotics