Srikanth Malla

Fremont, CA

srikanthmalla.com | linkedin/srikanthmalla | github/srikanthmalla
mallasrikanth004@gmail.com | Google Scholar

Experience

Samsung Semiconductor Feb 2023 – Present
Staff Machine Learning Engineer San Jose, CA

- Leading enterprise AI transformation: architected a centralized agentic AI platform with Model Context Protocol (MCP) tools, automating workflows across internal business units, including AI email and calendar assistants
- Drove AI-powered transformation of logistics operations, replacing manual processes with end-to-end intelligent automation
- Developed multi-objective reinforcement learning (MORL) methods for chiplet design, balancing competing design objectives
- Architected distributed LLM inference kernels on in-memory AI accelerators (hardware-software co-design); research on efficient inference via pruning (COPAL, ICML 2024) and dynamic early-exit for transformers
- Built and deployed an end-to-end RAG system for LLM-powered search over internal corporate knowledge bases

Kinetic Automation Oct 2021 – Feb 2023
Founding Engineer / Staff Research Engineer – Machine Learning Mountain View, CA

- Led ML pipeline end-to-end: simulation, synthetic data generation, 3D perception algorithm design, and production deployment for autonomous EV maintenance robotics
- Developed 3D machine vision algorithms for object detection and scene understanding in autonomous driving applications

Honda Research Institute Jan 2018 – Oct 2021
Research Engineer San Jose, CA

- Published 10+ papers at top venues (CVPR, ICCV, ICRA, CoRL) on multi-object tracking, trajectory forecasting, reinforcement learning (RL) based attention (RAIN), and risk-aware driving; 2 CVPR Oral papers
- Created H3D dataset and DRAMA benchmark – widely adopted in autonomous driving research; led 3D LiDAR mapping and multi-sensor fusion systems
- Filed 7 patents on prediction, localization, and calibration for autonomous systems

Carnegie Mellon University Sep 2015 – Aug 2017
Visiting Scholar / Research Associate, Field Robotics Center

- Deep learning-based ego-motion estimation for UAVs using monocular camera and IMU fusion (with Prof. Katerina Fragkiadaki)
- Built real-time coverage planning and control systems for industrial inspection UAVs (with Prof. Sebastian Scherer)

Education

Worcester Polytechnic Institute – M.Sc. Robotics Engineering (reinforcement learning/RL, motion planning), **GPA: 4.0/4.0** 2017 – 2018
Vellore Institute of Technology – B.Tech. Electronics & Instrumentation 2012 – 2016

Selected Publications

Full list on Google Scholar

- **COPAL**: Continual Pruning in Large Language Generative Models – **ICML 2024**
- **CAMILA**: Context-Aware Masking for Image Editing with Language Alignment – **NeurIPS 2025**
- **DRAMA**: Joint Risk Localization and Captioning in Driving – **WACV 2023**
- **TITAN**: Future Forecast using Action Priors – **CVPR 2020** [**Oral**]
- Shared Cross-Modal Trajectory Prediction for Autonomous Driving – **CVPR 2021** [**Oral**]
- **H3D**: Full-Surround 3D Multi-Object Detection and Tracking – **ICRA 2019**
- 16+ publications at CVPR, ICCV, ICRA, ICML, NeurIPS, CoRL, IROS; 8+ patents

Technical Skills

ML/AI: PyTorch, TensorFlow, Triton, LLMs, LLM inference optimization, pruning, RAG, agentic AI, MCP, multi-objective RL
Perception: LiDAR 3D detection, multi-sensor fusion, multi-object tracking, trajectory prediction, BEV segmentation
Robotics: ROS, MoveIt, Gazebo, MuJoCo, UAV systems, motion planning, SLAM
Languages & Tools: Python, C++, CUDA, Docker, AWS, Git, Linux

Reviewer

ICML '26 | AAAI '25,'26 | CVPR '22–'24 | ICCV '21,'23 | ECCV '22,'24 | TPAMI '24 | ICRA '21,'22,'24 | WACV '23,'26 | SNSF Grant '22