Noopur Koshta

(+1) 774-232-5148 | koshta.noopur@gmail.com | noopur2805.github.io/ | noopur2805 | noopur-koshta

Career Summary

Focused on the intersection of Computer Vision and Machine Learning, with an keen interest in optimizing neural networks and machine learning models on edge devices. My work aims to advance perception algorithms and lightweight model architectures to enable robust visual understanding on resource-constrained systems.

Professional Experience

MER Lab, Worcester Polytechnic Institute | Python, PyTorch, State Estimation *Worcester, MA, USA*

Graduate Research Assistant *Jan. 2024 - Dec. 2024*

- Advanced research in pose estimation and computer vision, yielding optimized neural network architectures with significant drop in convergence rate for real-time inference.
- Tracked manipulator arm pose through occlusions (<0.4 MSE) using Graph Convolution Network (GCN) for skeletal modeling and Unscented Kalman Filter (UKF) for end effector lacking adjacent joint constraints during visual obstruction.
- Achieved 20x faster neural network training convergence through network architecture refinement and custom loss function design.
- Explored warehouse robot navigation via decentralized EKF with Voronoi tessellation for coverage control and collision avoidance, achieving 95% localization precision with 6+ robots operating in spatially-allocated regions.

BioRobotics Lab, Carnegie Mellon University | C++, PID, ROS, RViz *Pittsburgh, PA, USA*

Graduate Research Assistant *Mar. 2024 - Apr. 2024*

- Investigated forward and inverse kinematics of surgical snake robots by analyzing link interdependencies, to enable precise 3D navigation and control during cardiac procedures.
- Developed initial B-spline trajectory planning using RViz simulation, creating obstacle avoidance strategies intended for implementation on robotic platforms.

Ignitarium Technology Solutions Pvt | C++, CUDA *Bangalore, India*

Senior AI Engineer *Nov. 2021 - Jun. 2022*

- Accelerated image processing on resource-constrained SoC devices by developing CUDA kernels that enabled parallel GPU execution of traditionally CPU-based vision functions.
- Reduced system latency in data transformations and vision tasks through optimized matrix multiplication (GEMM), convolution primitives, and pooling operations.
- Led a development team while contributing to create 100+ comprehensive test cases for low-level custom kernels, generalizing testing across diverse image patterns and sizes.
- Maintained and enhanced codebase by creating new abstractions that reduced code duplication and established a generalized framework for image processing kernels.

Persistent Systems Pvt Ltd | Python, Keras, CNN, LSTM, Flask *Pune, India*

Senior Machine Learning Engineer *May 2016 - Apr. 2020*

- Developed ML prototypes across image classification, text analytics, and facial recognition to replace outdated systems, successfully meeting business requirements.
- Engineered hierarchical CNN, for diabetic retinopathy lesion detection, achieving 96.5% detection accuracy across all classes and reduced false positives on IDRiD dataset.
- Automated attendance tracking via facial recognition using Haar-HOG and CNN, boosting accuracy from 85% to 96% when tested across 50+ employees, integrated anti-spoofing detection to eliminate photograph-based attendance fraud.
- Developed LSTM-based multi-stage ICD classification system for medical coding with 85% precision and designed intuitive Flask UI for physician workflow integration.

Skills

Research and Development Stack

Major Languages	Python, Modern C++, MATLAB, CUDA, Triton
Machine Learning	PyTorch, TensorFlow, TensorRT, Optuna, OpenAI Gym
Computer Vision	OpenCV, Eigen, PCL
Robotics Frameworks	ROS/ROS2, RViz
DL/RL Architectures	VLA, Transformers, Diffusion Models, CNN, GCN, GAN, DQN

Other Tools and Skills

Development Tools	Docker, Git, NVIDIA NSight Systems
Physics Simulation	Isaac Sim, MuJoCo, Gazebo, Blender
Cloud Platforms	AWS (Bedrock, SageMaker, Lambda, EC2)
Hardware Platforms	Turtlebot, Nvidia Jetson

Relevant Projects

Pick-and-Place Robot Localization C++, ROS

- Created a complete mobile robot demonstration that successfully navigated between pickup and dropoff locations with sub-meter localization accuracy in cluttered environments.
- Developed real-time robot localization through Adaptive Monte Carlo Localization (AMCL), achieving high-precision pose estimation through probabilistic resampling of particles weighted by sensor measurement likelihood.
- Integrated RTAB-Map SLAM for 3D mapping with loop closure detection, and optimal path planning using A* algorithms to generate collision-free trajectories.

Accelerating Transformers via Flash Attention with Triton Python, Triton, PyTorch

- Implemented Flash Attention, achieving 2.4x computational efficiency benchmarking TorchDynamo backend via GEMM optimization for transformer attention operations.
- Engineered fused kernel operations with SIMD-optimized online softmax and layer normalization, eliminating redundant memory transfers, and reducing numerical instability.
- Incorporated optimized attention mechanisms with quantization-aware training, enabling calibration-free INT8 inference through dynamic quantization of FP16 models.

3D scene reconstruction with Structure from Motion (SfM) Python, PyTorch, OpenCV

- Implemented using scale-invariant SIFT feature extractor with FLANN matching, to establish precise point correspondences across multiple viewpoints.
- Applied the 8-point algorithm with RANSAC for fundamental matrix estimation, recovered camera poses via PnP solving, and performed DLT triangulation with cheirality constraints.
- Implemented bundle adjustment using Levenberg-Marquardt optimization algorithm to jointly refine camera parameters and 3D positions by minimizing global reprojection error.

One-Shot Safety Alignment for Large Language Models (LLMs) via Optimal Dualization [In Progress] Python, Convex Optimization

- Reproducing MoCAN and PeCAN dual optimization algorithms for LLM safety alignment that eliminate costly primal-dual iterations through convex dual function pre-computation.

Exploring Query-Key Relationships in Vision Transformers Through Singular Value Decomposition (SVD) [In Progress] Python

- Reproducing research on transformer interpretability methodology, specifically focusing on the analysis of self-attention singular modes and their decomposition patterns.
- Analyzing complex feature direction interactions within attention layers, ultimately contributing to greater transparency and explainability of large language model (LLM) behaviors.

Multi-Sensor Fusion for Autonomous Driving C++, Python, OpenCV, MATLAB

- Developed sensor fusion pipeline integrating LiDAR, Camera, and Radar with Unscented Kalman Filter (UKF), achieving <0.6 RMSE for vehicle tracking using CTRV models.
- Implemented comprehensive perception pipeline utilizing RANSAC segmentation with Kd-tree clustering for LiDAR point clouds, YOLO v3 detection paired with Shi-Tomasi/BRIEF tracking for camera data, and FFT-based Range-Doppler mapping with CA-CFAR filtering for radar signal processing.
- Validated edge cases including narrow-profile object detection across diverse weather conditions, while optimizing Time-to-Collision to enhance safety-critical system performance.

Honors & Awards

2024 **2nd Winner,** Meta and AWS Hackathon for XR (AR/VR) and GenAI *Palo Alto, CA, USA*

2012 **1st Place,** Secured First Position in Third and Fourth Semesters (BS in Information Technology) *Nagpur, India*

Education

Worcester Polytechnic Institute *Worcester, MA, USA*

M.S. in Robotics Engineering *Aug. 2022 - Dec. 2024*

- Deep Learning-Based Visual Feature Tracking System / Advisor: Prof. Berk Calli
- Multi-Agent Localization / Advisor: Prof. Siavash Farzan
- **Relevant Courses:** Motion Planning, Control Theory, Computer Vision, Reinforcement Learning

Carnegie Mellon University *Pittsburgh, PA, USA*

Visiting Student *Jan. 2024 - May 2024*

- Shape Estimation of Snake Robot / Advisor: Prof. Howie Choset
- **Relevant Courses:** Convex Optimization