Diwakar Ravichandran

Email: diwakarjravi@gmail.com Location: Fremont, CA (Open to Relocation)
LinkedIn: linkedin.com/in/diwakar-ravichandran GitHub: github.com/bundle-adjuster Web: diwakars.pages.dev

Robotics perception engineer (M.S. Robotics, UC Riverside) with nearly 5 years of industry experience across visual SLAM, 3D reconstruction, multi-sensor fusion, and GPU-accelerated inference. I build perception systems end-to-end and write custom CUDA kernels for the optimization and inference layers underneath them.

Skills

Languages: Python, C++, MATLAB
Perception & SLAM: Visual SLAM, Visual-Inertial Odometry, Bundle Adjustment, 3D Reconstruction, Multi-View Stereo, Structure-from-Motion, Feature Detection & Matching, Camera Calibration, Object Detection, BEV / NeRF / Gaussian Splatting
Sensor Fusion & Estimation: Camera–IMU–LiDAR Fusion, Kalman & Extended Kalman Filters, Multi-Sensor State Estimation, GNSS/INS (loosely-coupled EKF), Nonlinear Least Squares (Gauss–Newton, Levenberg–Marquardt)
GPU & Inference: CUDA, TensorRT, DeepStream, NCCL, Thrust, MPI, Nsight Compute / Systems, Quantization, FlashAttention, Multi-GPU Distributed Optimization
Robotics & Frameworks: ROS, COLMAP, Ceres Solver, GTSAM, g2o, CARLA, Gazebo, PyTorch, TensorFlow, OpenCV, Open3D
Edge & Infra: NVIDIA Jetson, Docker, Azure, Git, CMake, FastAPI

Experience

Jio Platforms Ltd. July 2020 – August 2023
Data Scientist – Machine Vision (Robotics Research) Bangalore, India

- Designed and deployed multi-sensor SLAM and state-estimation pipelines fusing camera, IMU, and LiDAR (visual-inertial odometry), improving downstream motion-estimation accuracy by 40%; outputs consumed by motion-planning and control layers.
- Engineered a dense 3D reconstruction engine using Multi-View Stereo, increasing point-cloud density by 85% while preserving structural fidelity across large datasets (continued today as the *recon_engine* project below).
- Trained a Graph Neural Network-based keypoint detection and matching model, surpassing off-the-shelf state-of-the-art by 15% on in-house datasets; integrated learning-based perception with classical model-based estimation in production SLAM/SfM.
- Built an anchor-free person detection and tracking system on Azure: 15 ms inference-latency reduction and 30% accuracy improvement for real-time perception.
- Spearheaded a custom iOS multi-sensor data-collection app (iPhone 12 Pro and later), reducing data-collection cost by over 50% and standardizing dataset consistency; curated in-house SfM / inverse-rendering datasets for a 60% model-quality gain.

Amtrak Tech. Pvt. Ltd. September 2019 – July 2020
Solutions Architect – CUDA Inference Pipelines (NVIDIA Partner) Bangalore, India

- Engineered CUDA-accelerated computer-vision pipelines for real-time inference on NVIDIA GPUs and Arm-based Jetson edge platforms for customer demonstrations and POCs.
- Built DeepStream-based multi-stream perception, scaling inference to 16+ concurrent video streams with efficient GPU utilization.
- Implemented TensorRT quantization and model pruning, reducing end-to-end latency by 30% while validating accuracy through precision-calibration analysis.

NVIDIA Graphics Pvt. Ltd. January 2019 – September 2019
Deep Learning Research Intern Bangalore, India

- Refined CSRNet-based crowd-counting models for dense Indian scenes for the Government of India (via the National Informatics Center), improving accuracy by 20%; accelerated inference throughput by 200% with TensorRT for real-time Jetson deployment.
- Researched Hebbian learning mechanisms integrated into GAN architectures, improving generated-output quality by 7% through controlled experiments; built Dockerized Jetson deployment artifacts.

MathWorks India Pvt. Ltd. June 2018 – August 2018
Technical Support Intern Bangalore, India

- Built a regex-based code search-and-refactor utility to find and replace deprecated patterns across the codebase, reducing large-scale code-modification time by 50%.

Research

SoCAL Lab, UC Riverside October 2024 – June 2025
M.S. Thesis Research; Systems Optimization & Computer Architecture Lab; Advisor: Prof. Daniel Wong

- Authored *Celesta: A Fully Differentiable Optimization Framework* – a GPU-accelerated nonlinear optimization framework for distributed bundle adjustment in visual-SLAM backends.
- Integrated decentralized majorization–minimization (DABA) with Leiden graph partitioning for balanced multi-GPU workloads and improved convergence over the Louvain baseline; validated on BAL "Ladybug" (1,723 cameras, 678K measurements).
- Implemented custom CUDA kernels and Thrust GPU primitives with NCCL + MPI cross-device synchronization; shipped a Dockerized demo bundling five solver binaries: github.com/bundle-adjuster/celesta_demo.

Projects

LLM Inference Kernels (Custom CUDA) May 2026 – Present
CUDA, C++, PyTorch C++ extension, Nsight Compute, vLLM, FlashAttention – github.com/bundle-adjuster/llm-inference-kernels

- Custom-CUDA study of the three kernels that dominate LLM serving cost – fused attention, KV-cache compression (INT4 KIVI), and W4A16 weight-only GEMM – integrated end-to-end on Llama 3.1 8B Instruct (RTX 4090).
- Fused decode attention **1.91**× over PyTorch SDPA (189 GB/s achieved KV bandwidth); INT4 KIVI KV cache 0.27× memory and 1.29× faster; W4A16 GEMM up to **6.97**× over fp16 cuBLAS. End-to-end: **−51% peak VRAM** (18.5→9.05 GB) and **1.55**× decode tok/s, every step attributed to a specific Nsight Compute metric.

Dense Reconstruction Engine (recon_engine) September 2025 – Present

- From-scratch dense 3D reconstruction from image collections – direct continuation of the production reconstruction work at Jio, with the same co-author. Public companion visualizer (Open3D): github.com/bundle-adjuster/point_cloud_visualizer.

GNSS/INS Sensor Fusion (KITTI raw) May 2026

Python, NumPy, SciPy – github.com/bundle-adjuster/gnss-ins-ekf

- 15-state loosely-coupled Extended Kalman Filter fusing IMU with GNSS fixes: position RMS **2.26 m fused vs 1693 m** IMU-only dead-reckoning; demonstrated graceful degradation and recovery across a 30 s GPS-blackout window.

Collaborative Vehicle-to-Vehicle Perception (CARLA) September 2024 – December 2024

Python, PyTorch, CARLA, PointPillars – github.com/REGATTE/Collaborative-Vehicle-2-Vehicle-System

- Multi-agent cooperative perception: early-fused raw LiDAR from four collaborating vehicles into a unified ego frame, then ran PointPillars 3D detection to produce class-labeled BEV detections.

Parallel Bundle Adjustment (CUDA) September 2023 – December 2023

CUDA, C++, CMake – github.com/bundle-adjuster/bundle-adjustment-cuda

- CUDA-parallelized nonlinear least-squares bundle-adjustment solver, 10× over CPU on the Washington BAL dataset (RTX 4090); direct predecessor to the Celesta thesis.

Education

University of California, Riverside – M.S., Robotics (Vision & Perception) September 2023 – June 2025

GPA 3.83/4.00, top-ranked in cohort. Coursework: GPU Architecture & Parallel Processing, Advanced Computer Vision, State & Parameter Estimation Theory, Linear System Theory, Self-Driving Stack.

BMS College of Engineering – B.E., Mechanical Engineering (Top 5%) September 2015 – August 2019

GPA 8.1/10.00 (Scholaro-scaled 3.513/4.00).

Publications & Honors

- **Ravichandran, D.** (2025). *Celesta: A Fully Differentiable Optimization Framework.* M.S. Thesis, UC Riverside. Advisor: Prof. Daniel Wong.
- **Heartathon Hackathon – 2nd of 32 teams** (Heartfulness Innovation Lab, January 2025).