Arun Madhusudhanan

+1 857-381-3654 • madhusudhanan.a@northeastern.edu • LinkedIn • GitHub • Portfolio

EDUCATION

Master of Science in Robotics, Computer Science Concentration Sep 2022 - Dec 2024
Northeastern University, Boston, MA GPA: 4.0/4.0
Courses: Computer Vision, Robotic Sensing and Navigation, Reinforcement Learning, Artificial Intelligence

TECHNICAL SKILLS

- **Programming Languages:** C++, Python
- **Libraries:** PyTorch, OpenCV, TensorRT, ONNX, OpenVINO, vLLM, PCL, Open3D
- **Software Tools:** MATLAB, ROS, Docker, Git, Ubuntu, AWS ECS, FastAPI, OriginPro, SolidWorks

PROJECTS

LoRA Fine-Tuning & Inference of Quantized Qwen2-VL Models [Code] Jun 2025

- Fine-tuned two **4-bit** variants of the Qwen2-VL-2B on a LaTeX OCR dataset: (1) **GPTQ**-quantized model using **LoRA** with Hugging Face **TRL**, and (2) **NF4**-quantized model using **QLoRA** via **Unsloth**.
- Achieved over **2×** improvement in **BLEU** score and **4.5×** reduction in **Levenshtein Edit Distance** via LoRA fine-tuning.
- Enabled high-throughput, low-latency inference using **vLLM** to serve LoRA adapters without merging, benchmarked GPTQ vs QLoRA pipelines on **throughput**, **TTFT**, **TPOT**, and **ITL**, shared insights through detailed **medium articles**.
- Contributed to **open-source**: fixed model layer naming bugs in **vllm-project/llm-compressor** and **huggingface/optimum** and added support for two new datasets for multimodal benchmarking in **vllm-project/vllm**.

Open Vocabulary 3D Scene Understanding [Report] Dec 2024 - Apr 2025

- Designed a **multimodal** 3D scene understanding pipeline enabling robots to interpret complex natural language, image, or audio commands for tasks like grasping and navigation.
- Leveraged vision foundation models (**RAM, Grounding DINO, SAM, CLIP**) for object detection, segmentation, and semantic feature extraction from 2D images, and projected results onto point clouds for **3D semantic segmentation.**
- Incorporated **scene graphs** with object-room hierarchies and used **Qwen2.5-7B-Instruct** for user query understanding, improving object retrieval accuracy by **95%** on the **Replica dataset**.
- Integrated **GraspNet** for **open-vocabulary robot grasping** based on natural language commands, simulated in **PyBullet**.
- Enhanced the grasping pipeline with **Set of Mark (SoM) prompting** and a **Vision Language Model (LLaMA 4 Scout)** for improved reasoning and spatial relationship inference. [Code]

Other Projects

- Achieved **3–7× speedups** for **CLIP, CoCa,** and **YOLOv8** (with NMS TRT plugin) by **TensorRT** conversion. [Code]
- Benchmarked **OpenCV-DNN, ONNX Runtime,** and **OpenVINO** for YOLOv8 inference in C++ on a CPU-only device, selecting OpenVINO with **static INT8 quantization** for a **3×** speedup. [Code]
- Built a **multimodal RAG** pipeline to assist with research paper reading using **LLaMA-3.1-8B** (for text), **LLaMA-4-Scout** (for images), **gte-Qwen2-1.5B** (for embeddings) and **ChromaDB** for retrieval. [Code]
- Built ML Models from scratch: **Tiny-NeRF**, **Stable Diffusion, PointNet**, **Visual Transformer**, **GAN** & **VAE**. [Code]
- Developed a scalable **FastAPI** Image Classification Application using **AWS ECS** and **Docker** containerization. [Code]
- Built a robust **Vehicle Tracking system** by **fusing LiDAR** and **Radar** data with **Unscented Kalman Filter**. [Code]

WORK EXPERIENCE

Graduate Teaching Assistant, Computer Vision Jan 2024 -Apr 2024
Northeastern University, Boston, MA

- Reviewed code, debugged issues, and graded projects in C++, Python, OpenCV, and PyTorch for 120+ students.
- Mentored students on topics such as image processing, object recognition, camera calibration, key points, stereo vision.

Machine Learning Research Intern Jul 2023 - Dec 2023
Festo Corporation, Marlborough, MA

- Developed a robust machine learning pipeline to predict the output parameters of a high-precision liquid dosing unit achieving an error rate of less than **2.5%**; the device with incorporated ML model is currently under **patent application**.
- Handled **end-to-end ML pipeline**: data collection, labeling, feature extraction, model development, and live inference.
- Optimized software modules for sensor activation, hardware drivers, and data storage by refactoring code, fixing bugs, and implementing multi-threading, resulting in a significant reduction in system latency.

Wells Engineer Jul 2018 - Jun 2022
ExxonMobil, Bengaluru, India

- **Led beta testing** of company-wide tubular design software, ensuring robustness and reliability before release to users.
- Stewarded and improved the tubular design workflow for business divisions across the world in accordance with industry standard API 5C5, resulting in **$230k** immediate savings and long-term synergistic benefits.