Howard Wang

Seattle, WA

(+1) 425-351-7712 | howardw1120@gmail.com | howard-wang-hw/

Education

University of California, Santa Cruz — *Santa Cruz, CA*

B.S. in Computer Science — *Oct 2022 - June 2025*

- **Relevant Coursework**: Data Structure and Abstractions, Full Stack Web Dev, Computer Systems, Operating Systems, Algorithm Design, Applied Machine Learning, Computer Architecture, Dynamical Systems, Probability Theory, Algorithm Analysis, Software Eng

Skills

- **Languages:** Python, C/C++, JavaScript, TypeScript, Java, SQL, HTML/CSS
- **Frameworks & Libraries:** React, Node.js, Express.js, Flask, FastAPI, Socket.IO, TypeORM, Material-UI
- **AI & Machine Learning:** PyTorch/PyTorch Lightning, TensorFlow, OpenCV, HuggingFace, Ollama, Computer Vision, Deep Learning
- **Systems & Infrastructure:** Docker, Git, PostgreSQL, Ray, FAISS, Google Cloud, Azure, RESTful APIs, WebSocket, Agile/Scrum

Experience

Kurtz Robotics — *Boston, MA*

Software Eng Lead - Computer Vision — *Jan 2025 - May 2025*

- Led development of AI-powered tomato ripeness detection system achieving 99.24% accuracy using EfficientNetV2-S deep learning model.
- Designed computer vision pipeline with Intel RealSense D456 cameras for real-time RGB-D image processing and 3D object analysis.
- Built cross-platform training system using PyTorch Lightning with GPU acceleration support for CUDA, ROCm, and Apple Silicon.
- Developed lighting-adaptive model simulation to ensure robust performance across varying agricultural field conditions.
- Created Flask REST API for real-time image capture and processing, integrating seamlessly with robotic harvesting systems.

Disaggregated HNSW Lab — *Baskin Engineering, UCSC*

Research Assistant - Dr. Chen Qian — *Oct 2024 - Present*

- Designed Pyramid Search, a hierarchical vector search system using HNSW indexing that achieves 99.79% recall accuracy.
- Built Meta-HNSW routing system for distributing search queries intelligently across partitioned vector databases.
- Developed high-performance C++/FAISS integration with graph partitioning, processing 10,000+ queries per second.
- Optimized memory usage for similarity search operations on large-scale, high-dimensional vector datasets.
- Created comprehensive benchmarking framework using SIFT datasets to evaluate search accuracy versus speed performance.

ELVES Lab — *Baskin Engineering, UCSC*

Ray Distributed Systems Researcher - Dr. Liting Hu — *Jul 2024 - Oct 2024*

- Built fault-tolerant task scheduling system for Ray v2 using C++, reducing recovery time through predictive task replication.
- Optimized distributed memory management by improving reference counting, resolving object storage bottlenecks in large-scale clusters.
- Enhanced task failure recovery mechanisms, significantly improving system reliability for production machine learning workflows.

TouchBase Inc. — *Seattle, WA*

Full Stack Developer Intern — *Jul 2023 - Sep 2023*

- Developed scalable Node.js/Koa REST APIs deployed on Microsoft Azure Functions for serverless, event-driven web applications.
- Optimized MySQL database performance by redesigning schema structure, improving query indexing and data normalization.
- Delivered 15+ responsive front-end features using HTML/CSS in collaboration with design and product teams, handling complex edge cases.

Projects

FastServe - AI Inference System — *Jun 2025 - Present*

- Built FastServe AI inference server implementing Multi-Level Feedback Queue (MLFQ) algorithm with intelligent job profiling and dynamic queue assignment for optimized request scheduling.
- Designed dual-backend architecture supporting HuggingFace Transformers and Ollama models with Apple Silicon (MPS) optimization, memory management, and LRU cache system.
- Developed RESTful API with FastAPI and asynchronous processing, achieving 5.1x improvement in Job Completion Time and 6.4x in tail latency.

Waypoint - Realtime Party Navigation — *Jan 2025 - Apr 2025*

- Developed scalable backend with PostgreSQL and TypeORM, enabling efficient data handling for user and location management.
- Implemented WebSocket connections for real-time location sharing and group coordination, enhancing collaborative navigation.
- Integrated Google Maps API for optimized route planning, turn-by-turn directions, and location-based query performance.
- Containerized application using Docker and added thread-safe operations with async-mutex to prevent concurrent update conflicts.

Diligent - Messenger App — *Jul 2024 - Sep 2024*

- Built a full-stack messaging application using Node.js, Express, React, and PostgreSQL with real-time chat functionality.
- Implemented secure user authentication with JWT tokens and session management, storing user data in optimized JSONB format.
- Designed responsive UI components with Material-UI and achieved 95% test coverage using React Testing Library for component testing.
- Created comprehensive RESTful APIs documented with OpenAPI 3.0 and containerized the entire application using Docker.

Activities & Awards

- Peking University, Certificate of Completion in PKU International Summer Institute (2025) - Research Project under XUANZHE LIU.
- Won the Judge's Choice Award at the 2025 MassRobotics Form & Function Robotics Challenge, Boston Robotics Summit & Expo.
- Awarded the Dean's Award (2022-2025) for Academic Excellence.
- Won the Second Place Winner in Cruzhacks 2023.
- Participant of Amazon's Campus Summer Series.
- Languages: Native Speaker in Chinese (Mandarin/Traditional) and English.