

HAOWEI (HOWARD) CAO

San Jose, CA | (516)778-8868 | c.hw@me.com | [LinkedIn](#) | [GitHub](#) | chwhj.com

Summary

Software Engineer with experience building production systems in Python, C++, Java, and Kotlin across backend services, robotics, and ML infrastructure. Strong in performance optimization, distributed systems, and production observability.

Education

Johns Hopkins University

MSE, Electrical and Computer Engineering

Jan 2024 - May 2025

Baltimore, MD

University of Tennessee

B.S., Computer Science; Minor in Mathematics

Sep 2016 - May 2020

Knoxville, TN

• **Achievements:** Cum Laude; Dean's List (6 Semesters)

Experience

KeploreAI Inc | Software Engineer – AI Systems

Jan 2026 - Present

- Built and deployed containerized AI inference services using Python, Docker, and Kubernetes, exposing model APIs and instrumenting services with Prometheus for production monitoring.
- Profiled and optimized model-serving workflows to reduce inference latency and improve GPU utilization across containerized workloads.
- Developed backend services in Python and FastAPI, implemented REST APIs, and added OpenTelemetry instrumentation to improve production observability and debugging.
- Leveraged AI productivity tools including GitHub Copilot, Claude Code, and OpenAI Codex for codebase navigation, debugging, test generation, and refactoring, validating changes through testing and code review.

Amazon | Software Development Engineer Intern

Jun 2025 - Jan 2026

- Reduced end-to-end inference latency from 300 ms to 50 ms (6×) by optimizing the inference path and deploying TensorRT-accelerated models.
- Integrated OCR into a robotics barcode-recognition pipeline, increasing item-recognition accuracy from 86% to 90%.
- Shipped and tested production-ready perception and infrastructure components in Python and Kotlin, collaborating with robotics teams on integration and reliability.

University of Houston | Research Assistant & Teaching Assistant

Aug 2022 - Dec 2023

- Developed content-based video analysis and indexing algorithms to support efficient exploration of large video datasets and the videopoints.org platform.
- Contributed to the development and evaluation of graph-based algorithms for structural analysis of hex-dominant meshes, including sheet extraction, topology analysis, and visualization for mesh evaluation.
- Led review sessions on data structures and algorithms as a teaching assistant for data-systems courses; nominated for "Best Teaching Assistant."

Shanghai Floauto Valve Co. Ltd | Software Development Engineer

Oct 2020 - Jan 2022

- Refactored a legacy ERP system with a small engineering team, migrating monolithic code into modular C++ services with REST APIs to simplify data access and maintenance.
- Introduced multithreading and performance profiling to data-processing pipelines, reducing batch runtime from several days to a few hours.
- Integrated Elasticsearch for full-text search and operational-log indexing, improving data retrieval and system monitoring.

Selected Projects & Publications

EcoCAR Mobility Challenge – University of Tennessee

Sep 2019 - May 2020

ROS Software Developer, CAV Team · U.S. DOE Advanced Vehicle Technology Competition

- Led the design and implementation of a ROS-based remote visualization and diagnostic platform on NVIDIA DRIVE AGX Xavier, integrating Webviz and Ttyd for real-time vehicle monitoring and remote debugging.
- Integrated ROS data from radar, LiDAR, cameras, and ultrasonic sensors into browser-based diagnostic workflows, enabling engineers to monitor vehicle state, inspect logs and processes, and troubleshoot software during testing.

Hex-Dominant Mesh Structure Analysis – University of Houston

Aug 2023 - Dec 2023

- Analyzed hex-dominant mesh structures using graph-based methods and visualization workflows to compare topology across different mesh-generation approaches.
- Publication: [Hybrid Base Complex: Extract and Visualize Structure of Hex-Dominant Meshes](#), IEEE TVCG, 2024.

Facial Emotion Recognition System – Johns Hopkins University

May 2024 - Jul 2024

- Developed a facial emotion recognition model using U-Net, ResNet, and Transformer architectures with multilevel triplet loss.
- Publication: [Advanced Facial Emotion Classification with 135 Classes for Enhanced Cybersecurity Applications](#), IEEE AI in Cybersecurity, 2025.

Technical Skills

- **Languages:** Python, C++, Java, Kotlin, JavaScript
- **Systems & Infrastructure:** Linux, Docker, Kubernetes, REST APIs, Microservices, Distributed Systems, Parallel Systems, Multithreading, Git
- **ML & Observability:** PyTorch, TensorFlow, TensorRT, OpenCV, Prometheus, OpenTelemetry
- **Data:** SQL, Redis, MongoDB, Elasticsearch
- **AI Productivity Tools:** GitHub Copilot, Claude Code, OpenAI Codex, Cursor, Gemini CLI