Haozhou Xu

hzxuus@gmail.com • www.linkedin.com/in/haozhou-xu-858062288/

EDUCATION

University of California, San Diego San Diego, United States
Master of Science in Computer Science (CGPA: 4.0/4.0) *Sep 2024 – Jun 2026*

University of Hong Kong Hong Kong
Bachelor of Engineering (Major in *Computer Science*, Minor in *Statistics*) *Sep 2020 – Jun 2024*

- First Degree Honor (***Rosita King Ho Scholarship, Dean's Honours List***)

WORK EXPERIENCE

Amazon Austin, United States
Software Engineer Intern *Jun 2025 – Sep 2025*

- Build search **Infrastructure** and **Authorizer** by CDK for Admin operations' Dashboard.
- Created hierarchical **Search Algorithm** at **AWS Lambda** across **40+** workflows within **millisecond.**
- Architected and delivered a full-stack web application from scratch using the **AWS Coral**, **OpenSearch**.
- Reduced Security Team's **5-day** code change pipeline to **minutes** level user friendly operation.

University of Hong Kong Hong Kong
Research Assistant *Oct 2023 – Aug 2024*

- Utilized **OpenCV**, **VLM** tools for street condition checking and point cloud generation.
- Analysed the Hong Kong COVID-19 outbreak period mobility patterns by **Multimodal Models**.
- Developed ML **Localization** system to track human patterns and infection risks with **97%** accuracy.
- Implemented sensor quadratic self-alignment algorithm with **95.8%** accuracy.

University of Hong Kong – CS Department Hong Kong
Research Intern *May 2023 – Aug 2023*

- Explored methods for enhancing 3D perception in various Computer Vision applications.
- Modified **Open3D**, **Pointcept** libraries, integrated **DinoV2** for efficient 2D feature extraction.
- Adjusted **CUDA** code for projection, and achieved at least **3%** improvement in **ScanNet** segmentation.

AsiaInfo Technology Nanjing, China
Software Engineer Intern *May 2022 – Aug 2022*

- Java **API** development for a Fortune 500 company's customer system by **Spring Framework.**
- Involved in **JavaScrip**t writing, module testing, **Oracle DB** handling and customer info analysis.
- Implemented **Distributed** processing and **SQL Querying** to handle over **100/s** parallel requests.

SELECTED PROJECTS

Active RAG Research San Diego, United States

- Created quantitative representation of **LLM**'s Semantic **Uncertainty** for Open Q&A questions.
- Actively balancing **RAG agent** calling and answer **accuracy** based on Uncertainty decomposition.

Concept Learning Research San Diego, United States

- Utilized **Semantic Segmentation**, **Encoding** and Hierarchy **Clustering** for knowledge construction.
- Implemented **VLM** and **CLIP** for caption and **LLM** for data parsing, reasoning with **CBM structure**.

BidFlowAI Startup Project Site San Diego, United States

- Leveraged **RAG system** to transform natural language into standardized procurement documents.
- Implemented data retrieval and auto conversation systems by integrating **Pinecone** and **OpenAI**.
- Efficient Web development by utilizing **Flask**, **Uvicorn**, deployed project on **Heroku** server.
- Conducted module testing and **API** documentation utilizing **Postman** to ensure seamless services.

Personalized Career Consulting System (UG Final Year Project) Project Site Hong Kong

- Collected data by **Selenium** and trained Recommender System for career consulting (**0.82** F1 score).
- Implemented **NodeJS, ReactJS**, and **Flask** for Web design, and integrated into service bus.
- Established remote service on **MS Azure** together with **Azure DB** and **Azure AI** services.

PUBLICATIONS

SimulRAG: Simulator-based RAG for Grounding LLMs in Long-form Scientific QA *2025*
Infrastructural occupancy analytics with bluetooth low energy enabled smart lighting tracking system *2025*

SKILLS

✧ **Software:** Python, Java, Go, C/C++, R, Spring, Maven, Gradle, Netty, RabbitMQ, Apache Kafka, Redis
✧ **Machine Learning:** PyTorch, Tensorflow, OpenCV, Open3D, Keras, MCP, VLLM, RL, Vision, NLP
✧ **Tools:** AWS, MS Azure, MLOps, Kubeflow, MLflow, Flask/Node/React/, MySQL/Oracle/MongoDB, Kubernetes, Docker, Git, Linux, RESTful API, Tomcat, Lambda, Pinecone, OpenSearch, RPC, Jenkins