WUCHEN (AUBREY) LI

New York, NY | (669) 302-7032 | wl758@cornell.edu
GitHub/ LinkedIn/ Website

EDUCATION

Cornell University (Cornell Tech), New York, NY — Aug. 2024 – May 2026
Master of Science in Information Systems | GPA: 3.98
Relevant Coursework: Machine Learning Engineering, Natural Language Processing, Computer Vision, Data Science

ShanghaiTech University, Shanghai, CN — Aug. 2018 – Jun. 2022
Bachelor of Engineering in Computer Science | GPA: 3.50
Relevant Coursework: Software Engineering, Database System, Algorithm & Data Structure, Optimization & Machine Learning

PROFESSIONAL EXPERIENCE

Amazon Web Service (AWS), SDE Intern in AI/ML — May 2025 – Aug. 2025

- Work within Bedrock team to build GenAI infrastructure for training, inference, and deployment of foundation models on AWS.

Florens Asset Management Company, Data Engineer — Sept. 2022 – Jul. 2024

- **Developed Scalable Data Analysis Platform:** Designed and maintained data mid-platform to support business decisions.
  - Build ETL pipelines with HIVE, Greenplum, and PostgreSQL; created **40+** Tableau BI and FineBI dashboards.
- **Developed Asset Selection System:** Engineered a high-performance system to streamline portfolio asset selection.
  - Developed a linear optimization engine using Python, SQL, and CPLEX, supporting selection from **3+ million** assets.
  - Reduced selection time from several days to **5 minutes**, automating **90%** of the selection workflow.
- **Built AI-Powered Automation for Logistics:** Developed CV and LLM solutions to improve operational efficiency.
  - Developed a MaskRCNN-based CV model with **95+** accuracy to detect floor damages in the shipping container return process.
  - Designed an LLM-driven DAG workflow with ChatGPT/Llama and LangChain for automated order booking email processing.

Intel, Software Engineer Intern — Nov. 2021 - Feb. 2022

- **Open-Source Recommender Systems:** Contributed to **DeepRec**, co-developed with Alibaba's AI Recommendation Group.
  - Conducted performance evaluations and testing of BST, DIEN, and DSSM models to identify optimization opportunities.
  - Optimized model efficiency with BF16 precision and self-attention, leveraging Kubernetes for scalable training and testing.
- **Optimized LSTM Model Inference Speed:** Improved the inference performance of Intel's PyTorch LSTM Operator.
  - Optimized LSTM operator inference through profiling, memory alignment, integrating Intel dgemm library (C++/C) , and exp() approximation, achieving **3.5x** speedup. Ensured reproducible benchmarking with Docker.

TECHNICAL SKILLS

- **Coding Language:** Python, C++/C, C#, SQL, Shell Scripting, JavaScript, HTML, CSS, R
- **Tools & Frameworks:** Git, Docker, Linux, PyTorch, TensorFlow, Unity3D, OpenCV, SLURM
- **Professional Tools:** PostgreSQL, HIVE, Greenplum, CPLEX, Pandas, JSON, Figma, Tableau BI, Excel, CI/CD
- **Other Relevant Course:** Software Engineer, HCI, Building Startup Systems, Unity Game Development, Cryptography

PROJECTS

Cornell Tech: MiniTorch Machine Learning Framework Project (NumPy, Numba, Pytest) — Aug. 2024 – Dec. 2024
Course Project: Developed a PyTorch-like ML framework based on Python with auto-differentiation and GPU acceleration.

- Implemented broadcasting, backpropagation, and auto-differentiation for neural network training.
- Integrated GPU acceleration using Numba and operator fusion, achieving **100x** speedup in training and inference.
- Established a Python modular architecture with Pytest unit tests, ensuring reliability and maintainability.

MICCAI: Semi-Supervised Tooth Segmentation (PyTorch, nnUNet, ITK-SNAP) — Jul. 2023 – Sept. 2023
Research Project: Achieved **1st** place in MICCAI 2023 CBCT challenge with a novel two-stage training strategy for 3D tooth segmentation.

- Enhanced nnUNet with additional encoding layers for improved feature extraction and generalization.
- Designed a two-stage training strategy incorporating maxilla-mandible position prediction, data smoothing, and pseudo-labeling.
- Addressed dataset limitations by generating synthetic data to mitigate metal artifacts, improving model robustness.

ShanghaiTech: SPH Fluid Simulation with Temporal Graph Neural Networks (PyTorch, Unity) — Feb. 2022 – Jun. 2022
Research Project: Applied machine learning to physics-based fluid simulation, enhancing SPH simulation speed.

- Developed a temporal graph neural network (GNN) for SPH fluid simulation, modeling particle interactions over time. Integrated Gated Recurrent Units (GRU) and self-attention mechanisms to track fluid dynamics and predict future states.
- Built the simulation framework in Unity (C#) with ComputeShader for real-time rendering and PyTorch for deep learning.