

XIAORUI HUANG



Always Fascinated 

 Preferred Name: Richard
 richardxr.huang@mail.utoronto.ca
 +1 (289) 772-8682  Toronto, Canada
 xiaorui-richard-huang  Xiaorui-Huang

EXPERIENCE

eAI Machine Learning Engineer



Qualcomm

-  May 2023 — Aug 2023  Markham, ON
- Led efforts on **Neural Architecture Search (NAS)** and model compression within the **Edge AI R&D** team.
 - Developed a NAS framework, leveraging Qualcomm's patented NAS techniques, to optimize **arbitrary models** for **any profiled hardware**, harnessing Pytorch's **torch.fx** extensively.
 - Streamlined the NAS workflow for incoming client models, slashing **engineering time** by **80%**.
 - Achieved a **50% reduction** in **model size** and a **60% drop in inference latency** without compromising accuracy across benchmark models.
 - Engaged in lab meetings focused on cutting-edge model compression research, particularly **Quantization**.
 - Delivered a comprehensive presentation on the NAS framework to the broader eAI team.

NAS Quantization Pytorch torch.fx ONNX

RPA Backend Developer

IBM


-  May 2022 — Apr 2023  Markham, ON
- Worked on backend development for IBM's Robotics Process Automation (RPA) platform, written in **C#**.
 - Augmented IBM RPA's **WAL** programming language, introducing a reflection feature resembling Java and C#.
 - Collaborated with cross-functional teams, achieving a **15%** reduction in customer issues and defects per release.
 - Employed **agile methodologies**, showed both independent and collaborative competencies in a hybrid environment.
 - Articulated and presented solution strategies to RPA's senior architects and product teams.

C# Programming Language Design Agile

EDUCATION

University of Toronto

Candidate for HBSc. in Computer Science



-  Sep 2019 — Expected Jun 2024
- CSC367 **Parallel Computing** (In Progress) Parallel Arch & Algo, threading & OpenMP, Distributed Computing w/ MPI, **CUDA Arch & Reduction Algo**, Cloud Computing
 - CSC413 **Deep Learning** (96%) — **Transformers**, CNN, RNN, GAN, VAE, GNN, RL, Model Tuning techniques
 - ECE568 **Computer Security** (83%) — Buffer Overflow & Control Hijacking, **Cache Side-Channel** Attacks, Network Security, Cryptography, Web Security
 - CSC317 **Computer Graphics** (97%) — Ray Tracing, Mass Spring Systems, BVH, Meshes, Kinematics, OpenGL Shaders in **C++** using **Eigen** and **libigl**

CSC369 OS CSC401 NLP CSC412 Probabilistic ML



RESEARCH

Machine Learning Research Intern

embARC Research Group

-  Jan 2024 — Now  University of Toronto
- Research on **real-time Gaussian Splatting & NeRF** 3D reconstruction with data captured on embedded devices.
 - Provides incremental Point Cloud initialization and dataset sampling techniques to improve real-time reconstruction performance.
 - Supervised by **Prof. Nandita Vijaykumar**
- 3D Gaussian Splatting SLAM NeRF Pytorch



Linearly Explored Learning Rate Scheduler

-  Apr 2022  RolandGao/pycls
- We introduced the LES method to automate and refine the resource-intensive task of **learning rate tuning**.
 - LES achieves a final error rate of 8% on par with other commonly used optimizer and schedulers on pycls code base **without the need for learning rate tuning**.
 - Developed a custom **SGD with momentum** algorithm to facilitate exploration of various backpropagation strategies during LES creation.

PROJECTS

CUDA Ray Tracing



Almost Real Time Ray Tracing

-  Nov 2023  Xiaorui-Huang/cuda-ray-tracing
- Implemented a **CUDA** ray tracer with **BVH** acceleration structure, with **Blinn-Phong** shading.
 - Achieved **real-time** ray-tracing of **30 FPS** and **2000x Speedup** on RTX3060-Ti compared to CPU.
 - Incorporated dynamically loaded Scene generation to allow for future interactivity.

CUDA C/C++ Computer Graphics

Woodoku Learn












Reinforcement Learning Model

-  Jul 2022  EdwardHaoranLee/WoodokuLearn
- Replicated the mobile game Woodoku for the terminal using Python, enabling both human and AI gameplay through dedicated environment APIs.
 - Employed Q-Learning, a **Reinforcement Learning** approach with Pytorch, targeting top scores on the Woodoku leaderboard.

RL Pytorch OOP Agile CMake

SKILLS

Programming Languages

 Python  C/C++  CUDA  C#  Java  Rust
 LaTeX  R  TypeScript  HTML&CSS  SQL

Skills, Frameworks & Development Environments

3D Reconstruction Model Compression Parallel Algorithms
Pytorch Vim Docker WSL git VSCode