Martin Nguyen

(+1) 872-294-1416 | manhntm3@gmail.com | Linkedin: manhntm3 | Github: manhntm3 | W Flournoy St, Chicago, IL

A dedicated and results-driven machine learning engineer with problem-solving skills in algorithm and data structure. Knowledgeable about machine learning research areas. Experience in low-level optimizing of GPU model inference.

Education

The University of Illinois at Chicago

Master of Science in Computer Science, GPA: 4.0/4.0 — *Aug. 2024 – May. 2026*

Relevant coursework: Adv. Linux Kernel Programming, Cloud Computing, Distributed Systems

FPT University

Bachelor of Computer Science, Full-ride scholarship — *Aug. 2017 – Aug 2021*

Professional Experience

Founding Engineer (Machine Learning) — Jul. 2021 – Aug. 2024

Vizgard Ltd — *London, United Kingdom*

- Contributing to the company's growth from early prototype to securing over $2.5M in venture funding.
- Led the design and development of a real-time computer vision system for surveillance and unmanned system automation. Architected an event-driven, multi-queue data pipeline to optimize latency and minimize memory locality issues, enabling seamless integration of AI modules. Implemented in C++, the system supported up to 6 simultaneous camera streams on NVIDIA Jetson Orin and 30 streams on 4 NVIDIA Ada GPUs.
- Implemented, evaluated and optimized data pipeline for deep learning model including object tracking (SiameseRPN++ and DeepSORT), object detection(YOLO), pose estimation(AlphaPose) and face redaction. Trained models using PyTorch/TensorFlow on GCP; deployed with TensorRT for high-speed inference.
- Developed a low-latency WebRTC streaming server using GStreamer and Node.js to deliver real-time HD video to browsers and RTSP endpoints.
- Ported pre-processing, inference, and post-processing from CPU to CUDA kernels for various models, including transformer-based architectures with heatmap outputs, resulting in a 3× performance gain with no accuracy loss.

Software Research Engineer — Jan. 2020 – Jul. 2021

Qualcomm Research (previously VinAI Research)

- Built a face recognition SDK ran offline in mobile phones, which then scaled to a face check attendance system (included an annotation application on Android used to collect face data and send to AWS and Google FireBase to form the dataset and calculate the analytics)
- Designed a two-stage anti-spoofing module with infrared camera input which achieved $< 3\%$ FAR
- Improved existing face recognition system with knowledge distillation and network optimization. The final released model was 4 times faster with equivalent accuracy compared to the old model

Projects

Dynamic eBPF Firewall for Domain Filtering and DDoS Mitigation *eBPF, Rust, Linux Network*

Developed a Linux kernel module as a dynamic eBPF-based network firewall in Rust to block traffic and mitigate DDoS attacks via real-time IP filtering. **Github:** https://github.com/manhntm3/cs594-sp25-ebpf

Conversational Agent using AWS Bedrock *EC2, Scala, Go, gRPC, Python, AWS Lambda*

Developed a distributed LLM training pipeline using Apache Spark and implemented both RESTful and gRPC servers for cloud integration. Deployed services on AWS EC2 to route requests to AWS Lambda and Bedrock, powering a conversational agent built on the open-source LLaMA model.
Github: https://github.com/manhntm3/ConversationAgent

Reimplement FaceShifter *Python, Pytorch, Docker*

Reimplemented state-of-the-art computer vision framework FaceShifter, that performs face swapping in Pytorch. **Github:** https://github.com/manhntm3/FaceShifter-Pytorch

Skills Summary

Programming Languages:	C/C++, Python, Scala, CUDA, Java, JavaScript, Go
Databases & Bigdata:	Spark, Hadoop, MySQL, MongoDB
DevOps:	Docker, Kubernetes, Jenkins, Azure, AWS, GCP
Machine Learning:	CV systems, LLMs, Generative AI, PyTorch, TensorFlow
Web Development:	TypeScript, WebRTC, Node.js, HTML/CSS, Flask, Django
Soft Skills:	Problem Solving, Self-learning, Leadership