Yupeng Han

hanyupeng9406@gmail.com | +1 (765) 337-0063 | yupenghan.github.io

Summary

GPU performance engineer; strong in **CUDA optimization patterns**, kernel optimization, and **heterogeneous (CPU/GPU) computing**
Proven speedups: minutes to **50ms**, **20×** (170ms to 8ms), **10×** (150ms to 15ms) via kernel design and memory optimization
Strong in **Nsight Systems/Compute** profiling and system-level debugging across CPU/GPU pipelines

Hands-on LLM inference systems work spanning transformer serving bottlenecks and datacenter-scale runtime trade-offs, including continuous batching, chunked prefill, tensor parallelism, KV-cache behavior, roofline-guided analysis, GPU profiling, and distributed communication; implementation studies include `llama2.cpp`. GitHub: https://github.com/YupengHan/llm-scaling-notes

Professional Experience

PlusAI Inc. — **Jan 2024 – Present**

Staff Software Engineer – Perception System & Compute Efficiency

- **Architected the transition from open-loop to closed-loop simulation** for the secondary perception stack and established automated regression pipelines to quantify bottlenecks and validate latency-sensitive behavior under edge-compute constraints
- **Designed custom CUDA kernels** for fisheye camera stitching and implemented adaptive sensor-selection logic to enable new hardware integration, improving long-range traffic-light precision/recall by **5%** while balancing GPU compute load
- **Led release-candidate triage** across Perception, Prediction, Planning, and Control, isolating runtime regressions through replay, logging, and GDB-level debugging to maintain on-vehicle stability

EBots Inc. — **May 2022 – Jan 2024**

Senior GPU Engineer – High-Performance CUDA Kernel Design

- **Spearheaded GPU-based dense object retrieval**, cutting latency from minutes to **50ms** via custom CUDA kernels featuring **global memory coalescing and shared memory tiling**
- **Achieved 20× speedup** (170ms to 8ms) for large-scale 3D reconstruction by re-architecting compute patterns to **reduce warp divergence and improve SM occupancy**
- **Implemented a GPU-resident KD-tree** enabling 10× faster iterative nearest-neighbor searches (150ms to 15ms), minimizing host-device transfers and **reducing PCIe overhead**
- **Conducted deep-dive profiling** with Nsight Compute, identifying memory-bound kernels and optimizing instruction-level parallelism

Trifo Inc. — **Jun 2021 – May 2022**

R&D Engineer – Optimize SLAM & Local Feature Generation

- Developed a feature-voting strategy to filter noisy scans, improving mapping stability and real-world robustness

Carnegie Mellon University, Robotics Institute — **Oct 2019 – Jun 2021**

Research Engineer – Real-Time GPU Kernel Development

- Built real-time GPU object-detection pipelines, optimizing throughput and latency for deployment constraints
- Developed pose proposal generation for RGB-D 6-DOF pose estimation and accelerated matrix/tensor computations with **CUDA**; contributed to a system published at *IROS 2021*

Deptrum Co., Ltd — **Apr 2019 – Aug 2019**

Computer Vision Engineer

- Developed a high-precision face-detection pipeline on depth images and optimized CPU-side inference throughput

Education

Purdue University — **West Lafayette, IN**

M.S. in Engineering, GPA: 3.96/4.0 — *2017 – 2018*

Shanghai Jiao Tong University — **Shanghai, China**

B.S. in Engineering (Tsien-Hsue-Shen Honor Program), GPA: 3.75/4.0 — *2013 – 2017*

Publications

A. Agrawal, **Y. Han**, M. Likhachev, “PERCH 2.0: Fast and Accurate GPU-based Perception via Search for Object Pose Estimation,” *IROS 2021*
J. Thekinen, **Y. Han**, J. Panchal, “Designing Market Thickness and Optimal Frequency of Multi-Period Stable Matching in CBDM,” *ASME IDETC 2018*

Skills

Languages & Tools: C++, CUDA, Python, Nsight Systems/Compute, GDB, Git, Linux
Expertise: GPU Architecture (SIMT/Warp), Memory Optimization (Shared/Coalescing), Parallel Patterns, Heterogeneous Profiling

Honors

Dean’s List and Semester Honors (Purdue) · Outstanding Individual (SJTU, 2016) · First Prize, National Mathematical Olympiad (2013)