Yupeng Han

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

Summary

GPU performance engineer specializing in CUDA kernel optimization, profiling-driven performance engineering, reproducible benchmarking harnesses, and heterogeneous CPU/GPU systems. Proven speedups from minutes to 50 ms, 20× (170 ms to 8 ms), and 10× (150 ms to 15 ms).

Systems-oriented LLM inference and serving work focused on transformer bottlenecks, KV-cache behavior, continuous batching, chunked prefill, tensor parallelism, and roofline-guided performance analysis, including implementation-oriented studies of `llama2.cpp` and vLLM-style serving. GitHub: llm-scaling-notes.

Selected Project

CUDA Kernel Optimization Harness GitHub

- Engineered a profiling-driven, human-in-the-loop CUDA matmul optimization harness for a fixed BF16 GEMM on an RTX 3070 Laptop GPU, with correctness-gated benchmarking and structured iteration.
- Reduced a shape-specialized custom kernel from 802.8 ms to 24.2 ms; outperformed the local CUTLASS baseline (25.9 ms) by about 7% and reached 91.1% of cuBLAS performance.

Professional Experience

PlusAI Inc. *Staff Software Engineer – Perception System & Compute Efficiency* **Jan 2024 – Present**

- 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. *Senior GPU Engineer – High-Performance CUDA Kernel Design* **May 2022 – Jan 2024**

- Spearheaded GPU-based dense object retrieval, cutting latency from minutes to 50 ms via custom CUDA kernels featuring global memory coalescing and shared-memory tiling.
- Achieved a 20× speedup (170 ms to 8 ms) 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 (150 ms to 15 ms), 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. *R&D Engineer – Optimize SLAM & Local Feature Generation* **Jun 2021 – May 2022**

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

Carnegie Mellon University, Robotics Institute *Research Engineer – Real-Time GPU Kernel Development* **Oct 2019 – Jun 2021**

- 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 *Computer Vision Engineer* **Apr 2019 – Aug 2019**

- 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 Memory/Coalescing), Parallel Patterns, Heterogeneous Profiling