XI HAN

39 Montclair Dr, Selden, New York 11784, United States
(+1) 631-710-8313 | xihan1@cs.stonybrook.edu | https://axihixa.github.io/

EDUCATION

Department of Computer Science, Stony Brook University, New York, United States Aug 2019 – Present
Ph.D. in Computer Science (In progress, expected by Spring 2026) | GPA: 3.9/4.0

Department of Computer Science and Technology, Tsinghua University, Beijing, China Aug 2015 – Jul 2019
B.E. in Computer Science and Technology | GPA: 3.25/4.0

SKILLS

- **TL;DR.** C/C++/CUDA/Python; GPU Algorithms; Mamba; PDEs; Computer Graphics; Machine Learning.
- Expert in GPU algorithms for AI/HPC: Customized AI operators, AI model training/inference efficiency optimization. Involved techniques: PyTorch C++/CUDA extensions, GPU kernel profiling, fine-tuning, operator fusing, cache optimization, etc.
- Expert in Computer Graphics and Numerical Analysis: Neural PDE solvers, customized CUDA operators.
- Expert in languages: C/C++ (OOP, STL, Metaprogramming and Concurrency), CMake, CUDA (including PTX) and Python.
- Expert in tools: PyTorch Profiler, CUDA-GDB, Nsight Compute and NVIDIA Compute Sanitizer.
- Expert in frameworks: PyTorch and OpenGL.
- Familiar with: Linux systems and the Qt framework.
- Knows: Bash, Assembly, MATLAB, Java, Objective C/C++ and Swift.

PUBLICATIONS

- Jingwei Zhang*, Anh Tien Nguyen*, **Xi Han***, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S. Hosseini, “2DMamba: Efficient State Space Model for Image Representation with Applications on Giga-Pixel Whole Slide Image Classification”, In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. (*: **Equal Contribution**)
- **Xi Han**, Fei Hou and Hong Qin, “UGrid: An Efficient-And-Rigorous Neural Multigrid Solver for Linear PDEs”, In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.
- Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, **Xi Han**, Dingcheng Yang, Hao-Zhi Huang and Shi-Min Hu, “Pose2Seg: Detection Free Human Instance Segmentation”, In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

WORK EXPERIENCE

Research Assistant & Teaching Assistant Aug 2019 – Present
Stony Brook University, New York, United States

- Conducted research projects on GPU algorithms, neural PDE solvers, PDE-driven foundation models (e.g., Mamba), and Computer Graphics. Research work is published in top-venue conferences, including ICML and CVPR.
- Extensive project experience on AI training/inference efficiency optimization. Involved techniques: CUDA kernel fusing, performance profiling, and PyTorch CUDA extensions.
- Hosted lectures on OpenGL programming with C++/Python, the implementation details of computer graphics applications and algorithms, and the state-of-the-art research topics on graphics and physics-based modeling.

Research Assistant Jan 2017 – Jul 2019
Tsinghua University, Beijing, China

- Deployed a MobileNet module on IOS platform with Apple’s CoreML framework, and delivered an IOS app for a human segmentation (in Swift and Objective C++).
- Optimized the model used in the app (increased accuracy and added key point recognition) and achieved 10x speedup in FPS.

SELECTED PROJECTS

2DMamba: A Hardware-Aware 2D Selective State-Space Model with Applications in Image Processing

- **TL;DR**. 2DMamba employs a geometric-rigorous and hardware-aware 2D SSM formulation, which extends 1D Mamba into 2D while maintaining its modeling capabilities, high parallelism, and memory efficiency.
- Implements a warp-scan based 2D parallel scan routine which supports scanning prefix callbacks for global tiling. Extends 1D Mamba scanning operation into 2D while maintaining its training/inference efficiency. Compared to a naïve implementation, achieves a throughput of 10x, while the GPU memory consumption is only 10%.
- Involved techniques: Warp shuffle and parallel scans, 2D tiling and caching, HBM access optimization, CUDA kernel and AI model profiling, and PyTorch CUDA extension encapsulation.
- Code available at https://github.com/AtlasAnalyticsLab/2DMamba.

UGrid: An Efficient-And-Rigorous Neural Multigrid Solver for Linear PDEs

- **TL;DR.** UGrid is a neural solver for Partial Differential Equations (PDEs) with convergence guarantee.
- Built upon the combination of the U-Net architecture and the legacy MultiGrid PDE solver, provides users with high speed (up to 20x speedup against legacy solvers), high precision (relative residual as low as 1e-5), high robustness (against irregular and noisy input), high generalization power (to irregular boundary geometries and topology), and high scalability (without need for retraining).
- Involved techniques: Numerical analysis on convergence, customized AI operators (Python and CUDA based). Implements a customized CUDA convolution module to save computation for specific-shaped convolution kernels used in PDE solvers.
- Code available at https://github.com/AXIHIXA/UGrid.

CUDA Baseline Experiments: Performance Profiling and Optimization

- **TL;DR.** Implements and fine-tunes multiple CUDA baseline algorithms.
- Implemented baselines and their optimizations:
  - □ Parallel reduction (with loop unrolling and warp shuffle primitives).
  - □ Histogram and Copy-If (with atomic primitives).
  - □ Parallel scan (WarpScan and Raking variants).
  - □ Fused Biased-Mask-Scale-Add (fp32 and fp16, for fp16, with half precision primitives like __hadd2).
  - □ SGEMM an GEMV (loop unrolling, SMEM padding, warp tiling and double buffer optimizations, reaching 90% throughput of cuBLAS).
  - □ Dropout (with cuRAND APIs).
  - □ Fused SoftMax/LayerNorm/RMSNorm, Im2Col, Matrix transpose, etc.
- Code available at https://github.com/AXIHIXA/CudaDemo.

MISC

- Language Proficiencies:
  - □ Chinese (Mandarin): Native speaker.
  - □ English: Fluent, near-native proficiency. TOEFL: 106/120; GRE: 324/340 + Writing 3.5.
  - □ Japanese: Sufficient for basic working scenarios. JLPT: N1 173/180, N2 169/180.