Chenxia Han

+ (86) 176-1131-8773 | cxhan@link.cuhk.edu.hk | cxhan.com

Education

The Chinese University of Hong Kong — Hong Kong
M.Phil. in Computer Science and Engineering — *2020 - Nov. 2024*

Wuhan University — Wuhan
B.Eng. in Computer Science and Technology — *2014 - 2018*

Research

My previous work has primarily focused on three key areas: (1) developing efficient training frameworks, (2) accelerating model inference for video processing, and (3) optimizing GPU resource management.

Selected Publications

Scalable Complex Event Processing on Video Streams
Chenxia Han, Chaokun Chang, Srijan Srivastava, et al. — *SIGMOD 2025*

SGDRC — Software-Defined Dynamic Resource Control for Concurrent DNN Inference on NVIDIA GPUs
Yongkang Zhang, Haoxuan Yu, **Chenxia Han**, et al. — *PPoPP 2025*

Top-K Deep Video Analytics: A Probabilistic Approach
Ziliang Lai, **Chenxia Han**, Chris Liu, et al. — *SIGMOD 2021*

SimpleDet: A Simple and Versatile Distributed Framework for Object Detection and Instance Recognition
Yuntao Chen, **Chenxia Han**, Yanghao Li, et al. — *JMLR 2019*

Working Experiences

Research Assistant — Hong Kong
The Chinese University of Hong Kong — *Aug. 2019 - Nov. 2024*

- Developed Bobsled, an efficient system for video stream pattern matching. The solution combines algorithmic innovations (a draft model to reduce target model inference) with system optimizations (code generation, caching, and batching) to minimize overhead. Achieved speedups of 2.4× to 11.6× without accuracy degradation.
- Created SGDRC, a dynamic GPU resource controller that optimizes VRAM allocation by reverse-engineering black-box channel mappings. This novel approach eliminates resource contention during concurrent model inference, delivering throughput improvements up to 1.47× compared to existing GPU sharing solutions.
- Designed Everest, a high-performance system for video top-K query processing. Our solution combines a draft model with a probabilistic algorithm to achieve 14.3× to 20.6× speedups over traditional approaches.

Research Intern — Beijing
TuSimple — *May 2018 - June 2019*

- Developed SimpleDet (3.1k ☆), a high-performance object detection framework built on MXNet. The system implements cutting-edge optimizations including mixed-precision training, cross-GPU synchronized batch normalization, gradient checkpointing, and operator fusion, delivering 1.7× faster Mask R-CNN training throughput in FP16 compared to existing frameworks.

Honors and Awards

Best Demo Paper Runner Up, SIGMOD — 2022
Postgraduate Scholarship, CUHK — 2020-2022
Gold Medal, Google AI Open Images Object Detection Track — 2018
Silver Medal, China Collegiate Programming Contest — 2015

Skills

Programming Languages	C/C++, CUDA, Python, Golang, Java, Scala
Frameworks	PyTorch, MXNet, TVM, TensorRT, Horovod, MPI, Ray, Flink