Kele Shao

Hangzhou shaokele@gmail.com +86-15058041526 cokeshao.github.io

Research Interests

My research focuses on Efficient AI and ML Systems (MLSys), specializing in model compression and acceleration. I develop efficient deep learning techniques for computer vision, large language models (LLMs), and embodied AI, aiming to deploy resource-efficient AI systems that bridge research with real-world applications.

Education

Zhejiang University & Westlake University Incoming
Ph.D. in Computer Science and Technology *Advised by Prof. Huan Wang*

Zhejiang University Sept 2021 - June 2025
B.Eng. in Control Science and Engineering *GPA: 3.94/4.0 Rank: 23/121*

Publications

HoliTom: Holistic Token Merging for Fast Video Large Language Models
Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, Huan Wang† arXiv'25/05

LI-GS: Gaussian Splatting with LiDAR Incorporated for Accurate Large-Scale Reconstruction
Changjian Jiang, Ruilan Gao, ***Kele Shao***, Yue Wang, Rong Xiong, Yu Zhang† IEEE RA-L 2025

Projects

Token Compression for Fast Video Large Language Models April 2025 - May 2025

- Employed global redundancy-aware temporal segmentation and spatial-temporal merging for more than **90%** visual token reduction, complemented by a merging approach based on inner LLM token similarity.
- Achieved significant computational savings (6.9% FLOPs, 2.28× TTFT reduction, 1.32× decoding acceleration) on LLaVA-OneVision-7B with minimal performance drop (99.1% maintained).
- Advised by Huan Wang. **First author submitted to NeurIPS'25**.

Neural Pruning with Numerical Continuation Oct 2024 - Feb 2025

- Developed a new pruning paradigm that uses the continuation method for deep neural networks.
- Achieved state-of-the-art performance, including 77.96% accuracy on ResNet50 with only 2 GFLOPs, validating its effectiveness across vision and large language models.
- Advised by Huan Wang. **First author submitted to NeurIPS'25**.

ISC 2024 Student Cluster Competition Oct 2023 - May 2024

- Accelerated the micro-physics (μ phys) module from the ICON model by parallelizing its serial implementation of C++ with OpenMP for heterogeneous platforms, achieving optimal execution times.
- Achieved near-full bandwidth by optimizing data transfer through overlapping and prefetching. Boosted computational performance via loop reordering and CUDA intrinsics.

Autonomous Mapping Vehicle May 2023 - May 2024

- Zhejiang Students' Technology and Innovation Program supervised by Prof. Yu Zhang. Rating: **Excellent**.
- Generated centimeter-accurate color point clouds by integrating modified A-LOAM with image projection.
- Implemented modified 3DGS for novel view synthesis and adaptive TSDF for mesh texture extraction.

Experience

Research Intern *Hangzhou, ZJ*
ENCODE Lab, Westlake University July 2024 - Present

- Conducted research on Efficient AI (quantization, pruning), resulting in 2 publications.
- Contributed to lab infrastructure setup, including server management and website development.

- Advisors: Prof. Huan Wang.

Research Intern *Hangzhou, ZJ*
State Key Laboratory of Industrial Control Technology, Zhejiang University May 2023 - July 2024

- Designed and implemented a large-scale reconstruction method integrating LiDAR and Camera fusion, while performing joint hardware calibration to enhance system accuracy.
- Advisors: Prof. Yu Zhang.

Team Member *Hangzhou, ZJ*
ZJUSCT - Zhejiang University Supercomputing Team Oct 2022 - May 2024

- Led and participated in international supercomputing competitions (ASC, ISC), specializing in parallel acceleration of AI4S problems on multi-node heterogeneous platforms.
- Advisors: Prof. Jianhai Chen, Prof. Zeke Wang, Prof. Yin Zhang, Prof. Shuibing He.

Awards

Outstanding Graduates 2025, Zhejiang University
Zhejiang University First Prize Scholarship 2022, Zhejiang University
Xiaomi Scholarship 2023, Beijing Xiaomi Public Welfare Foundation
Li Yue Venture Capital Scholarship 2023, Shanghai Li Yue Venture Capital Partnership
The Second Prize of the 2024 ASC Student Supercomputer Challenge 2024, ASC24 committee
Bronze medal in the 12^{th} Asia-Pacific Informatics Olympiad (APIO) 2018, China Computer Federation

Technologies

Languages: English (CET-6 509), Mandarin (Native).
Programming: C/C++/CUDA, Python (Pytorch), Shell, Pascal, Matlab, VHDL.
Other Skills: LaTeX, Markdown, Git, 8051, STM32, FPGA, Arduino, Raspberry Pi.