Yuda Fan

Homepage: https://kurodakanbei.github.io/
Email : mistergalahad@gmail.com
Mobile : +86-1895-122-8326
+41-76-475-0337

Preferred Interview Language: C++

Education

- **ETH Zürich** Zürich, Switzerland
  Double enrolled in Direct Doctorate Program and M.Sc. in Computer Science *Sep. 2021 - May. 2024*
- **Shanghai Jiao Tong University** Shanghai, China
  B.Eng. in Computer Science, ACM Honor Class; GPA: 90.3/100, Summa cum laude *Sep. 2016 - Jun. 2020*

Experience

- **Theory Lab, Hong Kong Research Center** Hong Kong, China
  Senior Engineer in Information Theory Group *Nov. 2024 - Present*
  - **Omni-Infer**: An LLM inference architecture enhancement back-boned by vLLM and SGLang. Design and implement the following features: **Multi-Token Speculative Decoding, Rejection Sampling, NPU Accelerated Sampler, MultiStep**. Decrease TPOT by more than 50%. *Technical Report*
  - **CANN Adv Ops**: Improve the MTE1/MTE2 usage of GEMM on Atlas 800I A2.
  - **ArkData/GaussPD**: Improve CPU vector search with intrinsic and assembly instructions on Armv8-A Neon chip. Design and implement the heterogeneous batch vector search scheme with NPU on HiSilicon Kirin chipset.
  - **Skills**: **LLM Infrastructure, vLLM, SGLang, C, Ascend C, Assembly, Triton**
- **CADMO, ETH Zürich** Zürich, Switzerland
  Ph.D. Researcher in Prof. Emo Welzl's Group *Oct. 2022 - Apr. 2024*
  - **Hidden Points and Hidden Vertices**: Prove that the hidden point problem is in $\exists\mathbb{R}$. Introduce novel techniques such as convex/reflex chains, and find PTAS and efficient algorithms for spiral polygons, funnel polygons, pseudo-triangles, fan-shaped polygons, and staircase polygons. *Master's Thesis, JCDCG 2024*
  - **Skills**: **Graph Theory, Computational Geometry, Combinatorics, Scientific Writing**
- **Vision AI Department, Meituan** Beijing, China
  Machine Learning Engineer in Architecture Group *Jul. 2020 - Feb. 2021*
  - **AutoVision**: A platform to automatically conduct neural architecture search, on-device model compression and hyperparameters optimization based on the Alibaba MNN framework. *Highest level patent in 2020*
  - **Memory-Efficient Neural Architecture Search**: A training and inference scheme to eliminate the performance collapse in memory-efficient NAS. *ICCV 2023*
  - **Skills**: **Reinforcement Learning, Neural Architecture Search, PyTorch, Swift, iOS Dev**
- **MVIG, Shanghai Jiao Tong University** Shanghai, China
  Undergraduate Researcher in Prof. Cewu Lu's Group *Jul. 2018 - Jun. 2020*
  - **CyberPanda**: A novel universal robotic arm simulator with photorealistic visual feedback. Integrate the remote procedure call system, rendering pipeline and the physics engine in the software. *Bachelor's Thesis*
  - **Transferable Active Grasping**: Improve the viewpoint optimization strategy to deal with sparse reward issue. Propose a reliable grasping algorithm with higher success rate. *ICRA 2020*
  - **Skills**: **Computer Vision, Universal Robotics, Unreal Engine 4, C#, gRPC**

Selected Honors and Awards

Outstanding Graduate of Shanghai Jiao Tong University	Jun. 2020
Winner, 2025 Huawei Hackathon Software Challenge Final	Oct. 2025
2nd Place, 2025 Huawei Software Challenge Preliminary Contest	Aug. 2025
2nd Place, ICPC 2021-2022 Swiss Subregional Contest	Nov. 2021
2nd Place, ACM-ICPC 2017-2018 Hua-Lien Regional Contest	Nov. 2017
6th Place, ACM-ICPC 2017-2018 Xi'an Regional Contest	Oct. 2017