

Yuda Fan

Homepage: <https://kurodakanbei.github.io/>

Preferred Interview Language: C++

Email : mistergalahad@gmail.com

Mobile : +86-1895-122-8326

+41-76-475-0337

EDUCATION

• ETH Zürich

Double enrolled in Direct Doctorate Program and M.Sc. in Computer Science

Zürich, Switzerland

Sep. 2021 - May. 2024

• Shanghai Jiao Tong University

B.Eng. in Computer Science, ACM Honor Class; GPA: 90.2/100, Summa cum laude

Shanghai, China

Sep. 2016 - Jun. 2020

EXPERIENCE

• Theory Lab, Huawei Hong Kong Research Center

Senior Engineer in Information Theory Group

Hong Kong, China

Nov. 2024 - Present

- **Omni-Infer:** An LLM inference architecture enhancement backended 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*
- **Gauss DataBase:** Improve CPU vector search with intrinsic and assembly instructions on Armv8-A Neon chip. Design and implement the heterogeneous batch vector search scheme on HiSilicon Kirin chipset.
- **Skills:** **LLM Infrastructure, vLLM, C, Ascend C, Assembly, Triton**

• CADMO, ETH Zürich

Ph.D. Researcher in Prof. Emo Welzl's Group

Zürich, Switzerland

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 continuous visibility graph, and find PTAS and efficient algorithms for spiral polygons, funnel polygons, pseudo-triangles, fan-shaped polygons, and staircase polygons. *Master's Thesis*
- **Skills:** **Graph Theory, Computational Geometry, Combinatorics, Scientific Writing**

• Vision AI Department, Meituan

Machine Learning Engineer in Architecture Group

Beijing, China

Jul. 2020 - Feb. 2021

- **AutoVision:** A platform to automatically conduct neural architecture search, on-device model compression and hyperparameters optimization based on the MNN framework.
- **Memory-Efficient Neural Architecture Search:** A training and inference scheme to eliminate the performance collapse in memory-efficient NAS. *Highest level patent in 2020*
- **Skills:** **ML Training, Neural Architecture Search, Pytorch, Swift, iOS Dev**

• MVIG, Shanghai Jiao Tong University

Undergraduate Researcher in Prof. Cewu Lu's Group

Shanghai, China

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, 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
- **Winner, ICPC 2021-2022 Swiss Subregional Contest** Nov. 2021
- **2nd Place, 2025 Huawei Hackathon Software Challenge Preliminary Contest** Aug. 2025
- **2nd Place, ACM-ICPC 2017-2018 Hua-Lien Regional Contest** Nov. 2017
- **6th Place, ACM-ICPC 2017-2018 Xi'an Regional Contest** Oct. 2017