Qichen Fu

fuqichen1998@gmail.com | https://fuqichen1998.github.io

Education

Carnegie Mellon University, School of Computer Science Pittsburgh, PA
Master of Science in Robotics; GPA: 4.24/4.33 *Aug. 2020 - Aug. 2022*

University of Michigan - Ann Arbor, College of Engineering Ann Arbor, MI
Bachelor of Science in Computer Science (dual degree with SJTU); GPA: 4.00/4.00 *Aug. 2018 - Apr. 2020*

Shanghai Jiao Tong University Shanghai, China
Bachelor of Engineering in Electrical and Computer Engineering (dual degree with UM); GPA: 3.73/4.00 *Sept. 2016 - Aug. 2020*

Work Experience

Apple Seattle, WA
Machine Learning Engineer in AI/ML - Machine Intelligence Neural Design (MIND) *Aug. 2022 - Present*

Apple Intelligence – LLM Optimization

- Developed *LazyLLM*, a novel method that allows language models to dynamically select different subsets of tokens from the context in different generation steps, to significantly accelerate the generation without fine-tuning. For instance, in the multi-document QA task, *LazyLLM* accelerates the prefilling stage of the LLama 2 7B by 2.34x while maintaining accuracy. This work is accepted at Efficient Systems for Foundation Models workshop at ICML 2024.
- Developed Superposition Prompting, a novel RAG prompting method that allows the LLM to process input documents in parallel prompt paths, facilitates a 93× reduction in compute time while improving accuracy by 43% on the NaturalQuestions-Open dataset with the MPT-7B. This work is accepted at ICML 2024.
- Developed Speculative Streaming, a new speculative decoding paradigm that does not require a draft model, achieves 1.8× - 3.1× speedups in diverse tasks. This work is under review at EMNLP 2024.

Head Gestures for Airpods

- Led the development of machine learning models and algorithms for head gesture detection, allowing AirPods users to privately respond to Siri with a simple head nod yes or shake no. This feature is shipped at WWDC 2024.

Apple Vision Pro

- Built a multi-host multi-GPU distributed training infrastructure in PyTorch, supporting efficient large-scale training.
- Developed a temporal action classification model that predicts user activities from videos, which is used to mitigate the False Positives of Pinch Detection when the user is holding an object.

Leadership

- Mentored a Research Intern to develop FastSR-NeRF, an efficient NeRF+SR pipeline that speeds up NeRF training by 23× and inference by 18× while maintaining high quality. This work is accepted as an *Oral* paper at WACV 2024.

Selected Publications

LazyLLM: Dynamic Token Pruning for Efficient Long Context LLM Inference ES-FoMo @ ICML 2024
***Qichen Fu**, Minsik Cho, Thomas Merth, Sachin Mehta, Mohammad Rastegari, Mahyar Najibi*

Superposition Prompting: Improving and Accelerating Retrieval-Augmented Generation ICML 2024
*Thomas Merth, **Qichen Fu**, Mohammad Rastegari, Mahyar Najibi*

Speculative Streaming: Fast LLM Inference without Auxiliary Models Arxiv 2024
*Nikhil Bhendawade, Irina Belousova, **Qichen Fu**, Henry Mason, Mohammad Rastegari, Mahyar Najibi*

eDKM: An Efficient and Accurate Train-time Weight Clustering for Large Language Models IEEE CAL 2024
*Minsik Cho, Keivan A Vahid, **Qichen Fu**, ..., Peter Zatloukal*

Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation ICCV 2023
***Qichen Fu**, Xingyu Liu, Ran Xu, Juan Carlos Niebles, Kris M. Kitani*

Domain Adaptive Hand Keypoint and Pixel Localization in the Wild ECCV 2022
*Takehiko Ohkawa, Yu-Jhe Li, **Qichen Fu**, Ryosuke Furuta, Kris M. Kitani, Yoichi Sato*

Sequential Voting with Relational Box Fields for Active Object Detection CVPR 2022
***Qichen Fu**, Xingyu Liu, Kris M. Kitani*

Ego4D: Around the World in 3,000 Hours of Egocentric Video CVPR 2022 *Oral*, *Best Paper Finalist*
*Kristen Grauman, ..., **Qichen Fu**, ..., Jitendra Malik*

Research Experience

KLab, Carnegie Mellon University Pittsburgh, PA
Research Assistant; ***Advisor:*** *Prof. Kris Kitani* *Oct. 2020 - Aug. 2022*

- Led the video de-identification, state change object detection benchmark and challenge development of the EGO4D dataset.
- Developed a Dynamic Fusion Transformer framework for robust 3D hand pose estimation from videos.
- Developed a pixel-wise voting function with a Relational Box Field to robustly detect active objects under occlusions.

Fouhey AI Lab, University of Michigan Ann Arbor, MI
Research Assistant; ***Advisor:*** *Prof. David Fouhey* *May 2019 - May 2020*

- Developed an unsupervised object detection system predicting bounding boxes and articulation type for objects in video.
- Built an artificial object detection system for image filtering, reaching an accuracy of 95.06% and an AUC score of 0.92.

Fessler Research Group, University of Michigan Ann Arbor, MI
Research Assistant; ***Advisor:*** *Prof. Jeffrey A. Fessler, Prof. Yuni Dewaraja* *Oct. 2018 - May 2020*

- Developed complex-valued U-Net for MRI reconstruction, reducing parameters by 50% compared to the vanilla U-Net.
- Developed a novel method integrating back-projection and 3D U-Net for PET reconstruction directly from measurements.

Teaching Experience

Carnegie Mellon University Pittsburgh, PA

- 16-824: Visual Learning and Recognition (Spring 2022), advised by Prof. Deepak Pathak
- 16-720B: Computer Vision (Fall 2021), advised by Prof. Kris Kitani

University of Michigan Ann Arbor, MI

- EECS 442: Computer Vision (Winter 2020), advised by Prof. David Fouhey
- EECS 442: Computer Vision (Fall 2019), advised by Prof. David Fouhey

Honors

University of Michigan: Jackson and Muriel Lum Scholarship, James B. Angell Scholar, University Honors
Shanghai Jiao Tong University: National Scholarship, Undergraduate Excellent Scholarship, MiYuan Public Welfare Scholarship