XIJIE HUANG

Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong SAR
E-mail: huangxijie1108@gmail.com | Homepage: https://huangowen.github.io/ | Google Scholar: nFW2mqwAAAAJ&hl

EDUCATION

Hong Kong University of Science and Technology Hong Kong SAR
Ph.D. in Computer Science Engineering, HKUST Vision and System Design Lab *Sept 2020 - Present*

- Advisor: Prof. Tim Kwang-Ting CHENG
- HKUST Postgraduate Studentship and RedBird Scholarship

Shanghai Jiao Tong University Shanghai, China
B.E. in School of Electronics Information and Electrical Engineering *Sept 2016 - June 2020*

- Overall GPA: 89.4/100 (91.3/100 for junior year) Ranking:2/55
- Advisor: Prof. Cewu Lu, Machine Vision and Intelligence Group, SJTU

University of California, Los Angeles Los Angeles, USA
Visiting Research Student *June 2019 - Sept 2019*

- Research intern to UCLA ECE department (Cross-disciplinary Scholars in Science & Technology Program)
- Advisors: Prof. Mani B. Srivastava, Department of Electrical Computer Engineering, UCLA

RESEARCH INTERESTS

My research interests lie in the general area of artificial intelligence, particularly in efficient large-scale models (LLMs, Diffusion Models) and human-centric computer vision. More concretely, my research interests focus on designing quantization algorithms, algorithm-hardware co-design, human-object interaction, and healthcare.

PUBLICATIONS & PRE-PRINT

▷ Efficient AI Algorithm

Taming High-Resolution Text-to-Image Models for Mobile Devices with Efficient Architectures
Dongting Hu*, Jierun Chen*, **Xijie Huang***, Huseyin Coskun, Arpit Sahni, Aarush Gupta, Anujraaj Goyal, Dishani Lahiri, Rajesh Singh, Yerlan Idelbayev, Junli Cao, Yanyu Li, Kwang-Ting Cheng, Mingming Gong, S.-H. Chan, Sergey Tulyakov, Anil Kag, yanwu xu, Jian Ren (* indicates Equal Contribution)
IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2025 (Acceptance Rate: 22.1%)

Fewer is More: Boosting LLM Reasoning with Reinforced Context Pruning
Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, Fan Yang, Mao Yang
Conference on Empirical Methods in Natural Language Processing (EMNLP) 2024 (Acceptance Rate: 20.8%)

RoLoRA: Finetuning Outlier-free Model with Rotation for Weight-Activation Quantization
Xijie Huang, Zechun Liu, Shih-Yang Liu, Kwang-Ting Cheng
Conference on Empirical Methods in Natural Language Processing (EMNLP) Findings 2024

Efficient Quantization-aware Training with Adaptive Coreset Selection
Xijie Huang, Zechun Liu, Shih-Yang Liu, Kwang-Ting Cheng
Transactions on Machine Learning Research (TMLR)

Quantization Variation: A New Perspective on Training Transformers with Low-Bit Precision
Xijie Huang, Zhiqiang Shen, Pingcheng Dong, Kwang-Ting Cheng
Transactions on Machine Learning Research (TMLR)

LLM-FP4: 4-Bit Floating-Point Quantized Transformers
Shih-Yang Liu, Zechun Liu, **Xijie Huang**, Pingcheng Dong, Kwang-Ting Cheng
Conference on Empirical Methods in Natural Language Processing (EMNLP) 2023 (Acceptance Rate: 21.3%)

SDQ: Stochastic Differentiable Quantization with Mixed Precision
Xijie Huang, Zhiqiang Shen, Shichao Li, Zechun Liu, Xianghong Hu, Jeffry Wicaksana, Eric Xing, Kwang-Ting Cheng
International Conference on Machine Learning (ICML) 2022 (Acceptance Rate: 21.9%)

▷ Efficient AI Hardware

A 28nm 0.22J/Token Memory-Compute-Intensity-Aware CNN-Transformer Accelerator with Hybrid-Attention-Based Layer-Fusion and Cascaded Pruning for Semantic-Segmentation
Pingcheng Dong, Yonghao Tan, Xuejiao Liu, Peng Luo, Yu Liu, Luhong Liang, Yitong Zhou, Di Pang, Manto Yung, Dong Zhang, **Xijie Huang**, Shih-Yang Liu, Yongkun Wu, Fengshi Tian, Chi-Ying Tsui, Fengbin Tu, Kwang-Ting Cheng
IEEE International Solid-State Circuits Conference (ISSCC), 2025

Genetic Quantization-Aware Approximation for Non-Linear Operations in Transformers
Pingcheng Dong, Yonghao Tan, Dong Zhang, Tianwei Ni, Xuejiao Liu, Yu Liu, Peng Luo, Luhong Liang, Shih-Yang Liu, **Xijie Huang**, Huaiyu Zhu, Yun Pan, Fengwei An, Kwang-Ting Cheng
ACM/IEEE Design Automation Conference (DAC) 2024

A Tiny Accelerator for Mixed-bit Sparse CNN based on Efficient Fetch Method of SIMO SPad
Xianghong Hu, Xuejiao Liu, Yu Liu, Haowei Zhang, **Xijie Huang**, Xihao Guan, Luhong Liang, Chi Ying Tsui, Xiaomeng Xiong, Kwang-Ting Cheng
IEEE Transactions on Circuits and Systems II: Express Briefs (TCAS-II) 2023

▷ Human-Centric Vision

Automated Vision-Based Wellness Analysis for Elderly Care Centers
Xijie Huang, Jeffry Wicaksana, Shichao Li, Kwang-Ting Cheng
AAAI Conference on Artificial Intelligence (AAAI) W3PHIAI 2022

FedMix: Mixed Supervised Federated Learning for Medical Image Segmentation
Jeffry Wicaksana, Zengqiang Yan, Dong Zhang, **Xijie Huang**, Huimin Wu, Xing Yang, Kwang-Ting Cheng
IEEE Transactions on Medical Imaging (TMI)

Transferable Interactiveness Knowledge for Human-Object Interaction Detection
Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, **Xijie Huang**, Liang Xu, Cewu Lu
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

Latent Fingerprint Image Enhancement based on Progressive Generative Adversarial Network
Xijie Huang, Peng Qian, Manhua Liu
IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 Biometric Workshop

PaStaNet: Toward Human Activity Knowledge Engine
Yong-Lu Li, Liang Xu, Xinpeng Liu, **Xijie Huang**, Shiyi Wang, Hao-Shu Fang, Cewu Lu
IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020 (Acceptance Rate: 22.09%)

Transferable Interactiveness Knowledge for Human-Object Interaction Detection
Yong-Lu Li, Siyuan Zhou, **Xijie Huang**, Liang Xu, Ze Ma, Hao-shu Fang, Yanfeng Wang, Cewu Lu
IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019 (Acceptance Rate: 25.15%)

INTERNSHIP

Snap Research Santa Monica, CA
PhD Research Intern in Creative Vision Group *July 2024-*

- Mentor: Jian Ren, Anil Kag, and Yanwu Xu
- Project: Researching the efficiency of text-to-image (T2I) models and distillation scheme for diffusion models.
- Proposed an extremely small and fast T2I model SnapGen that generates high-quality images on mobile platforms. Our model, for the first time, demonstrates the generation of 1024^2 px images on a mobile device in $1.2 \sim 2.3$ seconds. On T2I benchmarks (i.e., GenEval and DPG-Bench), our model with merely 379M parameters, surpasses large-scale models with billions of parameters at a significantly smaller size (e.g., 7× smaller than SDXL, 14× smaller than IF-XL).

Microsoft Research Asia (MSRA) Beijing, China
PhD Research Intern in Systems Research Group *May 2023-Feb 2024*

- Mentor: Li Lyna Zhang
- Project: Researching the efficiency of in-context learning of Large Language Models (LLMs).
- Proposed CoT-Influx [EMNLP 2024], a novel approach to push the boundaries of few-shot CoT learning to improve LLM math reasoning capabilities. We propose a coarse-to-fine pruner as a plug-and-play module for LLMs, which first identifies crucial CoT examples from a large batch and then further prunes unimportant tokens

RESEARCH PROJECT

Hardware-Software Co-design of Model Compression, HKUST

- Proposed a LoRA-based scheme for weight-activation quantization RoLoRA [EMNLP 2024]. RoLoRA utilizes rotation for outlier elimination and proposes rotation-aware fine-tuning to preserve the outlier-free characteristics in rotated LLMs.
- Proposed Stochastic Differentiable Quantization (SDQ) [ICML 2022], an efficient and effective mixed-precision quantization technique outperforming full-precision ResNet/MobileNet on ImageNet with an average bitwidth lower than 4.
- Proposed an efficient variation-aware vision transformer (ViT) quantization framework [TMLR]. It is the first work to analyze and locate the variation in ViT quantization. Our solution to variation in ViTs leads to state-of-the-art accuracy on the ImageNet-1K dataset across different ViT models (DeiT, Swin, SReT).
- Propose a new angle through the coreset selection [TMLR] to improve the training efficiency of quantization-aware training. Our method can achieve 68.39% of 4-bit ResNet-18 on the ImageNet-1K dataset with only a 10% subset.

Automated Vision-Based Wellness Analysis for Elderly Care Centers, HKUST

- Building a vision-based elderly care system [AAAIW 2022] that can provide immediate assistance and useful insights for caretakers. Collaborating with the Heaven of Hope care center and built a healthcare dataset based on video recording.

Machine Vision and Intelligence Group, Department of Computer Science, SJTU

- Proposed Transferable Interactiveness Network [CVPR 2019] to tackle the imbalanced distribution in human action recognition problems, especially human-object interaction detection problems
- Built the state-of-the-art knowledge base and engine of human activity understanding HAKE. HAKE provides elaborate and abundant with 7 M+ fine-grained part level annotations in a large scale of images and videos.

Networked & Embedded Systems Laboratory, Department of Electrical Computer Engineering, UCLA

- Proposed a Trojan backdoors detection framework called NeuronInspect [arxiv], using visual interpretability technique to effectively detect Trojan backdoors in deep neural networks without restoring the trigger and any backdoor samples

SELECTED ACADEMIC ACHIEVEMENTS

National Scholarship (Top 2% students in Shanghai Jiao Tong University)	2017
A Class Scholarship (Top 2% students in Shanghai Jiao Tong University)	2017
CSST Scholarship (USD $5,343) University of California, Los Angeles	2019
RongChang Academic Scholarship (The highest honor in SJTU, Top 20 of 16000 students)	2019
Postgraduate Studentship, Hong Kong University of Science and Technology	2020-2024
AAAI-22 Student Scholarship	2022
Microsoft Research Star of Tomorrow	2023
EMNLP 2024 Travel Grant	2024

SERVICES AND EXPERIENCES

Open Source Project

- Awesome LLM Compression (1.5K stars, 4/3/2025)

Reviewer

- Conference: ICLR 2024, ACM MM 2024, EMNLP 2023-2024, NeurIPS 2023-2024, ICCV 2023, CVPR 2023-2025, AAAI 2022-2025, WACV 2022-2023, ICML 2022-2025 (Top 10% Reviewer in 2022), ECCV 2022
- Journal: TNNLS, TMLR

Program Committee

- ICCV 2023 Workshop on Low-Bit Quantized Neural Networks
- ICCV 2023 Workshop on Resource Efficient Deep Learning for Computer Vision

Teaching Assistant

- COMP 2211 (Exploring Artificial Intelligence), Fall 2022/Spring 2024, Lecture: Professor Desmond Tsoi
- COMP 5421 (Computer Vision), Spring 2021, Lecture: Professor Dan Xu
- COMP 1021 (Introduction to Computer Science), Fall 2021, Lecturer: Professor David Rossitor