# Kai Yao

PhD Student · Trustworthy ML

*Informatics Forum, 10 Crichton Street, EH8 9AB, Edinburgh, UK*

kai.yao@ed.ac.uk | https://kaikaiyao.github.io

## Education

**University of Edinburgh** — *Edinburgh, UK*

PhD - Cyber Security, Privacy and Trust — *2023 - present*

- Research: Privacy-Preserving ML, Fairness in ML
- Advisor: Dr. Marc Juarez

**Johns Hopkins University** — *Baltimore, US*

MS - Mechanical Engineering — *2020*

**Fudan University** — *Shanghai, CN*

BS - Theoretical and Applied Mechanics — *2017*

## Industry Experience

2021-2023 **AI Frameworks Engineer (Tech Lead)**, Intel (Full-Time)

2020-2021 **AI Algorithms Engineer**, Huawei (Full-Time)

## Publications

Rochman ND*, **Yao K***, Gonzalez NA*, Wirtz D, Sun SX. Single cell volume measurement utilizing the fluorescence exclusion method (FXm). Bio-protocol. 2020 Jun 20;10(12):e3652-.

**Yao K***, Rochman ND*, Sun SX. CTRL–a label-free artificial intelligence method for dynamic measurement of single-cell volume. Journal of cell science. 2020 Apr 1;133(7):jcs245050.

Perez-Gonzalez NA*, Rochman ND*, **Yao K***, Tao J, Le MT, Flanary S, Sablich L, Toler B, Crentsil E, Takaesu F, Lambrus B. YAP and TAZ regulate cell volume. Journal of Cell Biology. 2019 Oct 7;218(10):3472-88.

**Yao K***, Rochman ND*, Sun SX. Cell type classification and unsupervised morphological phenotyping from low-resolution images using deep learning. Scientific reports. 2019 Sep 17;9(1):1-3.

Zhang Q, Meng Z, Zhang Y, **Yao K**, Liu J, Zhang Y, Jing L, Yang X, Paliwal N, Meng H, Wang S. Phantom-based experimental validation of fast virtual deployment of self-expandable stents for cerebral aneurysms. BioMedical Engineering OnLine. 2016 Dec;15(2):431-7.

(Note: * denotes equal contributions to the paper.)

## Awards, Fellowships, & Grants

2023 **LFCS Travel Funds - PETS conference**, LFCS, University of Edinburgh
**School of Informatics PhD Scholarship**, University of Edinburgh
**Division Achievement Award - Fast Stable Diffusion on Intel CPU**, AIA, Intel
**Division Recognition Award - Neural Coder Partnering Alibaba Cloud**, CESG SW AI, Intel

2022 **Division Achievement Award - Innovation of Neural Coder**, AIA, Intel

2017 **Departmental Research Fellowship - AI in Biology**, Johns Hopkins University
**Fudan Outstanding Graduate of the Year 2017**, Fudan University

2014 **JASSO Full Scholarship for Exchange Students**, Japanese Government

## Research Experience

**CTRL - Image-based AI method for single-cell 3D morphology and size prediction** *Baltimore, US*

Johns Hopkins University *2019 - 2020*

- Developed a label-free and high-throughput AI-based technique that predicts single-cell 3D morphology and size from DIC microscopy images.
- Designed and implemented a microfluidics system that uses the fluorescence exclusion method to measure single-cell morphology by quantifying fluorescence exclusion.
- Designed image processing algorithms for pre-processing both DIC microscopy images and FXm fluorescent images to serve as CNN model input.
- Developed a CNN model structure based on U-Net structure and experimented with hyperparameter tuning to achieve the best prediction outcome.

**Cell type classification and morphological phenotyping via Deep Learning** *Baltimore, US*

Johns Hopkins University *2018 - 2019*

- Constructed a label-free and high-throughput AI-based technique that classifies normal cells vs. cancer cells using only low-res cell flask images.
- Designed image processing algorithms and established an automation pipeline for screening and pre-processing microscopy images for the CNN model.
- Developed a clustering method based on CNN feature maps from pre-trained models to group cells by morphology and study tumor cell shape.
- Analyzed the relationship between cell type, cell density, and cell morphology to understand how cancer cells move and grow in-vitro.

## Industry Experience

**AI Frameworks Engineer, Domain Lead (Org: AIA AIPC DL)** *Shanghai, CN*

Intel Corp. *2021 - 2023*

- Led a team of engineers in developing Neural Coder, an automation tool that inserts Deep Learning model optimization code into PyTorch and TensorFlow model scripts with one-click, streamlining the optimization process and reducing development time.
- Developed multiple features in Intel Extension for PyTorch, an Intel AI software that optimizes the computational efficiency of Deep Learning kernels and kernel fusions on Intel hardware, resulting in faster DL training and inference with PyTorch.
- Developed PyTorch adapter algorithms in Intel Neural Compressor, a toolkit for applying model compression techniques such as INT8 quantization, improving PyTorch model compression productivity while maintaining optimal model accuracy.
- Conducted out-of-box model-level and kernel-level benchmarks of various Deep Learning workloads on Intel CPU and GPU hardware, as well as on competitor hardware (e.g. NVIDIA GPU and AMD CPU), using tools such as Torch Profiler and BenchDNN, to identify areas of improvement and optimize AI software performance.
- Optimized CPU inference performance and conducted benchmarks for open-source AIGC models such as Stable Diffusion, DALLE, GPT-J, and BLOOM using model acceleration techniques such as mixed precision computation and SmoothQuant.
- Designed and developed a comprehensive Deep Learning workload benchmarking system for Intel's internal software products, encompassing hardware infrastructure setup, benchmark automation, data management, automatic data analysis, and web-based result visualization.
- Collaborated with Intel AI business partners (e.g. Alibaba PAI, AWS, HuggingFace, PyTorch) teams to integrate Intel's AI software product features into their Machine Learning platforms, resulting in increased market share and revenue.

**AI Algorithm Engineer (Org: 5G Solutions Design)** *Shanghai, CN*

Huawei Technologies Co., Ltd. *2020 - 2021*

- Designed and developed 5G Machine Learning learning algorithms, with a specific focus on 5G MU-MIMO feature, using key model structures such as MLP, CNN, and RNN.
- Optimized the Machine Learning algorithms and compressed ML models for efficient training and inference, and improved model prediction accuracy through hyperparameters tuning and data filtration scheme design.
- Collaborated closely with AI algorithm deployment and feature validation teams throughout the feature development cycle, ensuring seamless integration and smooth implementation of the final product.

## Teaching Experience

2019-2020 **Mathematical Image Analysis**, Teaching Assistant @Johns Hopkins University