# Zhihan Lu

+65 8542 8522 | zhihanl@alumni.cmu.edu | Website | Google Scholar

## Education

**Carnegie Mellon University**, Pittsburgh, US — August 2021 – December 2022
*Master of Science in Machine Learning* — GPA: 4.17/4.33

**Rice University**, Houston, US — August 2017 – May 2021
*B.A. in Computer Science and B.A. in Mathematics* — GPA: 3.9/4.0
*Awards: Louis J. Walsh Scholarship in Engineering, 2019 – 2021, President's Honor Roll, 2017 – 2020*

## Publications

**[1] Early-bird gcns: Graph-network co-optimization towards more efficient gcn training and inference via drawing early-bird lottery tickets**
Haoran You, **Zhihan Lu**, Zijian Zhou, Yonggan Fu, Yingyan Lin
*36th AAAI Conference on Artificial Intelligence (AAAI 2022)* [Paper], [Code]

**[2] Max-affine spline insights into deep network pruning**
Haoran You, Randall Balestriero, **Zhihan Lu**, Yutong Kou, Huihong Shi, Shunyao Zhang, Shang Wu, Yingyan Lin, Richard Baraniuk
*Transactions on Machine Learning Research (TMLR 2022)* [Paper], [Code]

**[3] SACoD: Sensor algorithm co-design towards efficient CNN-powered intelligent PhlatCam**
Yonggan Fu, Yang Zhang, Yue Wang, **Zhihan Lu**, Vivek Boominathan, Ashok Veeraraghavan, Yingyan Lin
*IEEE/CVF International Conference on Computer Vision (ICCV 2021)* [Paper], [Code]

## Research Experience

**Rice Efficient and Intelligent Computing Lab | Research Assistant** — January 2020 - August 2021
*Efficient model training and inference through early pruning and NAS* — *Advisor: Yingyan Lin*

- Proposed joint pruning of graph edges and model weights in Graph Convolutional Network during training, achieving 277x FLOPS reduction without compromising accuracy. [1]
- Developed a principled pruning approach with training-stopping criteria based on the Max-affine Spline theory to analyze model's decision boundaries. [2]
- Built an experimental codebase that enables Neural Architectural Search (NAS) for sensor-network co-deisn in vision tasks. [3]

## Work Experience

**Shopee | Senior Software Engineer** | Singapore, SG — January 2023 - present
*Unified Ranking Service for E-commerce, Livestream, and Short Videos*

- Developed a high-performance C++ graph engine powering all Shopee recommendations (>85k QPS).
- Simplified system architecture and optimized feature processing for more efficient memory access.
- Awarded *Outstanding Project* for saving >70k (33%) CPU cores and cutting latency by 30% (>50ms).
- Led seamless service migration through effective cross-functional collaboration and project planning.

**Waymo | Machine Learning Engineer Intern** | Mountain View, US — May 2022 – August 2022

- Productionized a gradient-based algorithm to identify influential training samples for evaluation sample.
- Built and deployed the backend service using Python, C++, Apache Beam, and Tensorflow.

**Amazon | Software Development Engineer Intern** | Seattle, US — May 2021 – August 2021

- Doubled the detection rate of invalid item details in Amazon's catalog via a rule-based classifier.
- Designed and deployed a scalable auditing API on AWS, enhancing human-in-the-loop reviews.
- Improved data preprocessing throughput 10x via Java multithreading and database performance tuning.

**LinkedIn | Software Engineering Intern** | Sunnyvale, CA — May 2020 – August 2020

- Improved the precision of LinkedIn's people-match model by 5% via targeted feature engineering.
- Halved Spark pipeline runtime by benchmarking and removing performance bottlenecks with Scala.

## Technical Skills

Python, C++, C, SQL, Pytorch, TensorFlow, CUDA, Keras, Polars, PySpark, gRPC