Bingyi Zhang

bingyizh@usc.edu
(571)-443-9187

EDUCATION

Fudan University — Shanghai, China
B.Eng., Micro Electronics — *Sept. 2013 - Jun. 2017*

Fudan University — Shanghai, China
Master of Engineering, Integrated circuit engineering — *Sept. 2017 - Jun. 2019*

University of Southern California — Los Angeles, USA
Ph.D. of Computer Engineering, Ming Hsieh Dept of ECE — *Sept. 2019 - present*

PUBLICATION

- Yi Chien Lin; **Zhang, Bingyi**; Prasanna, Viktor. GCN Inference Acceleration using High-Level Synthesis. 2021 IEEE High Performance Extreme Computing Virtual Conference (HPEC 2021).
- **Zhang, Bingyi**; Sanmukh R. Kuppannagari; Kannan, Rajgopal ; Prasanna, Viktor. BoostGCN: Efficient Neighbor-Sampling-based GNN Trainingon CPU-FPGA Heterogeneous Platform. 2021 IEEE High Performance Extreme Computing Virtual Conference (HPEC 2021).
- **Zhang, Bingyi**; Kannan, Rajgopal ; Prasanna, Viktor. BoostGCN: A Framework for Optimizing GCN Inference on FPGA. FCCM, 2021.
- **Bingyi Zhang**, Hanqing Zeng, Viktor Prasanna, "A Framework for Optimizing GCN Inference on FPGA", The 29th ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA 2021).
- **Zhang, Bingyi**; Zeng, Hanqing; Prasanna, Viktor, Hardware acceleration of large scale gcn inference, The 31st IEEE International Conference on Application-specific Systems, Architectures and Processors, 2020
- **Bingyi Zhang**, Hanqing Zeng, Viktor Prasanna, "Accelerating Large Scale GCN inference on FPGA", The 28th IEEE International Symposium On Field-Programmable Custom Computing Machines. 2020.
- **Bingyi Zhang**, Xin Li, Jun Han, Xiaoyang Zeng, "MiniTracker: A Lightweight CNN-based system for Visual Object Tracking on Embedded Device", International Conference on Digital Signal Processing, 2018.
- **Bingyi Zhang**, Jun Han, Zhize Huang, Jianwei Yang, Xiaoyang Zeng, "A Realtime and Hardware-efficient Processor for Skeleton-based Action Recognition with Lightweight Convolutional Neural Network", IEEE Transaction on Circuits and Systems II: Express Briefs
- J. Yin, J. Han, C. Wang, **Bingyi Zhang** and X. Zeng, "A Skeleton-based Action Recognition System for Medical Condition Detection," 2019 IEEE Biomedical Circuits and Systems Conference (BioCAS), Nara, Japan, 2019, pp. 1-4.

PROFESSIONAL SKILLS

Expertise: Parallel computing, digital circuit design, digital signal processing.
Hardware design:

- *Hardware description languages (HDL)*: Vivado HLS, Verilog.
- *Design automation Tool (EDA)*: Vivado, Design compiler, IC compiler, Cadence,

Programming Languages: C, C++, Python, Matlab, Perl, Latex
Software Tool: Caffe, Tensorflow, Pytorch

RESEARCH EXPERIENCE

FPGA/PARALLEL COMPUTING LAB, USC *Sep. 2019 - Jan. 2020*

Topic: Accelerating Large scale GCN inference on FPGA

- Under the mentorship of senior Ph.D. student Hanqing, develop a algorithm-architecture cooptimization framework to achieve high throughput GCN inference on FPGA.
- The algorithmic optimizations consist of graph sparsification and node reordering which can reduce the external memory traffic by 30%-40%.
- The proposed handware architecture achieve pipeline execution between two major computation kernels in GCN. It also support various GCN models and computational orders.
- The proposed framework can achieve 30 times speedup compared with 56-thread multi-core system and 2 times speedup compared with Titan Xp GPU platform.

State Key Laboratory of ASIC and System, Fudan University *Aug. 2017 - May. 2018*

Topic: skeleton-based human action recognition

- Design an lightweight 1D convolution neural network (NN) to perform real-time human action recognition, and achieve 20x acceleration compared to the baseline.
- Quantize and trim the proposed NN to increase hardware implementation suitability.
- Collaborate with colleagues to design a processor to perform the proposed algorithm, achieving low-power and hardware-efficient implementation.

State Key Laboratory of ASIC and System, Fudan University *Sep. 2016 - June. 2017*

Topic: visual object tracking on embedded system

- Design an Siamese neural network for object tracking, which is computation-saving and storage-saving.
- Collaboratively design a hardware architecture to deal with the proposed algorithm.
- Implement the proposed architecture on FPGA and achieve real-time (18.6 frame/s) and robust tracking with power consumption 1.284W, outperforming the baseline.

ACADEMIC COMPETITION

China Graduate Mathematical Modeling Competition *Nov. 2017*

Result: Second Prize (Top 10%)

- Personal contribution: algorithm design and software development

First National Undergraduate Integrated Circuit Innovation and Entrepreneurship Competition of China *Feb. 2017*

Topic: Neural network accelerator. Result: Rank 3rd out of 150+ teams

- Personal contribution: Team leader, algorithm designer and software developer
- Collaboratively design a neural network accelerator

Third National Undergraduate Integrated Circuit Innovation and Entrepreneurship Competition of China *Feb. 2019*

Topic: Hardware acceleration for convolutional neural networks. Result: Second price out of 500+ teams

- Personal contribution: hardware developer and software developer.
- Collaboratively build a hardware accelerator which can achieve 14 fps for VGG16 Net on FPGA.