Peng Zan

San Diego, CA zanpeng.pz@gmail.com

EDUCATION

Ph.D., Electrical and Computer Engineering	12/2019
University of Maryland, College Park, MD	GPA 3.6
M.Sc., Electrical and Computer Engineering	08/2019
University of Maryland, College Park, MD	GPA 3.8
B.Sc., Electrical Engineering	07/2014
Xi'an Jiaotong University, Xi'an, China	GPA 90/100
Exchange, Electronic Engineering	05/2013
Chinese University of Hong Kong, Hong Kong, China	GPA 3.7

WORK EXPERIENCE

Sr. Software Engineer	05/2025-present
Waymo, Mountain View	CA
Staff Engineer, AI Software	02/2024-05/2025
Samsung Semiconductor, Inc., San Diego	CA
Sr. Software Engineer - AI Framework	08/2021-09/2023
Black Sesame Technologies, San Jose	CA
Principal Scientist	02/2020-06/2021
Origin Wireless AI, Greenbelt	MD
DSP Research Intern	05/2019-08/2019
Starkey Hearing Technologies, Eden Prairie	MN
Graduate Research Assistant	08/2015-05/2019
University of Maryland, College Park	MD

EXPERTISE

AI Accelerator and Compiler Framework, Mathematical Optimization for Signal Processing

SKILLS

Programming: Python, C/C++
Software Tools: onnx-mlir, LLVM/MLIR, Pytorch, Tensorflow, Git/Github, Linux/Unix, Matlab, R, LaTeX
Software Skills: Neural Network Acceleration and Compilation on dedicated SoC, NPU, DSP, GPU and CPU
Data Science: Statistics, Digital Signal Processing, Machine Learning, Deep Learning

PROJECTS

GenAI on Edge 02/2024 - 05/2025
AI Framework R&D NPU Compiler Team, Samsung Semiconductor

- ➢ Developed a post-training quantization framework based on Tflite, reducing the EDSR model to 4-bit precision and enhancing its accuracy from 27dB to 33dB by a custom fine-tuning algorithm based on greedy search optimization.
- ➢ Led the edge deployment of a diffusion-based 3D object generative model (InstantMesh), managing all front-end tasks, including Torch-to-TFLite/ONNX model conversion, graph optimization, quantization, and accuracy enhancement with SmoothQuant, achieving high-quality on-device outputs.
- ➢ Optimized a distilled version of InstantMesh based on GECO, and reduced the diffusion steps to 1; reduced pipeline latency from 40s to 8.5s, about 5× faster.
- ➢ Architected and optimized pipeline workloads across heterogeneous backends, including NPU, GPU, and CPU, and collaborated with other teams to develop mobile applications for CES 2025, streamlining the deployment process.
- ➢ Optimized runtime for quantized Customized ONNX model in Exynos AI toolchain using graph optimization with QDQ and GPU utilization; decreased the latency from 191min to 6min for stable diffusion 1.5, $>$ 30× faster.

Efficient Deep Learning and LLM 09/2023 - 01/2024
AI Accelerator R&D Independent Projects

- ➢ Implemented Neural Architecture Search (NAS) based on MCUNetV2 using Evolutionary Search, achieved 90.7% accuracy on VWW dataset with only 30M MACs and 200kB peak memory.
- ➢ Quantized LLaMA2-7B with 4-bit weight and 8-bit activation based on Activation-aware Weight Quantization, and deployed it to an M1 MacBook with Parallel Computing techniques of Loop Unrolling, Multithreading and SIMD Programming, achieving real-time interaction.

AI Compiler Framework 02/2022 - 09/2023
AI Framework R&D Software Team, Black Sesame Technologies

- ➢ Devised and implemented graph optimization pipelines based on ONNX for efficient inference on A1000 Pro NPU.
- ➢ Designed and implemented Post-Training Quantization (PTQ) flow, quant-related A1000 Pro NPU ISA specs calculation and CodeGen based on ONNX.
- ➢ Created and deployed debugging tools for bit-exact testing, achieving 4× efficiency in identifying root causes.
- ➢ Architected and implemented next-generation compiler infrastructure for mixed-precision inference, including quantized data type definitions, related optimization passes with lowering and canonicalization, ISA specs calculation and CodeGen for BST A2000 NPU, utilizing onnx-mlir and LLVM/MLIR technologies.
- ➢ Led research on optimizing neural network inference scheduling for heterogeneous SoC architectures with Big.Little CPU + GPU, and achieved 2× faster runtime compared to Tflite-XNNPack.

Neural Network Compression and Acceleration[P1][P2][P3] 08/2021 - 05/2023
AI Framework R&D Software Team, Black Sesame Technologies

- ➢ Analyzed and identified factors affecting inference accuracy after quantization with A1000 Pro NPU constraints; optimized quantization algorithm to match floating-point model accuracy on a set of benchmark models[P1][P2][P3].
- ➢ Designed and optimized quantization pipelines and algorithms on A1000 Pro NPU using mathematical modeling and optimization theory, enhanced the compiler's stability and boosted testing process for the chip mass production.
- ➢ Designed and helped implement the Quantization-Aware Training (QAT) flow considering A1000 Pro NPU limitations on fixed-point arithmetic.
- ➢ Collaborated on developing an advanced NPU micro-architecture for A2000 that supports mixed-precision leveraging insights from the A1000 NPU, and expanded the data type support from (u)int8 to int4, (u)int8, fp16 and fp32.

WiFi Sensing and Internet of Things (IoT)[P4]-[P8] 02/2020 - 06/2021

- Led a research project on Activity of Daily Living (ADL), and developed a real-time algorithm for activity location and level estimation using WiFi sensing, achieving a detection rate of 95% and a false alarm rate of $< 5\%$.
- Won the CES 2021 Innovation Award with ADL algorithm, which was later deployed to HEX Home product.
- Designed and automated manufacture workflow for WiFi-sensing products by Python, and boosted production rate 12× faster from 1 per hour to 12 per hour; ensured on-time delivery to Verizon, Inc. and won us $\$14M$ investment.
- Developed a real-time tracking system with sub-meter accuracy based on a *Bayesian dynamic model on graph.*
- Built Android and iOS Apps of Origin Tracking, which work without WiFi (iOS demo and Android demo).

JOURNAL PUBLICATIONS

[J1] **Peng Zan**, Alessandro Presacco, Samira Anderson, and Jonathan Z. Simon. Exaggerated cortical representation of speech in older listeners: mutual information analysis. *Journal of Neurophys., 124(4):1152-1164*, Oct. 7, 2020.

[J2] **Peng Zan**, Alessandro Presacco, Samira Anderson, and Jonathan Z. Simon. Mutual information analysis of neural representations of speech in noise in the aging midbrain. *Journal of Neurophysiology Innovative Methodology, 122(6): 2372-2387*, Dec. 4, 2019.

[J3] Kai Lu, Wanyi Liu, Kelsey Dutta, **Peng Zan**, Jonathan B Fritz, and Shihab A. Shamma. Adaptive efficient coding of correlated acoustic properties. *Journal of Neuroscience, 39(44):8664-8678*, Oct. 30, 2019.

[J4] Kai Lu, Wanyi Liu, **Peng Zan**, Stephen V. David, Jonathan B Fritz, and Shihab A. Shamma. Implicit memory for complex sounds in higher auditory cortex of the ferret. *The Journel of Neuroscience, 38(46):9955-9966*, Nov. 14, 2018.

[J5] Junmin Liu, Yongchang Hui, and **Peng Zan**. Locally linear detail injection for pansharpening. *IEEE Access, 5:9728-9738*, June 7, 2017.

[J6] Dai Wang, Xiaohong Guan, Jiang Wu, Pan Li, **Peng Zan**, and Hui Xu. Integrated energy exchange scheduling for microgrids with electric vehicles. *IEEE Transaction on Smart Grid, 7(4):1762–1774*, July 10, 2016.

[J7] Xiaoming Du, Stephanie Hare, Ann Summerfelt, Bhim Adhikari, Laura Garcia, Wyatt Marshall, **Peng Zan**, Mark Kvarta, Eric Goldwaser, Heather Bruce, Si Gao, Hemalatha Sampath, Peter Kochunov, Jonathan Z. Simon, Elliot Hong. Cortical Connectomic Mediations on Gamma Band Synchronization in Schizophrenia. *Translational Psychiatry, Nature Publishing Group*, Jan. 19, 2023.

[J8] Yujie Zhang, Huiying Lan, Ehsan Aghapour, Zhiyuan Ning, **Peng Zan**, Weidong Shao. Para-Pipe: Exploiting Hierarchical Operator Parallelism of ML Computational Graphs on SoCs. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, May 8, 2025.

CONFERENCE PAPERS & POSTERS

[C1] Wenqiang Pu, **Peng Zan**, Jinjun Xiao, Tao Zhang, Zhi-Quan Luo. Evaluation of joint auditory attention decoding and adaptive binaural beamforming approach for hearing devices with attention switching. *2020 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 08, 2020.

[C2] **Peng Zan**, Alessandro Presacco, Samira Anderson, and Jonathan Z. Simon. Mutual information analysis of neural representations of speech in noise in the aging midbrain. *ARO 2019.*, Feb. 2019.

[C3] **Peng Zan**, Alessandro Presacco, Samira Anderson, and Jonathan Z. Simon. Cortical over-representation of speech in older listeners correlates with a reduction in both behavioral inhibition and speech intelligibility. *ARO*, Feb. 2019.

[C4] **Peng Zan**, Alessandro Presacco, Samira Anderson, and Jonathan Z. Simon. Mutual information analysis of neural representations of speech in noise in the aging midbrain. *Auditory SPLASH*, Sep. 8, 2018.

[C5] **Peng Zan**, Alessandro Presacco, Samira Anderson, and Jonathan Z. Simon. Mutual information analysis of neural representations of speech in noise in the aging midbrain. *EAR*, June 15, 2018.

PATENTS

[P1] **Peng Zan**. System and method for neural network structure-level quantization optimization. U.S. Application No. *18209932*, 12/19/2024.

[P2] **Peng Zan**. System and method for mathematical modeling of hardware quantization process. U.S. Application No. *18081515*, Patent No. *US20240202501A1*, 06/20/2024.

[P3] **Peng Zan**. System and method for mathematical modeling of hardware quantization process. CN Application No. *CN202311631256.8A*, Patent No. *CN117852596A*, 04/09/2024.

[P4] Beibei Wang, Zahid Ozturk Muhammed, Chenshu Wu, Xiaolu Zeng, Deepika Regani Sai, Yuqian Hu, K.J. Ray Liu, Chi-Lim Au Oscar, Yi Han, Hung-Quoc Duc Lai, David N. Claffey, Dan Bugos, **Peng Zan**. Method, device, and system for sound sensing and radio sensing. *JP2022191191A*, 12/17/2022.

[P5] Chenshu Wu, Beibei Wang, **Peng Zan**, Sai Deepika Regani, Xiaolu Zeng, Hung-Quoc Lai, Kj Ray Liu, Oscar Au. Method, apparatus, and system for wireless micro motion monitoring. *US20210311166A1*, 10/7/2021.

[P6] Beibei Wang, Muhammed Zahid Ozturk, Chenshu Wu, Xiaolu Zeng, Sai Deepika Regani, Yuqian Hu, K. J. Ray Liu, Oscar Chi-Lim Au, Yi Han, Hung-Quoc Duc Lai, David N. Claffey, Chun-I Chen, Dan Bugos and **Peng Zan**. Method, apparatus, and system for sound sensing and wireless sensing. EP Patent Application No. 22178761.7, filed June 13, 2022.

[P7] Yuqian Hu, Beibei Wang, Sai Deepika Regani, **Peng Zan**, Chenshu Wu, Dan Bugos, Xiaolu Zeng, Hung-Quoc Duc Lai, K. J. Ray Liu, Oscar Chi-Lim Au. Method, apparatus, and system for wireless sensing based on linkwise motion statistics. U.S. Patent Application No. 17/838,244, filed June 12, 2022.

[P8] Chenshu Wu, Beibei Wang, Oscar Chi-Lim Au, K.J. Ray Liu, Chao-Lun Mai, Dan Bugos, Hung-Quoc Duc Lai, Spencer Maid, Yuqian Hu, Sai Deepika Regani, Muhammed Zahid Ozturk, Xiaolu Zeng, Fengyu Wang, Jeng-Feng Lee and **Peng Zan**. Method, apparatus, and system for wireless monitoring to ensure security. EP Patent Application No. 21200823.9, filed October 4, 2021.

PEER REVIEWS

[R1]	IEEE Signal Processing Letters	05/2021, 06/2021, 08/2021, 10/2023, 02/2024, 05/2025
[R2]	IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing	01/2020, 03/2020
[R3]	IEEE Access	07/2019
[R4]	Neuroscience Letters	04/2021, 07/2021

SELECTED AWARDS & HONORS

Starkey Recognition Award	Starkey, 08/2019
NSF-Funded COMBINE Fellowship (Computational Biological Network Program)	UMD, 09/2017
Jimmy H. C. Lin Graduate Scholarship for Entrepreneurship	UMD, 09/2014
ECE Ph.D. Fellowship Award	UMD, 09/2014
National Scholarship, Ministry of Education of the P.R.C.	XJTU, 11/2011