4516 Wilderness Trl
Pueblo, CO 81008
LinkedIn

# AHMED HASSSAN

(602) 727-7346
ah2288@cornell.edu
Webpage

## RESEARCH INTEREST

- On-device AI, SSM and Transformer Language Models, Neural Radiance Fields (NeRF), 3D Gaussian Splatting, Spiking Neural Networks, Self-supervised Learning (SSL), Pruning, Quantization and Neural Architecture Search, AI-based Software/Hardware codesign, Neuromorphic Computing, Virtual and Augmented Reality.

## LANGUAGES AND TOOLS

- **Tools and Languages:** C, C++, GGML, Object-Oriented Programming, Python (PyTorch, Tensorflow, JAX, Keras), OpenGL, Verilog, System Verilog, VS Code, Linux, CUDA, Fast-AI, ML-Flow, GPU programming, MATLAB, Git

## EDUCATION

**Cornell University (Tech)** **NYC, New York** **Aug 2023 – Aug 2025**
- Ph.D. in Electrical Engineering CGPA: 3.70

**Arizona State University** **Tempe, AZ** **Aug 2020 – Aug 2023**
- Ph.D. in Electrical Engineering CGPA: 3.72

**Government College University** **Lahore, Pakistan** **Oct 2017 – Sep 2019**
- MS in Electrical Engineering CGPA: 3.75

**COMSATS University** **Islamabad, Pakistan** **Feb 2011 – Mar 2015**
- BS in Electrical Engineering CGPA: 3.37

**Relevant Courses**
- Advanced Applied ASIC Design, VLSI Design, Neuromorphic Computing Hardware Design, EffecientML.ai (Han Lab, MIT),MLHardware and Systems, Computer Vision and Pattern Recog, Machine Learning: Bayesian Perspective.

## EMPLOYMENT

**On-Device AI Research Intern** **AI-Center, Samsung Research America, Mountain View, CA** **May 2024 – Aug 2024**
- **Optimization of Prefill Time/ Time to First Token Generation of Mamba (SSM) for Faster On-device Inference**
  - Performed LLM latency profiling on CPU, GPU, and Samsung S23 CPU using Llamacpp to identify bottlenecks.
  - Compared the throughput and latency of Transformer and Mamba models on CPU, GPU, and **Samsung S23 CPU**.
  - Implemented divide and concur-based Parallel Scan on Pytorch and Llamacpp to reduce the on-device prefill/first token generation latency for Mamba1.1 and Mamba2.8 models with higher context length >5k.
  - Reduced the prompt eval (**TTFT**) latency by **Log(N)** and achieved **1.39X** speedup for **Mamba 1.1B** on a **Samsung S23 device** using Llamacpp.

**Graduate Research Assistant** **Cornell Tech (Cornell University), NY** **Aug 2023 – Present**
- **Shallow and Compressed Representation of Transformer and SSM Language Models (Mamba 1.1, Tiny Llama)**
  - Quantization and pruning of small language models for low latency and hardware-level optimization.
- **Low-precision and Memory Efficient Neural Radiance Fields (NeRF) with Hardware Accelerator Design**
  - Implemented fixed-grid-based quantization scheme for 3D Gaussian, statistically aware quantizationfor MLP activations, and adaptively rounded quantization for weights to make the complete NeRF flow low-precision.
  - Replaced the compute inefficient positional encoding with the fixed-precision (8-bit and 3-bit) look-uptables for software/hardware level efficiency. Achieved high PSNR and SSIM with very low rendering time.
- **Low-compute Spiking Neural Network (SpQuant-SNN) Architecture for DVS-based Computer Vision**
  - Designed a 3-bit integer-only membrane potential quantization scheme with pruning for SNN architecture.
  - Benchmarked SNN on complex datasets with >3X memory, >5X FLOPs reduction, and <1% accuracy drop.

**Design Technology Enablement Intern** **Intel Corporations, Hillsboro, OR** **May 2022 – March 2023**
- **Estimation of Local Layout Effect (LLEs) using Machine Readable Specs (MRS)**
  - Using LLE rules from the QA team, designed machine-readable specs to estimate the LLEs' presence in the layouts.
  - Translated MRS into Python utility and validated automated LLE estimation using different layouts.
  - Provided prototype version of python-based utility to QA and TTR team for small and large-scale layout testing.
- **Python-Based Automated Layout Generation for Different Cell Types**
  - Developed Python algorithm to generate automatic and DRC clean standard cells, and memory cells layouts.
  - Implemented an algorithm to identify the key requirements and discrepancies of run-sets for auto-correction.

**Research Associate** **Seo Lab, Arizona State University, AZ** **Aug 2021 – Aug 2023**
- **DARPA Project: Low Precision Autoencoder Design for Information Processing from Event-based Camera**
  - Designed low-precision sparse autoencoder architecture for Event (Prophesee-Gen1) data compression.

- Designed log of two-based quantization module to covert weights and activations to 4-bit and 2-bit precision.
- Achieved high accuracy of >91% and mean average precision of 0.30 with >10X image data compression ratio.

- **Hardware-efficient Spiking Neural Network (SNN) for DVS-based Computer Vision Applications**
  - Implemented learnable potential threshold-based efficient SNN algorithm (LT-SNN) for edge AI applications.
  - Implemented low-precision of weights, membrane potential, and activations (2-bit,4-bit, and 8-bit).
  - Improved state-of-the-art accuracy by more than 2.8% and mAP 0.11 with 10.68× less SNN model size.

**Research Associate** **Optoelectronics Lab, ASU, AZ** **Aug 2020 – July 2021**

- **DOE Funded Project** worth **$2 million: Polarization Camera based Drone Imaging System for Concentrated SolarPower Plants' Defects Detection**
  - Formulated polarization camera-based imaging setup using Nvidia Jetson Tx2 and Sony Imaging sensors.
  - Finished 90% of integration and testing of imaging setup with AltaX Freefly Drone.
- **DARPA Funded Project: Underwater Housing for the Object Localization in the Deep Sea**
  - Developed a polarimetric imaging system with Nvidia Jetson Xavier for robust imaging.

**Research Officer** **Computer Vision Lab, KICS LHR, PK** **Mar 2020 – Oct 2020**

- **Local Industry Project: Deep Leaning-based Face attendance system, Jetson Tx2 and PYNQ (XILINX)**
  - I worked on the algorithm design and customization of the YoloV3 model for inference on Jetson Tx2.

**Lab Engineer** **Sharif College of Engineering LHR, PK** **May 2015 – Feb 2020**

- **Lab Instructor in the Department of Electrical and Computer Engineering**

## PUBLICATIONS

- Hasssan, A., Meng, J., Anupreetham, A., Seo, J. "SpQuant-SNN: Ultra-low Precision Membrane Potential with Sparse Activations Unlock the Potential of On-device Spiking Neural Networks Applications*". Frontiers of Neuroscience, Section Neuromorphic Computing, 2024.*
- Hasssan, A., Meng, J., Seo, J. "Spiking Neural Networks with Learnable Threshold for Event-based Classification and Object Detection". *International Joint Conference on Neural Networks, IJCNN, 2024*.
- Hasssan, A., Meng, J., Anupreetham, A., Seo, J. "IM-SNN: Memory-Efficient Spiking Neural Networks with Low-precision Membrane Potentials and Weights". *International Conference on Neuromorphic Systems, ICONS, 2024*.
- Meng, J., Leo, Y., Anupreetham, A., Hasssan, A., et al. Torch2Chip: An End-to-end Customizable Deep Neural Network Compression and Deployment Toolkit for Prototype Hardware Accelerator Design". *Conference on Machine Learning and Systems*, *MLSys, 2024*.
- Nair, G.R., Nalla, P.S., Krishnan, G., Anupreetham, A., Hasssan, A., Yeo, I., et al. "3D In-Sensor Computing for Real-time DVS Data Compression: 65nm Hardware-Algorithm Co-design". *IEEE Solid-State Circuits Letters*, 2024.
- Krishnan, G., Nair, G.R., Oh, J., Anupreetham, A., Nalla, P.S., Hassan, A., Yeo, I., et al. "3D-ISC: A 65nm 3D Compatible In-Sensor Computing Accelerator with Reconfigurable Tile Architecture for Real-Time DVS Data Compression." *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2023.
- Rafique, M. Z. E., Faruque, H. M. R., Hassan, A., Tian, M., Das, N., & Yao, Y. "Field Deployable Mirror Soiling Detection Based on Polarimetric Imaging". SolarPACES Conference Proceedings, 2023.
- Seo, J.S., Saikia, J., Meng, J., Hasssan, A., et al. "Digital versus analog artificial intelligence accelerators: Advances, trends, and emerging designs." *IEEE Solid-State Circuits Magazine 14.3 (2022): 65-79*.
- Hasssan, A., Meng, J., Cao, Y., & Seo, J. S. "Spatial-temporal Data Compression of Dynamic Vision Sensor Output with High Pixel-level Saliency using Low-precision Sparse Autoencoder". *56th Asilomar Conference on Signals, Systems, and Computers, 2022 (pp. 344-348)*.
- Meng*, J., Hasssan*, A., and Seo, J. "LT-SNN: Self-adaptive Spiking Neural Network with Learnable Threshold". *Techcon, Semiconductor Research Corporation, 2022*.

## ACADEMIC PROJECTS

- **Optimization of Convolutional and Linear Layer Kernels for On-device Resource and Latency Optimization**
  - Implemented low-precision training of audio model on Arduino Nano with 4-bit min-max quantization.
  - Deployed TVM-based different Primitives including Blocking, Vectorization, and Parallelism on CPU and GPU.
  - Developed Convolution layers with Winograd and Fast Fourier Transform for resource optimization.
- **Low-Precision Convolution Neural Network-Based Processing Engine Design for Classification of MNIST Dataset**
  - Designed shallow (1 Conv and 1 Linear layer) low precision (2-bit weights and 2-bit activations) CNN architecture.
  - Designed and implemented log2-based quantization of weights and activations for better hardware efficiency.
  - Wrote RTL and developed a pipelined design for MAC units to optimize latency and synthesize the design.
  - Synthesized the design using Design Compiler, validated synthesized netlist, and achieved frequency 10Mhz.
- **Automatic assessment of atherosclerotic plaque features from intracoronary OCT images with deep learning**
  - Prepared intracoronary OCT cardiovascular data for six classes i-e calcified, fibrous, lipid, and thrombus.
  - Designed a GAN and ResNet-34-based classification engine to identify different plaques and their severity in the patients.