

Muhammad Ahmad Waseem

+1-716-939-9936 | ahmadwaseem648@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#) | Buffalo, NY

TECHNICAL SKILLS

Speech & ASR: Wav2Vec 2.0, Conformer, Whisper, CTC, Kaldi, NVIDIA NeMo, phoneme modeling, speaker diarization, TTS
GPU & Systems: CUDA, nvcc, Triton, TensorRT, NVIDIA A100, Jetson AGX Xavier, INT8/FP16 quantization
Quantization & Compression: QAT, AWQ, GPTQ, SmoothQuant, LLM.int8, model pruning, knowledge distillation
ML Frameworks & Libraries: PyTorch, HuggingFace Transformers, NVIDIA NeMo, TensorFlow, OpenCV, scikit-learn
Languages: Python, C++, C, CUDA
MLOps & Infra: Git, Docker, FastAPI, Linux, PostGIS, Google Earth Engine

EXPERIENCE

University at Buffalo (SUNY)

Graduate Research Assistant

Buffalo, NY

Aug. 2024 – Jul. 2026

- Reduced Phoneme Error Rate **from 80% to 15%** on children's speech (CSLU Kids, MyST, PhonBank) via Wav2Vec 2.0 fine-tuning and AutoPhon auto-annotation pipeline
- Developed outlier-aware QAT framework for Conformer ASR; achieved **7.17 WER vs. 93.7 naive INT8 baseline**, enabling standard INT8 deployment without custom runtime kernels
- Implemented CUDA kernels for scaled dot-product attention and feed-forward layers on **A100 GPU**; profiled memory- vs. compute-bound regimes against cuBLAS and PyTorch baselines
- Benchmarked AWQ, GPTQ, SmoothQuant across A100, Jetson, Raspberry Pi; derived device-specific strategies for edge deployment

CITY at LUMS

ML Engineer — Computer Vision & Systems

Lahore, Pakistan

Aug. 2021 – Jul. 2024

- Optimized YOLO+DeepSORT vehicle tracking on NVIDIA AGX Xavier via TensorRT; achieved **5× throughput increase (3→14 FPS)**; deployed city-wide with Punjab Safe Cities Authority
- Designed TFNet segmentation architecture (Dilated ResNet + dual decoders); achieved **94% F1** on SpaceNet2 and WHU benchmarks, surpassing DeepLabV3+ baselines; published at *IEEE IGARSS 2025*
- Built city-scale waste route optimization platform (Clarke–Wright savings algorithm, PostGIS, FastAPI, Vue.js); achieved **~100K liters/month fuel savings**; production deployment across Lahore
- Developed multi-sensor flood mapping pipeline on Google Earth Engine (Sentinel-1/2, Landsat-9); identified damage to **1,410 km of road infrastructure** during 2022 Pakistan floods

SELECTED PROJECTS

Speech-to-Phoneme Modeling for Children's ASR | PyTorch, Wav2Vec 2.0, Kaldi, NVIDIA NeMo 2025 – 2026

- Built AutoPhon annotation pipeline using pretrained models for auto-segmentation and phoneme-level transcription, reducing manual annotation time by **10×**
- Benchmarked S2P models (Wav2Vec 2.0, Whisper-style, Kaldi-HMM) on CSLU Kids, MyST, PhonBank; reduced PER **from 80% to 15%**; manuscript under review *IEEE SMC 2026*

CUDA Kernels for Transformer Workloads | CUDA, C++, NVIDIA A100, nvcc

2024 – 2025

- Implemented end-to-end attention (scaled dot-product) and feed-forward kernels with explicit thread/block mappings and memory-aware tiling targeting A100 shared memory hierarchy
- Profiled custom kernels against cuBLAS and PyTorch native ops; quantified compute- vs. memory-bound regimes and identified bottlenecks for transformer inference optimization

Hardware-Aware Quantization Benchmarks | PyTorch, Triton, torchao, AWQ, GPTQ

2024 – 2025

- Evaluated AWQ, GPTQ, and SmoothQuant across A100 GPU, Jetson, and Raspberry Pi CPU; exposed cases where GPU quantization gains do not transfer to edge and CPU-only hardware

SELECTED PUBLICATIONS

[S.1] M. A. Waseem et al. *reverseDAAM: Region-to-Token Interpretability for Text-to-Image Diffusion Models*. Under Review, **IEEE SMC 2026**.

[C.1] M. A. Waseem et al. *A Tuning-Fork Network for Improved Building Footprint Extraction*. **IEEE IGARSS 2025**.

[C.2] M. A. Waseem et al. *Evaluating the Cooling Efficacy of Urban Green Spaces During Extreme Heat Events*. **IEEE IGARSS 2025**.

[C.3] M. Awan, ..., M. A. Waseem et al. *Unsupervised Landmark Discovery Using Consistency Guided Bottleneck*. **BMVC 2023**.

EDUCATION

University at Buffalo (SUNY)

Master of Science in Computer Science — GPA: 3.945/4.0

Buffalo, NY

Aug. 2024 – Jul. 2026

Information Technology University (ITU)

Master of Science in Electrical Engineering — GPA: 3.83/4.0

Lahore, Pakistan

Sep. 2020 – Mar. 2023

University of Engineering & Technology (UET) Lahore

Bachelor of Science in Electrical Engineering — GPA: 3.52/4.0

Lahore, Pakistan

Oct. 2016 – Jun. 2020