

Reza Baharani

Senior AI Engineer | AI Infrastructure, Edge Intelligence & Agentic AI | PhD

📍 Charlotte, NC

✉ baharanireza@gmail.com

🌐 mbaharan.github.io

🌐 mbaharani

🌐 mbaharan

🎓 Google Scholar

Experience

ForesightCares Inc., Senior AI/ML Engineer

- Architected multi-agent AI workflows that automate complex healthcare compliance research and generate structured, review-ready documentation
- Built a cloud-native serverless AI service on AWS Lambda to deliver personalized exercise recommendations from curated clinical protocols
- Implemented scalable state-machine architecture for 5+ assessment protocols with automated monitoring and robust handling of multi-user scenarios
- Developed Field-of-View calibration model to correct 3D pose estimation across varying camera angles, improving measurement accuracy for at-home and clinical deployment scenarios
- Created 3D motion capture dataset using IMU-based mocap system for training and validation, collecting synchronized 60fps video and 400Hz motion data from 100+ subjects to train, validate and test our in-house pose estimator
- Designed a multi-stage signal processing pipeline that improved measurement stability and accuracy for joint-angle computation and gait analysis

Charlotte, NC
Apr 2025 – present
1 year 6 months

UNC Charlotte, Research Scientist

- Developed a self-supervised foundation model with tokenized Vision Transformers, achieving 89.4% action classification and 76.56% AUC-ROC on anomaly detection; accepted to BMVC 2026

Charlotte, NC
Oct 2024 – Mar 2025
6 months

ForesightCares Inc., ML/AI Engineer

- Architected and deployed an end-to-end AI/ML inference pipeline achieving 30fps real-time 3D human pose estimation, from data collection through production deployment and monitoring
- Engineered a cross-platform edge inference system with model optimization and deployment automation, eliminating cloud dependencies and reducing model bundle size through native infrastructure
- Co-invented AI-powered mobility assessment system (US Patent 12,343,138 B2), enabling automated fall risk evaluation and real-time exercise coaching via 3D skeleton modeling from monocular video

Charlotte, NC
Jan 2022 – Sept 2024
2 years 9 months

UNC Charlotte, Research Scientist

- Architected enterprise-scale distributed AI infrastructure enabling real-time multi-camera inference across hybrid edge/cloud deployments with privacy-aware data governance

Charlotte, NC
Aug 2021 – Dec 2022
1 year 5 months

Astranis Space Technologies, FPGA Deployment Systems Engineer (Internship)

- Supported FPGA deployment engineering for satellite systems, applying reconfigurable-computing and hardware/software integration expertise in an industry production environment

Dec 2020 – Mar 2021
4 months

UNC Charlotte, Graduate Research Assistant

Charlotte, NC

Aug 2017 – Aug 2021

4 years 1 month

- Designed an algorithm-architecture co-design framework for neural networks on heterogeneous compute platforms, enabling power-efficient inference through hardware-software optimization
- Developed a scalable real-time monitoring and reliability-prediction system for IoT infrastructure operating within a 1.87W power budget
- Created lightweight temporal convolutional network for vehicle trajectory prediction in highways, matching state-of-the-art accuracy with significantly reduced computational complexity (published in IEEE TITS, 97 citations)
- Co-developed an LLVM-based system architecture simulator for accelerator modeling (74 citations, IEEE/ACM MICRO 2020)
- Published 12+ peer-reviewed papers across IEEE and ACM venues with 657 citations (h-index: 11) spanning AI systems, model optimization, and distributed computing
- Designed attention-based temporal convolutional networks (ATCN) enabling fast time-series classification on resource-constrained embedded devices with formalized hyperparameters for application-specific architecture tuning (published in ACM TECS)
- Deployed ATCN on Cortex-M7 microcontroller consuming only 49% of 320KB RAM and 15% of 1MB flash; validated across 70 UCR 2018 time-series benchmarks using data augmentation (jittering, magnitude warping, window warping, scaling)
- Developed real-time person re-identification system at the edge using mixed precision (FP16/FP32) inference, achieving efficient tracking on embedded GPUs with MobileNetV2 backbone (published at ICIAR 2019, 17 citations)

Education

PhD University of North Carolina at Charlotte, Electrical and Computer Engineering

Charlotte, NC

Aug 2017 – Aug 2021

- Advisor: Prof. Hamed Tabkhi
- Focus: Computer Architecture and Deep Learning

Publications

A Feasibility Study on Self-Supervised LLM-Inspired Training for Generalizable Human Motion Understanding

Aug 2026

Reza Baharani, G.A. Noghre, A.D. Pazho, G. Maldonado, H. Tabkhi
(British Machine Vision Conference (BMVC), accepted; arXiv:2502.05432)

Ancilia: Scalable Intelligent Video Surveillance for the AIoT

Jan 2023

A.D. Pazho, C. Neff, G.A. Noghre, B.R. Ardabili, S. Yao, Reza Baharani, H. Tabkhi
(IEEE IoT Journal)

ATCN: Resource-Efficient Processing of Time Series on Edge

Jan 2022

Reza Baharani, H. Tabkhi
(ACM Trans. Embedded Computing Systems (TECS))

DeepTrack: Lightweight Deep Learning for Vehicle Trajectory Prediction in Highways

Jan 2022

V. Katariya, Reza Baharani, et al.
(IEEE Trans. Intelligent Transportation Systems)

Real-Time Deep Learning at the Edge for Scalable Reliability Modeling of Si-MOSFET Power Electronics Converters

Jan 2019

Reza Baharani, et al.
(IEEE IoT Journal)

REVAMP2T: Real-Time Edge Video Analytics for Multicamera Privacy-Aware Pedestrian Tracking

Jan 2019

C. Neff, M. Mendieta, S. Mohan, Reza Baharani, et al.
(IEEE IoT Journal)

Patents

- US 12,343,138 B2: AI Powered Mobility Assessment System, ForesightCares Inc., Jul. 2025. Inventors: H. Tabkhivayghan, M. Azarbayjani, R. Baharani.

Skills

AI & Machine Learning: Deep Learning, Computer Vision, Human Motion Understanding, 3D Pose Estimation, Self-Supervised Learning, Vision Transformers, Time-Series Modeling, Signal Processing, Generative AI, Agentic AI, RAG, Multimodal AI

AI Systems & Optimization: AI Inference & Serving, Distributed AI Systems, Edge AI, MLOps, Multi-Agent Orchestration, Model Optimization, Quantization, Knowledge Distillation, Pruning, Hardware/Software Co-design

Languages & Frameworks: Python, C/C++, CUDA, Java, SQL, PyTorch, TensorFlow, ONNX, LangChain, Flask

Infrastructure & Platforms: AWS, AWS Lambda, Docker, Kubernetes, Terraform, GitHub Actions, Linux, Git, iOS/iPadOS

Hardware & Compilers: NVIDIA Jetson, FPGA, ASIC, Cortex-M, HLS, HDL, MLIR