

Contact

faraz.kayani2322@gmail.com

www.linkedin.com/in/faraz-kayani
(LinkedIn)

Top Skills

Computer Vision

Python (Programming Language)

Machine Learning

Certifications

Generative AI with Large Language Models

AI Skills Fest 2026

Network Security

CyberOps Associate

Publications

Vision Model Inference on Mobile Devices: A Large-Scale Delegate-Aware Benchmark Beyond FLOPs

Understanding the Performance of Deep Computer Vision Models: A Symbolic Regression Approach to Accuracy and Latency Prediction

Real Image Denoising with Knowledge Distillation for High-Performance Mobile NPUs

Faraz Kayani

Computer Vision & Machine Learning Engineer | Deep Learning & Model Optimization | BMVC & CVPRW 2026 Author

Würzburg, Bavaria, Germany

Summary

I build and deploy computer vision models that run efficiently on real hardware, from training in PyTorch to compressing and shipping to mobile NPUs at production-grade latency. First author on a paper published at CVPR Workshops 2026, introducing LiteDenoiseNet: a 1.96M-parameter denoising network distilled from a 41.6M-parameter teacher, running at 34ms per frame on-device (a 3.9x speedup over GPU inference on the same hardware). Also co-author on a paper accepted to BMVC 2026 benchmarking vision model inference across mobile hardware at scale, and a Sensors 2026 paper on predicting deep vision model accuracy and latency via symbolic regression, with an additional manuscript currently under review at Neurocomputing. At the Computer Vision Lab (CAIDAS & IFI), University of Würzburg, I build the infrastructure that makes this repeatable: automated PyTorch to TFLite/ONNX conversion pipelines, cross-device CPU/GPU/NPU benchmarking, and CI-driven model evaluation on physical hardware, not simulators. Before my M.Sc., I spent two years as a software engineer (Zenkoders) shipping full-stack web and mobile products in Python, JavaScript, React, and React Native, including AI/chatbot integrations for production clients. That background means I bring shipping discipline, not just research instincts, to ML work. Core stack: PyTorch · TensorFlow/TFLite · ONNX · Python · C++ · Docker · Git · Linux M.Sc. Computer Science, University of Würzburg, 2026. Available for full-time Computer Vision / Machine Learning Engineer roles in Germany.

Experience

Computer Vision Laboratory Wuerzburg

Research Engineer, Computer Vision & Machine Learning

October 2025 - August 2026 (11 months)

Würzburg

Researched and developed efficient computer vision models for on-device deployment, working across model architecture design, knowledge distillation, quantization, and mobile deployment pipelines. Collaborated with faculty and PhD researchers on multiple publications spanning image denoising, latency prediction, and neural architecture search.

- Designed LiteDenoiseNet, a 1.96M-parameter student network distilled from a 41.6M-parameter teacher, achieving 37.58 dB PSNR, first-authored at IEEE/CVF CVPR Workshops 2026, where our team placed 1st in the MAI Image Denoising Challenge
- Built a full PyTorch-to-TFLite deployment pipeline via AI Edge Torch, benchmarking inference at 34ms on Dimensity 9500 and 46.1ms on Snapdragon 8 Elite
- Discovered an "Inference Inversion" effect (NPU inference ~3.9x faster than mobile GPU on identical models), now guiding hardware-aware design decisions across the lab
- Engineered open-source nn-dataset/nn-lite benchmarking pipelines, automating ADB-based deployment and CPU/GPU/NPU evaluation across physical devices
- Co-authored a large-scale mobile inference benchmark accepted to BMVC 2026, demonstrating that delegate selection outperforms FLOPs as a predictor of real on-device latency

Zenkoders

Software Engineer

July 2021 - June 2023 (2 years)

Karāchi

Software engineer at a client-services agency delivering web and mobile applications across multiple industries.

- Built and shipped full-stack web and mobile features, including REST APIs and database integration, across concurrent client engagements
- Integrated AI and chatbot functionality into production applications using Python, JavaScript, React, and React Native

- Collaborated with designers, QA, and PMs in Agile sprints, contributing to code reviews and technical scoping

Bitrupt

Full-Stack Developer Intern

February 2020 - April 2020 (3 months)

Karāchi

- Built React and React Native applications from prototype through feature-complete delivery using JavaScript, React Native, Firebase, and SQLite

Education

The Julius Maximilians University of Würzburg

Master of Science (M.Sc.) in Computer Science · (2023 - 2026)

DHA Suffa University

Bachelor of Science (B.Sc.) in Computer Science · (2017 - 2021)