# Anh Tran

Philadelphia, PA, United States | tranhuyhoanganh220702@gmail.com | (+1) 267 336 5564

Website: anh-thh.github.io | LinkedIn: Huy Hoang Anh Tran | GitHub: anh-thh

## EDUCATION

**University of Pennsylvania**, M.S.E. in Electrical Engineering - GPA: 3.92/4.0 Sep 2024 – May 2026

**Relevant coursework**: System-on-a-Chip Architecture, Computer Organization & Design, General-Purpose GPU Architecture & Programming, HW/SW Co-Design for ML, Digital IC & VLSI Fundamentals, Hardware Security (Spring 2026)

**VinUniversity**, B.Sc in Electrical Engineering - GPA: 3.83/4.0 Sep 2021 – Jun 2025

**Relevant coursework:** Digital Logic & Computer Organization, Computer System Programming, Artificial Intelligence, Natural Language Processing, Digital Signal and Image Processing

## RESEARCH EXPERIENCE

**Research Assistant | A Dedicated MicroBlaze SoC for Deep Neural Networks | EDABK LAB** Jul 2025 – Present

**Supervised by:** Assoc. Prof. Duc Minh Nguyen

- Developed a custom System-on-Chip (SoC) on the UltraScale+ ZCU104 featuring a MicroBlaze soft processor, with full support for DDR4, UART, and GPIO interfaces.
- Actively working on deploying convolutional neural network (CNN) models on FPGA and integrating the accelerator into the custom SoC architecture.

**Research Assistant | Federated Learning for Graph-based ICD auto coding | VinUniversity** Jul 2025 – Present

**Supervised by:** Assistant Prof. Danh Cuong Do

- Leveraged Graph Neural Networks to enhance the diagnosis of International Classification of Diseases (ICD) codes in language models.
- Ongoing effort to develop a federated learning architecture that supports collaborative training across multiple hospital clients while preserving data privacy and training efficiency.

**Research Assistant | Implicit Deep Learning | VinUniversity** May 2024 – Oct 2024

**Supervised by:** Prof. Laurent El Ghaoui

- Designed experiments to verify and evaluate the generalization performance of implicit models across various architectures (fully connected, residual, attention-based, RNNs), a new class of deep learning models proposed by Laurent El Ghaoui.
- Explored and deployed various solvers (MOSEK, ADMM, and projected gradient descent) for fixed-point equations, a key component in training implicit models.
- Examined the sparsity and representational capacity of implicit models by analyzing patterns in weight matrices.

**Research Assistant | Satellite Imagery Super-resolution for Carbon Stocks Estimation | VinUniversity** Mar 2024 – Oct 2024

**Supervised by:** Assoc. Prof. Nidal Kamel

- Utilized deep learning methods to super-resolve satellite images, incorporating dynamic high-pass filtering and channel attention to enhance image generation.
- Trained an image super-resolution model using data from Vietnam's mountainous and forest regions; used quality-enhanced images as input for a carbon stock estimator, integrating neural networks to refine predictions.

## PROFESSIONAL EXPERIENCE

**AI/Data Engineer Intern | AlphaAsimov Robotics** Mar 2024 – Oct 2024

- Performed data preprocessing and analysis to ensure readiness for AI model training; assessed the alignment of various modalities (camera, LIDAR, SONAR, IMU, GPS, etc.) in the dataset.
- Designed tools using Python and Bash scripting, to streamline and partially automate the data verification process, incorporating anomaly detection models and descriptive visualizations.

## TEACHING EXPERIENCE

**CIS 5710: Computer Organization & Design | University of Pennsylvania** Spring 2026

- Incoming responsibilities include answering students' online questions, holding office hours, and grading exams

**ESE 5060: Introduction to Optimization Theory | University of Pennsylvania** Fall 2025

- Responsibilities included answering students' online questions, grading homework, and grading/proctoring exams

## PROJECTS

**High-Performance CUDA Kernel for CSR-Dense Matrix Multiplication** Oct 2025 – Present

- Designed a custom Compressed Sparse Row (CSR) × dense CUDA kernel with an optimized thread layout for coalesced access and minimized bank conflicts, along with shared-memory tiling to improve data reuse.
- Achieved up to 1.5× speedup over cuSPARSE and 1.39× speedup over torch.sparse baseline.
- Ongoing effort to explore additional optimization strategies, including load-balancing schemes across warps/threads and alternative work partitioning to further accelerate the kernel.

**System-on-Chip (SoC) Design for Real-Time Data Deduplication and Compression** Oct 2025 – Dec 2025

- Developed a comprehensive deduplication-compression pipeline integrating content-defined chunking (CDC), SHA-256–based hashing, and LZW (Lempel-Ziv-Welch) compression.
- Enhanced cryptographic hashing performance through ARM NEON SIMD intrinsics and implemented a FPGA-based hardware accelerator for LZW.
- Demonstrated system-level performance exceeding 800 Mb/s throughput with a 0.65 compression ratio on Ultra96-V2 platform.

**Pipelined RISC-V Processor** Jan 2024 – May 2025

- Developed a custom 32-bit RISC-V core using SystemVerilog with a fully pipelined datapath, incorporating multicycle operators, direct-mapped instruction and data caches, and AXI4-Lite protocol for streamlined memory communication.
- Synthesized using the Yosys toolchain and deployed on Lattice ECP5 FPGA, achieving a 31MHz maximum clock frequency with resource utilization of 30.9% LUTs and 4.1% flip-flops.

**Fast, Compact and Efficient DNN via Pruning and Sparse Matrix Compression** Oct 2024 - Dec 2024

- Explored various pruning strategies (global, channel-wise, hard pruning), combined with quantization, to reduce model size while maintaining accuracy and accelerating inference.
- Developed custom sparse linear layer leveraging Compressed Sparse Row (CSR) format for efficient storage and inference.
- Achieved a 1.52× speedup in the most pruned layer and reduced the model size by 43% on VGG16 architecture.

**Configurable Logic Block (CLB) Design and Optimization** Oct 2024 - Dec 2024

- Designed and verified a transistor-level 16-bit CLB circuit using 45nm Salicide CMOS technology in the Cadence environment.
- Performed transistor sizing, mitigated timing hazards, and optimized the circuit for minimal delay and improved energy efficiency.
- Achieved a maximum operating frequency of 1 GHz and an average power consumption of 134.9 µW.

## SKILLS & INTERESTS

**Programming:** C/C++, Python, Verilog, Assembly (x86, RISC-V), CUDA, MATLAB, Shell
**Hardware Platforms:** Avnet Ultra96-V2, Xilinx PYNQ-Z2, Xilinx ZCU104, ULX3S
**Hardware Design & EDA Tools:** Vitis HLS, Vivado Design Suite, Cadence Toolchain
**Software & Systems Tools:** Linux, Git, Docker, CUDA Toolkit
**ML & Computing Frameworks:** PyTorch, TensorFlow, TensorRT, JAX, OpenCV, OpenCL
**Research Interests:** ML Systems, Reconfigurable Computing, GPU Computing, Computer Architecture

## HONORS & ACHIEVEMENTS

**Vingroup Science & Technology Scholarship**, Full funding for Master's study at the University of Pennsylvania. 2024
**Excel Award for Exceptional Capability**, VinUniversity 2023
**Dean's List**, VinUniversity 2021 – 2024
**Undergraduate Merit-based Full-Tuition Scholarship**, VinUniversity 2021
**First Prize**, Vietnamese National Physics Competition for High School Students 2020
**Gold Medal**, Physics Competition for Specialized Students in the Northern Delta and Coastal Areas in Vietnam 2019