
EXPERIENCE

• Sony

Zurich, Switzerland

*Senior AI Research Engineer**Aug 2023 - Present*

- **Sparse Attention for Vision Transformer:** Overhauled FlashAttention-2's per-block attention in PyTorch with sparse masking, achieving 50% FLOP reduction during inference and higher accuracy.
- **Sparse Data Embedding:** Innovated an asynchronous PointNet-based embedding to transform spatio-temporal data into dense tensor representations, enabling seamless Transformer model integration.
- **Research Event-Based SLAM:** Advanced ego-motion compensation techniques for event-based data to optimize SLAM pipeline performance.
- **Model Conversion Tooling:** Developed a tool for efficient PyTorch to ONNX model conversion, streamlining model deployment processes.
- **Energy-Latency Estimator:** Created a tool using ZigZag software for predicting energy consumption and latency of AI models on diverse hardware platforms.
- **SW/HW Co-Design Automation with Neural Architecture Search:** Built an AI-driven, hardware-aware neural architecture search framework. Enabled reducing a model by TODO.
- **Technical Supervision:** Mentored interns in AI-based visual odometry and cost-effective data embedding for event-based cameras.
- **Scrum Master:** Managed project initiatives as a repository maintainer and Scrum Master, ensuring efficient and streamlined project progress.
- **Communication:** Facilitated global knowledge sharing through patents, papers, thesis contributions, speaking engagements, and extensive internal presentations.
- **AI Hardware Survey:** Conducted in-depth research and developed a tool for estimating energy and latency of hardware accelerators for attention-based networks.

*AI Research Engineer**June 2022 - Aug 2023*

- **TODO:** TODO

• STMicroelectronics

Grenoble, France

*Hardware Design Engineer**Apr 2021 - May 2023*

- **CPU:** Innovatively built the Trace and Debug subsystem from scratch for an ARM's Armv9-A SoC. Developed a python library that automates the components instantiation and assembly.
- **Reset and Clock Controller:** Crafted an RTL generator to automate the generation the reset and clock controller for a multi-clock domain STM32 MPU SoC with over 300 peripheral resets and clocks.
- **CPU Benchmarking:** Conducted extensive benchmarking of an ARM cortex on a multi-core MPU SoC from C-coded CoreMark benchmark. Demonstrated significant impacts of compiler choices on system performance.
- **Training:** Gained specialized training on the latest ARM IPs, focusing on V9 architecture and AMBA5 protocols, enhancing technical proficiency in cutting-edge CPU architecture.

• CEA LETI

Grenoble, France

*Doctoral Researcher on AI and Hardware Design**Apr 2018 - Apr 2021*

- **Neuromorphic Hardware Survey:** Conducted a comprehensive bibliographic study on scalable, multi-chip, distributed neuromorphic hardware, leading to a widely cited survey in ACM JETC.
- **HW: ULP IC Design for Near-Sensor Sparse CNN:** Developed an ultra-low power neuromorphic accelerator in 28nm FDSOI, achieving record-breaking energy efficiency and a minimal footprint for integration with EB pixel grids.
- **SW: Moving Object Detection:** Engineered an innovative Event-Based (EB) VIO/SLAM pipeline with integrated ego-motion compensation, adaptable to standard VIO/SLAM pipelines without requiring conventional frames. Devised a novel method for object detection using reconstructed ego-motion compensated EB frames, leading to a patent and enhancing perception accuracy.
- **Simulation and Dataset Generation:** Created two specialized datasets for in-house use, one virtually using Blender and another with a CeleX-V imager, enhancing data accuracy and relevance.
- **Communication:** TODO

• IBM Research

Yorktown Heights, NY, USA

*Hardware Engineer (Intern)**Feb 2017 - Aug 2017*

- **NVM NPU Characterization Automation:** Spearheaded the automation of Phase Change Memory crossbar characterization at wafer-scale, drastically reducing the full characterization time for thousands of crossbars to mere hours.
- **Neuromorphic Application Pioneer:** Contributed to the optimization of PCM device technologies for neuromorphic applications, resulting in patented methods for enhancing conductance linearity.

NOTABLE PUBLICATIONS

Scalable Pitch-Constrained Neural Processing Unit for 3D Integration with Event-Based Imagers	2021
Spiking Neural Network Hardware Implementations and Challenges: A Survey	2019 - 200 cites
On-Chip Trainable 1.4M 6T2R PCM Synaptic Array with 1.6K Stochastic LIF Neurons for Spiking RBM	2019

EDUCATION

- | | |
|---|---|
| <ul style="list-style-type: none"> • Grenoble Alpes University
<i>Ph.D. in Computer Science</i> | Grenoble, France
<i>Apr 2018 – Apr 2021</i> |
| <ul style="list-style-type: none"> • EPFL
<i>Master of Engineering in Electronics (Highest Honors)</i> | Lausanne, Switzerland
<i>Sep 2015 – Sep 2017</i> |
| <ul style="list-style-type: none"> • Grenoble Institute of Technology
<i>Bachelor of Engineering in Electronics</i> | Grenoble, France
<i>Sep 2015 – Sep 2017</i> |