Yifan (Ethan) Yang

LinkedIn ·GitHub ·Google Scholar

Email : yyang29@stanford.edu
Mobile : +1-650-788-9903

Education

- **Stanford University** — Stanford, CA
  B.S. in Electrical Engineering — *Sep 2021 – Jun 2024*
  - **GPA**: 3.9/4.0
  - **Selected Coursework**: Operating Systems, Compilers, Computer Architecture, Parallel Computing, Machine Learning, Computer Security, Blockchain, Digital System Design, VLSI, Web Development.

Publications & Preprints

- **Early Termination for Hyperdimensional Computing Using Inferential Statistics (OMEN)** 2025
  Pu (Luke) Yi, Yifan Yang, Chae Young Lee, Sara Achour. *ASPLOS.*
- **Exchangeability in Neural Network Architectures and its Application to Dynamic Pruning** 2025
  Pu (Luke) Yi, Tianlang Chen, Yifan Yang, Sara Achour. *ICML 2025 Workshop (ES-FoMo III); NeurIPS submission under review.*
- **IoT System for Collecting Vital Signs and Geographic Location Data of Mobile Users** 2020
  Yifan Yang, Yujie Wang, Yangkai Lin, Liye Jia. *2020 International Conference on Communications, Information System and Computer Engineering (CISCE).*

Research Experience

- **Stanford University** — Stanford, CA
  Research Assistant (Advisor: Prof. Sara Achour) — *Sep 2022 – Present*
  - **OMEN**: Co-developed OMEN, reframing HDC inference with inferential statistics to reduce computation while preserving accuracy guarantees.
  - **GPU parallelism**: Re-engineered the experimental stack to leverage GPU parallelism, shrinking per-experiment turnaround from days to ¡ 1 minute, enabling richer ablations and broader dataset coverage.
  - **Artifacts & reproducibility**: Led artifact engineering (one-command runners; fixed seeds; data/log parsers -¿ auto-generated LaTeX tables) and helped secure artifact-quality recognition.
  - **LearningHD**: Proposed and implemented learned encoders and learned class hypervectors to densify representations and improve downstream accuracy versus random hypervectors.
  - **On-device pipeline**: Built a microcontroller-based on-device evaluation pipeline (precise timing & memory accounting) to support edge-intelligence claims and documentation.
- **Stanford University** — Stanford, CA
  Post-OMEN Research — Dynamic Pruning via Exchangeability (with Prof. Sara Achour) — *2024 – 2025*
  - **Exchangeability**: Formalized exchangeability in NN parameters/activations, showing exploitable redundancy at inference time; derived ExPrune, a per-input dynamic pruning rule.
  - **Evaluation harness**: Built an evaluation harness over CNN/Transformer backbones with controlled sparsity–accuracy–latency ablations; standardized early-exit vs. dynamic-pruning baselines for fair comparisons.
  - **Reproducibility**: Refactored the codebase for reproducibility (scripts, seed discipline, reporting) and prepared public artifacts/experiment sheets.
- **Stanford University** — Stanford, CA
  Exploratory Project — Music Note Prediction (with Prof. Sara Achour) — *2023 – 2024*
  - **Task formulation**: Formulated symbolic music note prediction as a time-series task; benchmarked HDC encoders vs. RNN/LSTM & lightweight Transformers under strict memory/latency constraints.
  - **Pipeline**: Implemented dataset curation (MIDI -¿ event sequences), on-device profiling, and anytime/early-halt evaluators; recorded negative results and introduced a hybrid encoder with stabilized voting.
  - **Outcomes**: Clarified regimes where HDC wins (short horizon, noisy labels, tight memory) and where learned encoders dominate; distilled insights into follow-up ablations for ExPrune/OMEN.
- **Stanford University (CS107E)** — Stanford, CA
  RISC-V Course Migration & Infrastructure — *Aug 2023 – Jun 2024; ongoing support through Aug 2025*

- **Course migration**: Migrated the course from ARM/Raspberry Pi to RISC-V boards: rebuilt the teaching codebase (display pipeline, PS/2 keyboard, build/flash workflows), studied schematics/ISA, and recreated peripheral drivers (I^2C/SPI/SD, etc.).
- **Toolchain enablement**: Researched and enabled RVV 0.7.1 and FPU in the MangoPi toolchain; authored activation guides & intrinsics notes integrated into the official project guide.
- **Support & maintenance**: Provided ongoing office hours/debugging support post-graduation and maintained issues/patches with course staff.

Industry & Engineering Experience

- **NVIDIA** Santa Clara, CA
  Deep Learning Infrastructure Engineer *Full-time: Jun 2024 – Present; Intern: Jun–Sep 2022, Jun–Sep 2023*
  - **Platform ownership**: Project owner/architect & primary implementer of a distributed, asynchronous log analytics platform for silicon verification teams (design → APIs → implementation → rollout).
  - **Capabilities**: Real-time tail & regex, multi-line parsing/structured extraction, error detection & alerting, precise stack-trace mapping, free-text search with millisecond-level latency, and elastic autoscaling at TBs/day scale.
  - **Latency**: Typical tail-to-query delay at our scale was minute-level with off-the-shelf engines; our design achieves millisecond-level end-to-end query latency under production load (measured internally).
  - **Additional contributions**: Containerized & migrated chip-verification services to Kubernetes; deployed Redis HA; built an OpenTelemetry observability pipeline & an internal tracing library; designed and implemented Tasker, a job scheduling/management platform for chiplet workflows.
- **Si-Tech (Urumqi)** Urumqi, China
  Software Engineer Intern *Jan 2021 – Jun 2021*
  - **AI work-order system**: Built an AI work-order system (Java/Web + NLP), reducing human triage by ¿ 90%; modularized backend to decouple subsystems; automated workflows to cut processing time by 80%.
- **China Unicom (Urumqi)** Urumqi, China
  O&M Engineer Intern *Jan 2020 – Mar 2020*

Selected Projects

- **EE374 Blockchain Node & Mining Pool (Team Lead)**: Implemented a full blockchain node and a pool coordinator; developed a cross-platform miner in C++/OpenCV+CUDA using CPU+GPU. When a fork/DoS-like incident stalled the class chain, open-sourced the miner & pool to aggregate peers' compute and restore chain liveness. (Spring 2022)
- **CS229 → CS194W: Music Genre / Time-series & HDC**: Explored audio time-series; HDC surpassed RNN/LSTM baselines in our setting. Productized the pipeline and applied OMEN-derived ideas to improve accuracy while remaining mobile-friendly.
- **CS224W: Graph Neural Networks & HDC**: Explored GNN structures and how they capture/represent graph structures. Explored expressivity of HDC representations vs. GNNs across basic tasks. (Fall 2023)
- **PiAuto (Open-source)**: Turned iPads into portable head units via Raspberry Pi: iOS client + on-board server (12V power), Wi-Fi AP, AirPlay audio to car speakers, and OBD integration. **View Project**
- **IoT System for Pandemic Control (with Prof. Fouad Tobagi)**: IoT research project designed as a system prototype for pandemic control. Initiated in Jan 2020 at the emergence of COVID-19. Led a 4-person team in developing the system for collecting vital signs and geographic location data of mobile users. Published at IEEE CISCE 2020.

Teaching & Service

- **CS107E contributions**: Contributed write-ups to the CS107E project guide (RISC-V RVV/FPU activation, intrinsics usage, peripheral drivers & lab materials); continued office hours & debugging support after graduation; contribution spanned over a year.

Research Interests

- **Low-power ML for edge devices**: Hardware–software co-design for on-device intelligence.
- **Hyperdimensional Computing (HDC)**: Statistical early termination, anytime/early-halt inference, learned encoders & class hypervectors.
- **Dynamic pruning & efficient inference**: Dynamic pruning/elastic inference for neural networks; time-series/audio ML under tight latency & memory budgets.

Skills

- **Programming**: C/C++, Python, Go, JavaScript/TypeScript, CUDA, Verilog/RTL
- **Systems & Infra**: RISC-V/ARM bare-metal, RVV, GPU optimization, Linux, Kubernetes, Docker, OpenTelemetry, CI/CD
- **ML**: on-device benchmarking, PyTorch, CNN, GNN, RNN
- **Tools**: LaTeX, Git