Yan Ru Pei

yanrpei@gmail.com | https://github.com/PeaBrane

Research Interests

spin glass, artificial intelligence, quantum computing, neuromorphic computing, cryptography

Academic Background

University of California, San Diego — Advisor: *Massimiliano Di Ventra*
PhD in Physics (Computational Science Specialization) — *Sept 2017 – Sept 2021*

University of California, Los Angeles — Advisor: *Robert Cousins*
BS in Physics and Applied Mathematics (summa cum laude) — *Sept 2014 – Sept 2016*

Publications/Preprints (Google Scholar)

Artificial Intelligence:

- Tiyasa Mitra, Ritika Borkar, Nidhi Bhatia, Shivam Raj, Hongkuan Zhou, **Yan Ru Pei**, et al. (2026). Beyond the Buzz: A Pragmatic Take on Inference Disaggregation. *Conference on Machine Learning and Systems (MLSys)*.
- **Yan Ru Pei**, Olivier Coenen. (2025). PLEIADES: Building Temporal Kernels with Orthogonal Polynomials. *NeurIPS 2025*.
- **Yan Ru Pei**, Ritik Shrivastava, FNU Sidharth. (2025). Real-time Speech Enhancement on Raw Signals with Deep state-space Modeling. *Interspeech 2025*.
- **Yan Ru Pei**. (2025). Let SSMs be ConvNets: State-space Modeling with Optimal Tensor Contractions. *ICLR spotlight*.
- **Yan Ru Pei**, Sasskia Brüers, Sébastien Crouzet, Douglas McLelland, Olivier Coenen. (2024). A Lightweight Spatiotemporal Network for Online Eye Tracking with Event Camera. *CVPR*.
- Zuowen Wang, Chang Gao, Zongwei Wu, Marcos V. Conde, Radu Timofte, Shih-Chii Liu, Qinyu Chen, Zheng-jun Zha, Wei Zhai, Han Han, Bohao Liao, Yuliang Wu, Zengyu Wan, Zhong Wang, Yang Cao, Ganchao Tan, Jinze Chen, **Yan Ru Pei**, et al. (2024). Event-Based Eye Tracking. AIS 2024 Challenge Survey. *CVPR*.
- Haik Manukian*, **Yan Ru Pei***, Sean Bearden, M. Di Ventra. (2020). Mode-assisted unsupervised learning of restricted Boltzmann machines. *Nature Communication Physics*.
- **Yan Ru Pei**, Haik Manukian, M. Di Ventra. (2020). Generating Weighted MAX-2-SAT Instances with Frustrated Loops: an RBM Case Study. *Journal of Machine Learning Research*.

Physics:

- **Yan Ru Pei**, M. Di Ventra. (2022). A Finite-temperature Phase Transition for the Ising Spin-glass in $d \geq 2$. *preprint*.
- **Yan Ru Pei**, M. Di Ventra. (2022). Non-equilibrium criticality and efficient exploration of glassy landscapes with memory dynamics. *Physica A*.

Other:

- Yuki Wang, **Yan Ru Pei**. (2019). The Optimal Deterrence of Crime: A Focus on the Time Preference of DWI Offenders. *preprint*.
- **Yan Ru Pei**, Fabio L. Traversa, Massimiliano Di Ventra. (2019). On the Universality of Memcomputing Machines. *IEEE TNNLS*

Patents

Accepted:

- Methods and System for Improved Processing of Sequential Data in a Neural Network
- Method and system for implementing temporal convolution in spatiotemporal neural networks

Submitted:

- System of Interconnected State Space Models with Jointly Driven State Matrices
- System And Method For Efficient Execution of Large Generative Artificial Intelligence Models on Edge Devices Using State-Space Models
- Real-time Speech Enhancement on Raw Signals with Deep State-space Modeling
- Method and System for Implementing Encoder Projection in Neural Network

Research Experience

NVIDIA Santa Clara, CA

Senior Deep Learning Algorithm Engineer *Feb 2025 – present*

- **Distributed inference at scale – Dynamo**
  - Architected a distributed router for large-scale LLM inference, applying computational complexity principles to optimize scheduling and caching. Implemented in Rust, achieving a 3× reduction in time-to-first-token and throughput latency. Results were presented at NVIDIA GTC 2025.
  - Designed and deployed a mock engine emulating vLLM's block manager and scheduler, enabling end-to-end CI pipelines without GPUs. This framework also supported large-scale testing, debugging, and performance tuning in GPU-free environments for the Dynamo repository.
  - Developed a graph-theory–based data synthesizer to generate workloads with controlled prefix hit rates. Enhanced load testing by producing realistic traffic patterns that stress engines and pipelines under varying cache utilization scenarios.

BrainChip, Inc. (neuromorphic AI startup) Laguna Hills, CA

Senior Machine Learning Scientist *Oct 2021 – Feb 2025*

- **Developed a SOTA spatiotemporal network for computer vision at the edge [Paper]**
  - Designed a spatiotemporal neural network interfacing with RGB and event cameras for efficient online object detection, object tracking, and monocular depth estimation.
  - Implemented complex data pipelines (pre-processing event signals) and wrote vectorized routines for object detection in functorch for evaluating the network, eliminating all CPU bottlenecks.
  - Applied the network on the Prophesee event-camera road-scene dataset, and achieved a 30% increase in mAP over the previous SOTA network (ConvLSTM) with **100x fewer** parameters and compute.
  - Combined the network with GNN and applied it for open-loop and closed-loop motion planning on the NuPlan dataset, reaching near first-place solutions on the leaderboard.
  - Prepared four papers and three patents for an in-house hardware implementation of the network (Akida 2.0).
- **Developed a deep state-space model for sequence prediction**
  - Designed a network that can be dual-configured as convolution and recurrent networks, with **a million times less parameters and compute** than transformer models for NLP tasks. Integrated concepts from modern SSMs (e.g. H3, hyena, mamba) to build an LLM model for the edge.
  - Collaborated with SW engineers to train, quantize, and sparsify this network down to fixed-point 8-bit and 90% sparsity, and supported HW engineers with its RTL simulation.
  - Wrote custom fused kernels with Triton to eliminate memory bottlenecks for training the model on Nvidia Ampere GPUs, yielding a training speed up of $\sim 20\%$.
  - Coordinated multi-gpu training runs on the cloud using AWS EC2 with 8 A100 instances.

- Achieved near-SOTA results on audio denoising, vital signs prediction, speech recognition, and Wikitext-103.

Graduate Student Researcher Sept 2017 – Present
UCSD *San Diego, CA*

- Developed continuous dynamical approach to constrained optimization and simulation of spin glasses. Applied the approach for solving boolean satisfiability problems, encompassing problems such as prime factorization (a basis for RSA encryption).
- Developed a new pre-training method for RBMs based on modal sampling, significantly improving the stability and efficiency of the training process.
- Studied the possibility of applying machine learning techniques to quantum many-body systems. Drew a connection between neural networks and quantum computing benchmarks.
- Explored a mathematical approach for describing general computing architectures.

Reesarch Support Associate Sept 2016 – Sept 2017
MIT *Boston, MA*

- Worked on a modular component for high precision magnetic field control for Dysprosium MOT chambers (under supervision of Nobel laureate Wolfgang Ketterle).

Undergraduate Research Assistant December 2015 – Sept 2016
UCLA *Los Angeles, CA*

- Analyzed the rigor of the method of data unfolding in high energy experiments in a Bayesian context (under supervision of Robert Cousins).
- Designed and simulated a voltage array for collimating ion beams (under the supervision of Eric Hudson).

Skills

Theory: AI training and inference, quantum/neuromorphic computing, stochastic/dynamical systems
Edge Computing: event-based algorithms, network quantization, online spatiotemporal inference
Libraries: PyTorch, TF, JAX, Hydra, wandb, scikit-learn, Numba, pandas, OpenCV
Engineering: signal processing, audio denoising, SLAM, circuit design, laser optics

Invited Talks

IROS EdgeAI4R Workshop, 2025

Conference

The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS), 2025
International Conference on Learning Representations (ICLR), 2025
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024
APS Physics March Meeting, Nashville, 2021
APS Physics March Meeting, Denver (Virtual), 2020
Harvard-MIT CUA Winter Retreat, 2017

Refereed Journals/Conferences

AAAI 2026 Workshop XAI4Science
ACM TheWebConf 2026
Applied Intelligence
CVPR 2025 Workshop on Event-based Vision, **Program Committee**
Environmental Research Letters

The 28th European Conference on Artificial Intelligence (ECAI 2025), **Program Committee**
Graph Signal Processing Workshop (GSP)
IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)
IEEE Symposium on Computers & Informatics (ISCI) 2025
IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI)
IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2025
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2026, **Area Chair**
IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA) 2025
IEEE World Congress on Computational Intelligence 2026 (WCCI 2026), **Area Chair**
International Conference on Learning Representations (ICLR)
ICLR 2025 Workshops: AI4CHL, DeLTa, FM-Wild, XAI4Science
The 34th International Joint Conference on Artificial Intelligence (IJCAI)
The 35th International Joint Conference on Artificial Intelligence and the 29th European Conference on Artificial Intelligence (IJCAI-ECAI 2026)
International Joint Conference on Neural Networks (IJCNN), **Area Chair**
International Journal of Circuit Theory and Applications
International Journal of Image and Data Fusion
International Symposium on Circuits and Systems (ISCAS) 2026
Interspeech 2025
Journal of Environmental Engineering and Landscape Management (JEELM)
KDD 2025 August ADS Track, **Outstanding Reviewer**
KDD 2026 August ADS Track, **Senior Area Chair**
Land Degradation & Development
Machine Learning: Science and Technology
NeurIPS 2025
NeurIPS 2025 Workshops: COLM
Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2026
PLOS ONE
Remote Sensing Letters
SciPy 2025 Proceedings
Society & Natural Resources
International Conference of Students of Systematic Musicology (SysMus)
The Journal of Physical Chemistry
Transactions on Machine Learning Research (TMLR)

Graduate Level Teaching Experience

Spring 2019: UCSD Phyiscs 212C - Quantum Mechanics III
Winter 2019: UCSD Phyiscs 200B - Theoretical Mechanics II
Fall 2018: UCSD Phyiscs 243 - Stochastic Methods