Joshua Fan

Email: jyf6@cornell.edu – **Website**: http://joshuafan.github.io

EDUCATION

Ph.D. Computer Science, Cornell University, *Ithaca, NY, USA (in progress)* **2019 – present**
- Advisor: Prof. Carla Gomes
- GPA: 4.10/4.3

(Expected graduation: Dec. 2025)

M.S. Computer Science, University of Washington, *Seattle, WA, USA* **2017 – 2019**
- Advisor: Prof. Sreeram Kannan
- GPA: 3.84/4.0

B.S. Computer Science, University of Washington, *Seattle, WA, USA* **2013 – 2017**
- GPA: 3.97/4.0 *(summa cum laude)*

RESEARCH EXPERIENCE

Research Assistant, Cornell University

Institute for Computational Sustainability *(advised by Prof. Carla Gomes)* **Aug. 2019 – present**
- Integrated **scientific models** into **deep learning** in an **interpretable** way to simulate the soil carbon cycle. Our approaches (ScIReN/BINN) fully respect scientific laws, infer *unlabeled* physical process parameters that govern the soil carbon cycle, and learn transparent relationships between these physical processes and input variables.
- Designed novel **contrastive learning** methods for **satellite images** to map fish farms in the Amazon. Used masked pooling and multiple embedding spaces to overcome difficult **distribution shifts** that threw off standard models.
- Created an **interpretable spatiotemporal deep learning** framework (GNN-RNN) to forecast crop yields in US counties from weather/soil data. Using only public data, we achieve accuracy comparable to USDA's expensive manual surveys.
- Proposed novel **Transformer** variants which emphasize locality and relative positioning, surpassing SOTA on many **time series** regression tasks.
- Developed Coarsely-Supervised Smooth U-Net, a **weakly-supervised learning** method which can predict at extremely **fine resolutions** (30m) even when training labels are only available at a much coarser resolution (3km). We use this to predict SIF (a measure of vegetation productivity) at unprecedented fine resolutions.

Research Assistant, University of Washington

Information Theory Lab *(advised by Prof. Sreeram Kannan)* **Mar. 2017 – Jun. 2018**
- Developed scalable algorithms inspired by Latent Dirichlet Allocation and matrix factorization to discover cell types and find structure in large single-cell RNA-seq datasets (over 1 million cells) (Poster, Paper, Code)

Computing for Development Lab *(advised by Prof. Richard Anderson)* **Mar. 2015 – Jun. 2016**
- Redesigned a survey app which helps health workers collect patient data and suggest follow-up actions

PEER-REVIEWED PUBLICATIONS *(* denotes equal contribution)*

Monitoring Vegetation from Space at Extremely Fine Resolutions via Coarsely-Supervised Smooth U-Net.
Joshua Fan, Di Chen, Jiaming Wen, Ying Sun, Carla Gomes.
IJCAI-22: *International Joint Conference on Artificial Intelligence (AI for Good track),* 2022.
- *Also presented at the "Tackling Climate Change with Machine Learning" workshop at NeurIPS 2021*

A GNN-RNN Approach for Harnessing Geospatial and Temporal Information: Application to Crop Yield Prediction.
Joshua Fan*, Junwen Bai*, Zhiyun Li*, Ariel Ortiz-Bobea, Carla Gomes.
AAAI-22: *AAAI Conference on Artificial Intelligence (AI for Social Impact track),* 2022.
- *Workshop version got **best paper award (ML Innovation)** at "Tackling Climate Change with ML" workshop, NeurIPS 2021.*

Towards Sustainable Aquaculture in the Amazon.
Felipe S. Pacheco, Sebastian A. Heilpern, Claire DiLeo, Rafael M. Almeida, Suresh A. Sethi, Marcela Miranda…, **Joshua Fan**, … , Carla P. Gomes, & Alexander S. Flecker. *Nature Sustainability*, 2025.

Scalable preprocessing for sparse scRNA-seq data exploiting prior knowledge.
Sumit Mukherjee, Yue Zhang, **Joshua Fan**, Georg Seelig, Sreeram Kannan. *Bioinformatics,* 2018.

PREPRINTS UNDER REVIEW

Scientifically-Interpretable Reasoning Network (ScIReN): Uncovering the Black-Box of Nature. *Submitted to NeurIPS 2025.*
Joshua Fan*, Haodi Xu*, Feng Tao*, Md Nasim, Marc Grimson, Yiqi Luo, Carla P. Gomes.

Biogeochemistry-Informed Neural Network (BINN) for Improving Accuracy of Model Prediction and Scientific Understanding of Soil Organic Carbon. *Submitted to Geoscientific Model Development (GMD).*
Haodi Xu*, **Joshua Fan***, Feng Tao*, Lifen Jiang, Fengqi You, Benjamin Z. Houlton, Ying Sun, Carla P. Gomes, Yiqi Luo.

WORKSHOP PAPERS

Detecting Aquaculture with Deep Learning in a Low-Data Setting.
L. Greenstreet, **J. Fan**, F. Pacheco, Y. Bai, M. Ummus, C. Doria, N. Barros, B. Forsberg, X. Xu, A. Flecker, C. Gomes.
Fragile Earth workshop at KDD 2023. Poster presented at AGU (American Geophysical Union) Annual Meeting, Dec 2023.

SELECTED POSTERS & TALKS

Scientifically-Interpretable Reasoning Network (ScIReN) for Understanding the Soil Carbon Cycle.
Joshua Fan*, Haodi Xu*, Feng Tao, Md Nasim, Yiqi Luo, Carla Gomes.
Poster/talk at AI-LEAF Institute Annual Meeting, Fort Collins, CO, USA. May 2025.

Using Deep Learning to Monitor and Forecast Vegetation Growth.
Talk at Soil and Crop Sciences Seminar, Cornell University, Ithaca, NY, USA. Feb 2022.

INDUSTRY EXPERIENCE

Research & Development Intern, Kitware *(Computer Vision team)* **Summer 2024**
- Researched techniques to mitigate loss of plasticity in deep continual learning – to allow pretrained models to learn incrementally from new datasets and tasks

NLP Research Intern, Docugami *(AI Document Engineering startup)* **Summer 2018, Summer 2019**
- Researched and implemented state-of-the-art NLP algorithms (including topic modeling, clustering, and question-answering techniques), and adapted them in novel ways for enterprise document analysis

Software Engineer Intern, Meta *(Integrity Computer Vision Team)* **Fall 2018**
- Trained a clip-based 3D CNN to detect graphic and violent content in videos, outperforming previous approaches

Software Engineer Intern, Meta *(Search, Whole Page Ranking Team)* **Fall 2017**
- Trained sequence neural networks to predict search behavior based on recent query history; improved user click rate

Software Engineer Intern, Meta *(Search Indexing Team)* **Summer 2016**
- Built a web tool to help engineers debug and simulate changes to the search indexing pipeline

Software Design Engineer Intern, BitTitan **Summer 2015**
- Built infrastructure to allow mailbox migrations to be tested in memory; optimized SQL queries

TEACHING EXPERIENCE

- **TA, Application of Machine Learning to Plant Science** (Cornell University, PLSCI 7202): Fall 2022
- **TA, Intro to Artificial Intelligence** (Cornell University, CS 4700): Fall 2019
- **TA, Probability & Statistics** (UW, CSE 312): Fall 2015, Winter 2016, Spring 2017, Winter 2018, Winter 2019
- **TA, Foundations of Computing I/Discrete Math** (Univ of Washington, CSE 311): Fall 2016, Spring 2018
- **TA, Intro to Machine Learning for Non-Majors** (Univ of Washington, CSE 416): Spring 2019

ADVISING

Kaitlyn Chen (BS Cornell 2023: next position: Amazon)
- Topic: Improved Transformer architecture with inductive biases for continuous time-series regression.

Saraswathy Amjith (high school, 1st place in Earth & Env. Sciences at WA State Science Fair – next position: MIT)
- Topic: Detecting deforestation using deep learning with radar and optical satellite imagery

SERVICE

Reviewer for:
- *AAAI:* 2024, 2025
- *NeurIPS Computational Sustainability workshop*: 2023
- *Environmental Research Letters* (2 journal papers): 2022-2023

Organizer for Cornell Computer Science Visit Day, 2020

TECHNICAL SKILLS

- **Significant experience:** Python, Java, C#, SQL, C++, PHP/Hack
- **Working knowledge:** R, Matlab, HTML/CSS, JavaScript
- **Libraries/tools:** PyTorch, DGL, Pandas, Matplotlib, Eclipse, Git, Visual Studio, Linux, Nuclide

HONORS & AWARDS

- **NSF Research Training (NRT) Fellowship**: Digital Plant Science (2021-2024), AI for Sustainability (2025-2026)
- **Bob Bandes Memorial Excellence in Teaching Award**, University of Washington, 2019 (one of 3 winners)