Joshua Fan

jyf6@cornell.edu | 425-628-7677 | http://joshuafan.github.io | Work authorization: US citizen

SUMMARY

PhD candidate (Cornell University, Computer Science) with 7+ yrs experience leading applied ML projects – data engineering, understanding customer needs, research, model design, training, evaluation. Open to ML Scientist or Engineer roles where I can use my AI/ML expertise for real-world impact. Available to start in **Jan. 2026**. Prefer remote but open to relocation.

EDUCATION

PhD Computer Science, Cornell University – *Ithaca, NY* **Aug 2019 – Dec 2025 (expected)**

- Thesis: Spatio-Temporal Deep Learning with Limited Labels: Applications in Computational Sustainability
- Advisor: Prof. Carla Gomes GPA: 4.10/4.3

MS Computer Science, University of Washington – *Seattle, WA* **Mar 2017 – June 2019**

- Advisor: Prof. Sreeram Kannan GPA: 3.84/4.0

BS Computer Science, University of Washington – *Seattle, WA* **Sep 2013 – Mar 2017**

- Advisor: Prof. Richard Anderson GPA: 3.97/4.0 (summa cum laude)

RESEARCH EXPERIENCE

Research Assistant, Cornell University (Institute for Computational Sustainability) – *Ithaca, NY* **Aug 2019 – Present**

- Integrated **scientific models** into **deep learning** in an **interpretable** way to simulate the soil carbon cycle. Our approaches (ScIReN/BINN) respect scientific laws, infer *unlabeled* biogeochemical processes, and help scientists gain insight into how input variables influence biogeochemical processes [7][8].
- Designed novel **self-supervised (contrastive) learning** methods for **satellite images** to map fish farms in the Amazon. Used masked pooling and custom augmentations to overcome difficult **distribution shifts**, increasing F1 by 20% [5].
- Created an **interpretable spatiotemporal deep learning** framework (GNN-RNN) to forecast crop yields in US counties [2]. With only public weather/soil data, we achieve accuracy comparable to USDA's expensive on-the-ground forecasts.
- Developed novel **Transformer** variants which emphasize locality and relative positioning, reducing prediction error (RMSE) by an average of 18% on multiple **time series regression** tasks over the previous state-of-the-art [6].
- Devised a **weakly-supervised U-Net** 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 (vegetation productivity) from **remote sensing** data at unprecedented fine resolutions, reducing RMSE by 9% over the strongest baselines [1].

Research Assistant, University of Washington (Information Theory Lab) – *Seattle, WA* **Mar 2017 – June 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)

Research Assistant, University of Washington (Computing for Development Lab) – *Seattle, WA* **Mar 2015 – June 2016**

- Redesigned a survey app which helps health workers collect patient data and suggest follow-up actions

INDUSTRY EXPERIENCE

Research & Development Intern, Kitware *(Computer Vision team) – Clifton Park, NY* **June 2024 – Aug 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) – Kirkland, WA* **Mar 2018 – Sept 2018, Jan 2019 – Aug 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) – Seattle, WA* **Sept 2018 – Dec 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) – Menlo Park, CA* **Sept 2017 – Dec 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) – Seattle, WA* **June 2016 – Sept 2016**

- Built a web tool to help engineers debug and simulate changes to the search indexing pipeline

Software Design Engineer Intern, BitTitan – Bellevue, WA **June 2015 – Sept 2015**

- Built infrastructure to test mailbox migrations in memory; optimized SQL queries (reducing runtime by 60%)

PUBLICATIONS (peer-reviewed conference and journal papers; * denotes equal contribution)

[1] 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. **(Acceptance rate: 19.3%)**

- *Also presented at the "Tackling Climate Change with Machine Learning" workshop at NeurIPS 2021*

[2] 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. **(Acceptance rate: 15.0%)**

- *Workshop version got **best paper award (ML Innovation)** at "Tackling Climate Change with ML" workshop, NeurIPS 2021.*

[3] 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.

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

WORKSHOP PAPERS (peer-reviewed)

[5] Detecting Aquaculture with Deep Learning in a Low-Data Setting.
L. Greenstreet, **Joshua 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.*

PUBLICATIONS UNDER REVIEW (* denotes equal contribution; underline denotes mentored students)

[6] Locality and Distance-Aware Attention for Time Series Extrinsic Regression. *Submitted to AAAI 2026.*
Joshua Fan*, Kaitlyn Chen*, Shufeng Kong, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, Carla Gomes.

[7] Scientifically-Interpretable Reasoning Network (ScIReN): Discovering Hidden Relationships in the Carbon Cycle and Beyond. *Submitted to AAAI 2026.*
Joshua Fan*, Haodi Xu*, Feng Tao*, Md Nasim, Marc Grimson, Yiqi Luo, Carla P. Gomes.

[8] Biogeochemistry-Informed Neural Network (BINN) for Improving Accuracy of Model Prediction and Scientific Understanding of Soil Organic Carbon. *Submitted to Geoscientific Model Development (GMD), 2025.*
Haodi Xu*, **Joshua Fan***, Feng Tao*, Lifen Jiang, Fengqi You, Benjamin Z. Houlton, Ying Sun, Carla P. Gomes, Yiqi Luo.

[9] AISciVision: A Framework for Specializing Large Multimodal Models in Scientific Image Classification. *Arxiv, 2024.*
B. Hogan, A. Kabra, F. Pacheco, L. Greenstreet, **Joshua Fan**, A. Ferber, M. Ummus, A. Brito, … , A. Flecker, C. Gomes.

SELECTED TALKS

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

TEACHING EXPERIENCE

- **TA, Cornell:** Application of Machine Learning to Plant Science (Fall 2022), Intro to Artificial Intelligence (Fall 2019)
- **TA, University of Washington:** Intro to Machine Learning (Spring 2019), Probability & Statistics (Fall 2015, Winter 2016, Spring 2017, Winter 2018, Winter 2019), Discrete Math (Fall 2016, Spring 2018)

MENTORSHIP

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

TECHNICAL SKILLS

- **Significant experience:** Python, Java, C#, SQL, C++, PHP/Hack
- **Working knowledge:** R, Matlab, HTML/CSS, JavaScript, Bash
- **Tools:** PyTorch, DGL, Pandas, Matplotlib, scikit-learn, rasterio, Git, Visual Studio, xarray, Google Earth Engine, ArcGIS

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)