Ruth Fong

✉ ruthfong@princeton.edu
www.ruthfong.com
Last updated: Dec 2022

Current Appointment

2021 – now **Lecturer**, *Princeton University*, Department of Computer Science
- COS324: Introduction to Machine Learning (Fall 2022 & Spring 2023)
- COS126: Computer Science: An Interdisciplinary Approach (Fall 2021 & Spring 2022)

Education

2016–2020 **D.Phil. Engineering Science**, University of Oxford
Advisor: Professor Andrea Vedaldi
Examiners: Professors Antonio Torralba (MIT) and Andrew Zisserman (Oxford)
Thesis: Understanding Convolutional Neural Networks

2015-2016 **M.Sc. Neuroscience**, *with distinction*, University of Oxford
Rotation 1 Advisor: Professor Rafal Bogacz
Rotation 2 Advisors: Professor Andy King, Dr. Ben Willmore, and Dr. Nicol Harper
Thesis 1: Optimizing Deep Brain Stimulation to Dampen Tremor
Thesis 2: Modeling Blind Single Channel Sound Separation Using Predictive Neural Networks

2011-2015 **A.B. Computer Science**, *magna cum laude with Highest Honors*, Harvard University
Advisors: Professors David Cox and Walter Scheirer
Thesis: Leveraging Human Brain Activity to Improve Object Classification

Teaching Experience

2020 – 2021 **Teacher**, *Timothy Christian School*
- High School Courses: Beginner Computer Programming, AP CS A, AP CS Principles
- Middle School Courses: 6th and 8th Grade Computer Science
- FIRST Tech Challenge (FTC) Head Coach
- American Computer Science League (ACSL) Coach

2019 – 2020 **Tutorial Instructor**, *University of Oxford*, Department of Engineering Science
- B14: Information Engineering (3rd year undergrad course on machine learning and computer vision)

Summer 2015 **Course Instructor**, *NJ Governor's School in Engineering and Technology*
- Mathematics in the World

2012, 2014 **Course Assistant & Teaching Fellow**, *Harvard University*, Department of Computer Science
- CS121: Introduction to the Theory of Computation (Fall 2014)
- CS20: Introduction to Discrete Math (Spring 2014)
- CS50: Introduction to Computer Science I (Fall 2012)

Industry Experience

Summer 2019 **Research Collaborator**, *Pro Unlimited @ Facebook*, with Andrea Vedaldi

Summer 2018 **Research Intern**, *Google Research*, with Vitto Ferrari

Summer 2014 **Quantitative Software Engineer Intern**, *D.E. Shaw, Co.*

Summer 2013 **Software Engineer Intern**, *Apple*, with Safari Webkit team

Summer 2012 **Explore Intern**, *Microsoft*, with Windows 8 team

Publications

Preprints

P6 **Interactive Visual Feature Search**
D. Ulrich, R. Fong *arXiv 2022*

P5 **Improving Fine-Grain Segmentation via Interpretable Modifications: A Case Study in Fossil Segmentation**
I. Panigrahi, R. Manzuk, A. Maloof, R. Fong, *arXiv 2022*

P4 **"Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction**
S.S.Y. Kim, E.A. Watkins, O. Russakovsky, R. Fong, A. Monroy-Hernández, *arXiv 2022*

P3 **Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability**
V.V. Ramaswamy, S.S.Y. Kim, R. Fong, O. Russakovsky, *arXiv 2022*

P2 **Gender Artifacts in Visual Datasets**
N Meister*, D. Zhao*, A. Wang, V.V. Ramaswamy, R. Fong, O. Russakovsky, *arXiv 2022*

P1 **ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features**
V.V. Ramaswamy, S.S.Y. Kim, N. Meister, R. Fong, O. Russakovsky, *arXiv 2022*

Refereed Conference Papers

C8 **HIVE: Evaluating the Human Interpretability of Visual Explanations**
S.S.Y. Kim, N. Meister, V.V. Ramaswamy, R. Fong, O. Russakovsky, *ECCV 2022*

C7 **On Compositions of Transformations in Contrastive Self-Supervised Learning**
M. Patrick*, Y.M. Asano*, P. Kuznetsova, R. Fong, J.F. Henriques, and A. Vedaldi, *ICCV 2021*

C6 **Quantifying Learnability and Describability of Visual Concepts Emerging in Representation Learning**
I. Laina, R. Fong, and A. Vedaldi, *NeurIPS 2020*

C5 **Contextual Semantic Interpretability**
D. Marcos, R. Fong, S. Lobry, R. Flamary, N. Courty, and D. Tuia, *ACCV 2020*

C4 **There and Back Again: Revisiting Backpropagation Saliency Methods**
S.-A. Rebuffi*, R. Fong*, X. Ji*, and A. Vedaldi, *CVPR 2020*

C3 **Understanding Deep Networks via Extremal Perturbations and Smooth Masks**, *Oral*
R. Fong*, M. Patrick*, and A. Vedaldi, *ICCV 2019*

C2 **Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks**, *Spotlight*
R. Fong and A. Vedaldi, *CVPR 2018*

C1 **Interpretable Explanations of Black Boxes by Meaningful Perturbation**
R. Fong and A. Vedaldi, *ICCV 2017*

Refereed Workshop Papers

W3 **Interactive Similarity Overlays [link]**
R. Fong, A. Mordvintsev, A. Vedaldi, and C. Olah, *VISxAI 2021*

W2 **Debiasing Convolutional Neural Networks via Meta Orthogonalization**
K. David, Q. Liu, and R. Fong, *NeurIPSW (Workshop on Algorithmic Fairness through the Lens of Causality and Interpretability) 2020*

W1 **Occlusions for Effective Data Augmentation in Image Classification**
R. Fong and A. Vedaldi, *ICCVW (Workshop on Interpreting and Explaining Visual Artificial Intelligence Models) 2019*

Edited Volumes

E1 **XXAI – Beyond Explainable Artificial Intelligence**
A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, editors, *Springer LNAI 2022*

Book Chapters

B1 **Explanations for Attributing Deep Neural Network Predictions**
R. Fong and A. Vedaldi, in *Interpretable AI: Interpreting, Explaining, and Visualizing Deep Learning*, edited by W. Samek, G. Montavon, A. Vedaldi, L. Hansen, and K.-R. Müller, *Springer LNCS 2019*

Technical Reports

T1 **Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims**
M. Brundage, S. Avin,J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Tonor, R. Fong, et al., *arXiv 2020*

Talks

All talks were invited talks unless otherwise noted in parentheses.

Directions in Interpretability

Nov 2022 HEIBRiDS (Berlin) – HEIBRiDs Lecture Series
Oct 2022 MICCAI 2022 – Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC)
June 2022 CVPR 2022 – Tutorial on Human-Centered AI for Computer Vision

Understanding Deep Neural Networks

Dec 2021 Princeton University – Guest Lecture in COS324: Intro. to Machine Learning
Nov 2021 Princeton University – Guest Lecture in COS429: Intro. to Computer Vision
July 2021 Princeton University – AI4ALL Guest Lecture
June 2021 Princeton University – Visual AI Lab
June 2020 CVPR 2020 – Tutorial on Interpretable Machine Learning for Computer Vision
Jan 2020 University of Notre Dame – Department of Computer Science and Engineering

Tutorial on TorchRay: a PyTorch interpretability library for reproducible research

Nov 2019 ICCV 2019 – Workshop on Interpreting and Explaining Visual AI Models

Understanding Deep Networks via Extremal Perturbations

Nov 2019 ICCV 2019 (oral presentation)
April 2019 Continental
April 2019 OpenAI
April 2019 Stanford University – Stanford Vision and Learning Lab

Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in DNNs

July 2018 CVPR 2018 (spotlight presentation)

Using Human Brain Activity to Guide Machine Learning

Oct 2017 ICCV – Workshop on Mutual Benefits of Cognitive and Computer Vision (contributing talk)
April 2015 MIT Broad Institute – Girls Advancing in STEM (GAINS) Network Conference

Service

Research community service

2023 **CVPR 2023**, *Workshop Organizer*, XAI4CV: Explainable AI for Computer Vision
2018 – 2021 **Black in AI**, *Mentor and Program Committee*
2020 **ICML 2020**, *Workshop Organizer*, XXAI: Extending Explainable AI

University service

2022 – now **Princeton CS**, *Committee Member*, Climate & Inclusion committee

2021 – now **Princeton CS**, *Committee Member*, Lecturer hiring committee (2 academic cycles)

Other service

2018 – 2019 **Harvard Women in CS (WiCS)**, *Alumni Mentor*

2015 – 2016 **Oxford St. John's College MCR**, *Black and Minority Ethnic (BME) Representative*

2012 – 2015 **Harvard Women in CS (WiCS)**, *Webmaster, Board Member, and Mentor*

Student Advising

Princeton senior theses

2022 – now **Creston Brooks**

2022 – now **Alexis Sursock**

2022 – now **Devon Ulrich**

2022 – now **Indu Panigrahi**
Won the following: Computing Research Association (CRA) Outstanding Undergraduate Research Award Nominee (2022), Outstanding Independent Work Award (2022), Princeton Research Day Orange & Black Undergraduate Presentation Award (2022).

2021 – 2022 **Vedant Dhopte**, *Holistically Interpreting Deep Neural Networks via Channel Ablation*

2021 – 2022 **Frelicia Tucker**, *The Virtual Black Hair Experience: Evaluating Hairstyle Transform Generative Adversarial Networks on Black Women*

Other

2019 – 2021 **Kurtis David**, *Debiasing Convolutional Neural Networks via Meta Orthogonalization*, masters thesis at University of Austin, co-advised by Dr. Qiang Liu
First author on NeurIPS Workshop 2020 paper

Student Mentoring

2021 – 2022 **Nicole Meister**, *undergrad at Princeton University*, advised by Dr. Olga Russakovsky
First and second author on several papers under review

2021 – now **Sunnie S.Y. Kim**, *PhD at Princeton University*, advised by Dr. Olga Russakovsky
First and second author on several papers under review; first author on ECCV 2022 paper

2021 – now **Vikram V. Ramaswamy**, *PhD at Princeton University*, advised by Dr. Olga Russakovsky
First author and co-author on several papers under review

2018 – 2020 **Mandela Patrick**, *PhD at University of Oxford*, advised by Dr. Andrea Vedaldi
Co-lead author on paper on ICCV 2019 paper

Awards

2022 **Open Philanthropy AI Alignment Grant**

2022 **Princeton SEAS Innovation Grant**, *Project X Fund*

2018 **Open Philanthropy AI Fellow**

2018 CVPR Outstanding Reviewer

2017, 2018 Women in Computer Vision CVPR Travel Grant

2017, 2018 Murray Speight Research Grant,

2017 International Computer Vision Summer School Best Poster Award

2017, 2018 Murray Speight Research Grant

2016, 2017 Women in Machine Learning NeurIPS Travel Grant, *declined in 2017*

2015 **Rhodes Scholar**

2015 **NSF Graduate Research Fellowship**, *declined*

2015 **Fulbright Scholar to Tanzania**, *declined*

2015 **Hoopes Prize**, *for outstanding undergraduate thesis*

2015 **Derek Bok Certificate of Distinction**, *for outstanding CS121 teaching evaluations*

2014 Tech in the World Fellowship

2013 Apple iOS Scholarship

Reviewing

Journals Distill, IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI), IEEE Signal Processing Letters, Proceedings of the National Academy of Sciences (PNAS)

Conferences Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR), European Conference in Computer Vision (ECCV), IEEE Conference on Computer Vision & Pattern Recognition (CVPR), IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)

Workshops Black in AI (BAI), Women in Machine Learning (WiML), Women in Computer Vision (WiCV), Neural Architects

Skills

Languages Python, MATLAB, Java

Frameworks PyTorch, TensorFlow, Keras, MatConvNet, PyCaffe

Web Dev HTML/CSS, JavaScript, Svelte

References

Andrea Vedaldi, *U. of Oxford*

Olga Russakovsky, *Princeton*

Walter Scheirer, *U. of Notre Dame*

Harry Lewis, *Harvard*

Andrew Zisserman, *U. of Oxford*

Antonio Torralba, *MIT*