

Suraj Srinivas

ssrinivas@seas.harvard.edu · suuraj.srinivas@gmail.com · [suraj-srinivas.github.io](https://github.com/suraj-srinivas)

research interests

Trustworthy Machine Learning (Interpretability, Explainability and Robustness);
Computationally Efficient Machine Learning; Large Language Models; Generative Models

work experience

2022 - **Postdoctoral Research Fellow**,
Harvard University, USA,
Faculty Advisor: Prof. Himabindu Lakkaraju
Research Topics: Interpretable Machine Learning, Robustness.

education

2021 **Doctor of Philosophy**,
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland,
Faculty Advisor: Prof. François Fleuret
Thesis: Gradient-based Methods for Deep Model Interpretability.

2017 **Master of Science (Engineering)**,
Indian Institute of Science, Bangalore, India,
Faculty Advisor: Prof. R. Venkatesh Babu
Thesis: Learning Compact Architectures for Deep Neural Networks.

internships

winter 2020 **Research Intern**, *Qualcomm AI Research, Netherlands*,
Research on algorithms to sparsify neural networks.

summer 2016 **Research Intern**, *DataGrokr, India / Verisk Analytics, USA*,
Speeding up inference on deep neural networks using tensor factorization.

fall 2014 **Engineering Intern**, *Tonbo Imaging, Bangalore*,
Implemented image processing algorithms on FPGA for a thermal imaging camera.

summer 2013 **Research Intern**, *Indian Institute of Science, Bangalore*,
Research on computational photography to perform camera jitter compensation.

awards and honors

2022 **Best paper award** at ICML *Interpretable ML for Healthcare* Workshop

2022 **Highlighted reviewer** at *International Conference on Learning Representations (ICLR)*

2021 EPFL EDEE **PhD thesis distinction award** for top 8% thesis in EE

2019 ICML travel grant for ICML 2019

2017 **Best paper award** at NeurIPS *Learning with Limited Data* Workshop

2015 Xerox Research India travel grant for BMVC 2015

2014 **All India Rank 399** (99.8%ile) in the Graduate Aptitude Test in Engineering (GATE) for entrance to graduate school in electronics and communications engineering

2010 **State Rank 191** (99.8%ile) in the Karnataka Common Entrance Test (CET) for entrance to undergraduate engineering programmes.

research articles

Total citations: 1850+ | **h-index:** 11

highlighted publications

- 2023 **Suraj Srinivas***, Sebastian Bordt*, Hima Lakkaraju. (*co-first-author)
Which Models have Perceptually-Aligned Gradients? An Explanation via Off-Manifold Robustness.
Neural Information Processing Systems (NeurIPS) - **Spotlight (Top 3%)**
- 2022 **Suraj Srinivas***, Kyle Matoba*, Hima Lakkaraju, François Fleuret. (*co-first-author)
Efficient Training of Low-Curvature Neural Networks.
Neural Information Processing Systems (NeurIPS)
- 2022 Tessa Han, **Suraj Srinivas**, Hima Lakkaraju.
Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations.
Neural Information Processing Systems (NeurIPS)
ICML Interpretable ML for Healthcare Workshop - **Best Paper Award**
- 2021 **Suraj Srinivas**, François Fleuret.
Rethinking the Role of Gradient-based Attribution Methods in Model Interpretability.
International Conference on Learning Representations (ICLR) - **Oral (Top 1%)**
- 2018 **Suraj Srinivas**, François Fleuret.
Knowledge Transfer with Jacobian Matching.
International Conference on Machine Learning (ICML)
NeurIPS Learning with Limited Data (LLD) Workshop - **Best Paper Award**

additional peer-reviewed publications

- 2023 Usha Bhalla*, **Suraj Srinivas***, Hima Lakkaraju. (*co-first-author)
Discriminative feature attributions: Bridging post hoc explainability and inherent interpretability.
Neural Information Processing Systems (NeurIPS)
- 2023 Anna Meyer*, Dan Ley*, **Suraj Srinivas**, Hima Lakkaraju.
On Minimizing the Impact of Dataset Shifts on Actionable Explanations.
Uncertainty in Artificial Intelligence (UAI) - **Oral (Top 5%)**
- 2022 Marwa El Halabi, **Suraj Srinivas**, Simon Lacoste-Julien.
Data-Efficient Structured Pruning via Submodular Optimization.
Neural Information Processing Systems (NeurIPS)
- 2022 **Suraj Srinivas**, Andrey Kuzmin, Markus Nagel, Mart van Baalen, Andrii Skliar, Tijmen Blankevoort.
Cyclical Pruning for Sparse Neural Networks.
Computer Vision and Pattern Recognition Workshops (CVPRW) - **Oral**
- 2019 **Suraj Srinivas**, François Fleuret.
Full-Gradient Representation for Neural Network Visualization.
Neural Information Processing Systems (NeurIPS)
- 2018 Akshayvarun Subramanya, **Suraj Srinivas**, R. Venkatesh Babu.
Estimating Confidence for Deep Neural Networks through Density Modelling.
IEEE Conference on Signal Processing and Communications (SPCOM)
- 2017 **Suraj Srinivas**, Akshayvarun Subramanya, R. Venkatesh Babu.
Training Sparse Neural Networks.
Computer Vision and Pattern Recognition Workshops (CVPRW) - **Oral**

- 2017 Lokesh Boominathan, **Suraj Srinivas**, R. Venkatesh Babu.
Compensating for Large In-plane Rotations in Natural Images.
Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)
- 2016 **Suraj Srinivas**, R. Venkatesh Babu.
Learning the Architecture of Deep Neural Networks.
British Computer Vision Conference (BMVC)
- 2015 **Suraj Srinivas**, R. Venkatesh Babu.
Data-free Parameter Pruning for Deep Neural Networks.
British Computer Vision Conference (BMVC)

preprints & workshop papers

- 2023 Aounon Kumar, Chirag Agarwal, **Suraj Srinivas**, Soheil Feizi, Hima Lakkaraju.
Certifying LLM safety against adversarial prompting.
- 2023 Tessa Han, **Suraj Srinivas**, Hima Lakkaraju.
Efficient estimation of average-case robustness for multi-class classification.
ICML 2023 workshop on Formal Verification of Machine Learning
- 2023 Dan Ley, Leonard Tang, Matthew Nazari, Hongjin Lin, **Suraj Srinivas**, and Hima Lakkaraju.
Consistent explanations in the face of model indeterminacy via ensembling.
ICML 2023 workshop on Interpretable Machine Learning for Healthcare
- 2023 Alex Lin, Lucas Paes, Sreeharsha Tanneru, **Suraj Srinivas**, Hima Lakkaraju.
Word-Level Explanations for Analyzing Bias in Text-to-Image Models.
ICML 2023 Workshop on Challenges in Deploying Generative AI

book chapters

- 2017 **Suraj Srinivas**, Ravi Kiran Sarvadevabhatla, Konda Reddy Mopuri, Nikita Prabhu, Srinivas SS Kruthiventi, R. Venkatesh Babu.
A taxonomy of deep convolutional neural nets for computer vision.
Book chapter: *Deep Learning for Medical Image Analysis*, Elsevier
Journal version: *Frontiers in Robotics and AI*

talks

- 11/2023 *On the Missing Conceptual Foundations of Interpretable Machine Learning*
Indian Institute of Technology, Hyderabad
- 03/2023 *Pitfalls and Opportunities with Feature Importance Methods*
[MERL seminar series](#), Boston
- 07/2022 *Pitfalls and Opportunities with Feature Attribution Methods*
Simons Institute, UC Berkeley
- 06/2022 *Pitfalls and Opportunities with Feature Attribution Methods*
Vanderbilt University, USA
- 03/2022 *Cyclical Pruning for Neural Network Sparsity*
Google Sparsity Reading Group
- 08/2021 *Pitfalls of Saliency Map Interpretation in Deep Neural Networks*
HES-SO, Sierre, Switzerland
- 04/2021 *Rethinking the Role of Gradient-based Attribution Methods for Model Interpretability*
ICLR (virtual)

- 01/2020 *Neural Network Interpretability using Full-Gradient Representation*
Indian Institute of Science, Bangalore
- 01/2020 *Full-Gradient Representation for Neural Network Visualization*
[ML for Astrophysicists Club](#)
- 11/2019 *Full-Gradient Representation for Neural Network Visualization*
Swiss Machine Learning Day, Lausanne
- 05/2019 *Complete Saliency Maps using Full-Jacobians*
Valais / Wallis AI workshop, Martigny
- 07/2018 *Knowledge Transfer with Jacobian Matching*
ICML, Stockholm
- 07/2016 *Making Deep Neural Networks Smaller and Faster*
Deep Learning Conf, Bangalore

reviewing

- Conferences AAAI, CVPR, ECCV, NeurIPS (2020) ; WACV, ICML, ICCV, NeurIPS (2021);
ICLR, ICML, NeurIPS (2022); ICLR, AISTATS (2023)
- Journals IEEE SP-Letters, Elsevier Neural Networks, IEEE T-PAMI, Nature Communications

teaching

- 2023 **Co-instructor** for *Interpretability and Explainability in ML*
Instructors: Prof. Hima Lakkaraju, Jiaqi Ma, Suraj Srinivas
Harvard University, USA
Webpage: <https://interpretable-ml-class.github.io/>
- 2018, '19, '21 **Teaching Assistant** for *Deep Learning*
Instructor: Prof. François Fleuret
EPFL, Switzerland
- 2021 **Guest Lecturer** on Interpretability for *Deep Learning for Computer Vision*
Instructor: Prof. R. Venkatesh Babu
Indian Institute of Science, Bangalore

research mentoring

- 2023 Usha Bhalla & Alex Oesterling (PhD students, Harvard)
Concept Decompositions with CLIP, ongoing
- 2022-23 Tessa Han (PhD candidate, Harvard)
Local Function Approximation to Characterize Explanations, NeurIPS 2022
Efficient Estimation of Local Robustness, ICML Workshops, 2023
- 2023 Usha Bhalla (PhD student, Harvard)
Verifiable Feature Attributions, NeurIPS 2023
- 2023 Daniel Ley (PhD student, Harvard)
On Minimizing the Impact of Dataset Shifts on Actionable Explanations, UAI 2023
- 2017 Akshayvarun Subramanya (Research Assistant, IISc)
Estimating Confidence for Deep Neural Networks via Density Modeling, SPCOM 2017
- 2016 Lokesh Boominathan (Research Assistant, IISc)
Compensating for Large In-plane Rotations in Natural Images, ICVGIP 2016

service

- 2023 Co-organizer of "XAI in Action: Past, Present, and Future Applications"
NeurIPS 2023 workshop

references

Prof. Himabindu Lakkaraju
Harvard University, USA
hlakkaraju@hbs.edu

Prof. François Fleuret
University of Geneva, Switzerland
francois.fleuret@unige.ch

Prof. R. Venkatesh Babu
Indian Institute of Science, Bangalore
venky@iisc.ac.in

Prof. Flavio du Pin Calmon
Harvard University, USA
flavio@seas.harvard.edu