Francesco Croce

E-mail: francesco91.croce@gmail.com **Google Scholar** **GitHub:** https://github.com/fra31

EDUCATION

University of Tübingen Germany Sept. 2017 - Sept. 2023	**PhD in Computer Science** Machine Learning group, supervised by Prof. Matthias Hein Thesis: "Evaluating and Improving the Robustness of Image Classifiers against Adversarial Attacks" Overall grade: Excellent (Summa cum Laude)
University of Torino Italy Sept. 2014 - July 2016	**Master's Degree in Mathematics** Final grade: 110/110 cum Laude and honor mention Thesis: "Generalized Poisson processes: estimation techniques"
University of Torino Italy Sept. 2010 - Oct. 2013	**Bachelor's Degree in Mathematics for Finance and Insurance** Final grade: 110/110 cum Laude Thesis: "Tic-Tac-Toe on hypercubes: draws and victories"

WORK EXPERIENCE

EPFL, Switzerland Oct. 2023 - now	**Postdoctoral Researcher** Theory of Machine Learning group, supervised by Prof. Nicolas Flammarion
DeepMind London, UK June - Nov. 2022	**Research Internship** Hosted by Dr. Sven Gowal Main project: "Seasoning model soups for robustness to adversarial and natural distribution shifts" (CVPR 23)

AWARDS

PhD Thesis Awards

➢ **DAGM MVTec Dissertation Award 2024** from the German Association for Pattern Recognition (DAGM, Deutsche Arbeitsgemeinschaft für Mustererkennung), which honors an outstanding dissertation in the fields of pattern recognition, image processing, machine vision, and machine learning.

➢ **Wilhelm Schickard Dissertation Award 2024** as best dissertation of the Department of Computer Science at the Eberhard Karls University of Tübingen

Paper Awards

➢ **Best Paper Honorable Mention Award** for "RobustBench: a standardized adversarial robustness benchmark" at ICLR 21 Workshop on Security and Safety in ML Systems

➢ **Honorable Mention Award** for "A randomized gradient-free attack on ReLU networks" at GCPR 18

Competitions **1st place** at the "Find the Trojan: Universal Backdoor Detection in Aligned LLMs" competition, co-located with SaTML 24

Grants (co-author) ➢ "Safe GenAI via Robust Content Moderation Models" ($100k funded by Google, 2024)
➢ "Robust LLM-based Scoring of Agent Alignment" ($200k funded by Schmidt Sciences, 2025)

Scholarships Winner of INdAM's (National Institute for High Mathematics) scholarship for top 40 students in a national math contest, confirmed for the three years of BSc

ACADEMIC SERVICE

Reviewer ➢ Conferences: ICML 24, 23, 22, 21, 20 (**top 33% reviewer**), NeurIPS 24, 23, 22, 21, 20, 19 (**top 400 reviewer**), ICLR 25, 24, 23, 22, 21 (**outstanding reviewer**), CVPR 25, 24, 23 (**outstanding reviewer**), 22, 21, ICCV 23, 21, AAAI 22, ECCV 24, SaTML 25
➢ Journals: Artificial Intelligence (ARTINT), Machine Learning (MACH), Circuits, Systems & Signal Processing (CSSP), IEEE Transactions on Neural Networks and Learning Systems, Information, Forensics & Security, Pattern Analysis and Machine Intelligence (TPAMI), Transactions on Machine Learning Research (TMLR)

Co-organizer ➢ 1st Workshop on Test-Time Adaptation: Model, Adapt Thyself! (MAT) at CVPR 24
➢ "A Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning" Workshop at ICML 21

TEACHING AND SUPERVISION EXPERIENCE

Students supervision I supervised students on several projects:
➢ PhD students: Sara Ghazanfari (New York University)
➢ Master's students: Naman D. Singh (University of Tübingen, contributed to work **appeared at AAAI 22**), Hao Zhao (EPFL, work **published at ICML 24**), Rishika Bhagwatkar (MILA and Université de Montréal)
➢ Bachelor's students: Kari Gustedt, Sascha Bielawski (University of Tübingen)

Teaching assistant ➢ Tutor for "Statistical Machine Learning" (2021), "Mathematics for Machine Learning" (2019/20) at the University of Tübingen
➢ Tutor for "Discrete Mathematics and Logic" (2017), "Algebra and Geometry" (2015/16), "Probability and Statistics" (2015/16, 2012/13), "Mathematical Analysis 2" (2012/13) at the University of Torino

LIST OF PUBLICATIONS

Selected publications

S. Ghazanfari, S. Garg, N. Flammarion, P. Krishnamurthy, F. Khorrami, **F. Croce**. Towards unified benchmark and models for multi-modal perceptual metrics. arXiv (December 2024) [paper]

F. Croce*, C. Schlarmann*, N. D. Singh*, M. Hein. Adversarially robust CLIP models induce better (robust) perceptual metrics. SaTML 25 [paper]

F. Croce, S. Rebuffi, E. Shelhamer, S. Gowal. Seasoning model soups for robustness to adversarial and natural distribution shifts. CVPR 23 [paper]

F. Croce*, M. Andriushchenko*, V. Sehwag*, E. Debenedetti*, N. Flammarion, M. Chiang, P. Mittal, M. Hein. RobustBench: a standardized adversarial robustness benchmark. NeurIPS 21 Datasets and Benchmarks Track [paper, website]

F. Croce, M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. ICML 20 [paper, code]

Full list

S. Ghazanfari, S. Garg, N. Flammarion, P. Krishnamurthy, F. Khorrami, **F. Croce**. Towards unified benchmark and models for multi-modal perceptual metrics. arXiv (December 2024) [paper]

N. D. Singh, **F. Croce**, M. Hein. Perturb and recover: fine-tuning for effective backdoor removal from CLIP. arXiv (December 2024) [paper]

F. D'Angelo, **F. Croce**, N. Flammarion. Selective induction Heads: How Transformers Select Causal Structures in Context. ICLR 25 [paper]

H. Zhao, M. Andriushchenko, **F. Croce**, N. Flammarion. Is in-context learning sufficient for instruction following in LLMs? ICLR 25 [paper]

M. Andriushchenko, **F. Croce**, N. Flammarion. Jailbreaking leading safety-aligned LLMs with simple adaptive attacks. ICLR 25 [paper]

F. Croce*, C. Schlarmann*, N. D. Singh*, M. Hein. Adversarially robust CLIP models induce better (robust) perceptual metrics. SaTML 25 [paper]

J. Rando, **F. Croce**, K. Mitka, S. Shabalin, M. Andriushchenko, N. Flammarion, F. Tramèr. Competition report: finding universal jailbreak backdoors in aligned LLMs. arXiv (April 2024) [paper]

P. Chao*, E. Debenedetti*, A. Robey*, M. Andriushchenko*, **F. Croce**, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, E. Wong. JailbreakBench: an open robustness benchmark for jailbreaking large language models. NeurIPS 24 Datasets and Benchmarks Track [paper]

F. Croce*, N. D. Singh*, M. Hein. Robust semantic segmentation: strong adversarial attacks and fast training of robust models. ECCV 24 [paper]

H. Zhao, M. Andriushchenko, **F. Croce**, N. Flammarion. Long is more for alignment: a simple but tough-to-beat baseline for instruction fine-tuning. ICML 24 [paper]

C. Schlarmann, N. D. Singh, **F. Croce**, M. Hein. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. ICML 24 [paper]

F. Croce, M. Hein. Segment (almost) nothing: prompt-agnostic adversarial attacks on segmentation models. SaTML 24 [paper]

N. D. Singh*, **F. Croce***, M. Hein. Revisiting adversarial training for ImageNet: architectures, training and generalization across threat models. NeurIPS 23 [paper]

M. Andriushchenko, **F. Croce**, M. Müller, M. Hein, N. Flammarion. A modern look at the relationship between sharpness and generalization. ICML 23 [paper]

F. Croce, S. Rebuffi, E. Shelhamer, S. Gowal. Seasoning model soups for robustness to adversarial and natural distribution shifts. CVPR 23 [paper]

S. Rebuffi, **F. Croce**, S. Gowal. Revisiting adapters with adversarial training. ICLR 23 [paper]

M. Augustin*, V. Boreiko*, **F. Croce**, M. Hein. Diffusion visual counterfactual explanations. NeurIPS 22 [paper]

V. Boreiko, M. Augustin, **F. Croce**, P. Berens, M.Hein. Sparse visual counterfactual explanations in image space. GCPR 22 [paper]

F. Croce, M. Hein. On the interplay of adversarial robustness and architecture components: patches, convolution and attention. "New Frontiers in Adversarial Machine Learning" Workshop at ICML 22 [paper]

F. Croce*, S. Gowal*, T. Brunner*, E. Shelhamer*, M. Hein, T. Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. ICML 22 [paper, code]

F. Croce, M. Hein. Adversarial robustness against multiple and single lp-threat models via quick fine-tuning of robust classifiers. ICML 22 [paper, code]

F. Croce, M. Andriushchenko, N. D. Singh, N. Flammarion, M. Hein. Sparse-RS: a versatile framework for query-efficient sparse black-box adversarial attacks. AAAI 22 [paper, code]

F. Croce*, M. Andriushchenko*, V. Sehwag*, E. Debenedetti*, N. Flammarion, M. Chiang, P. Mittal, M. Hein. RobustBench: a standardized adversarial robustness benchmark. NeurIPS 21 Datasets and Benchmarks Track [paper, website]

F. Croce, M. Hein. Mind the box: l1-APGD for sparse adversarial attacks on image classifiers. ICML 21 [paper, code]

M. Andriushchenko*, **F. Croce***, N. Flammarion, M. Hein. Square Attack: a query-efficient black-box adversarial attack via random search. ECCV 20 [paper, code]

F. Croce, M. Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. ICML 20 [paper, code]

F. Croce, M. Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. ICML 20 [paper, code]

F. Croce, M. Hein. Provable robustness against all adversarial lp-perturbations for p ≥ 1. ICLR 20 [paper, code]

F. Croce, M. Hein. Sparse and imperceivable adversarial attacks. ICCV 19 [paper, code]

F. Croce*, J. Rauber*, M. Hein. Scaling up the randomized gradient-free adversarial attack reveals overestimation of robustness using established attacks. Intl. Journal of Computer Vision, 2019. [paper, code]

F. Croce*, M. Andriushchenko*, M. Hein. Provable robustness of ReLU networks via maximization of linear regions. AISTATS 19 [paper, code]

F. Croce, M. Hein. A randomized gradient-free attack on ReLU networks. GCPR 18 [paper]