Dr. Yinpeng Dong

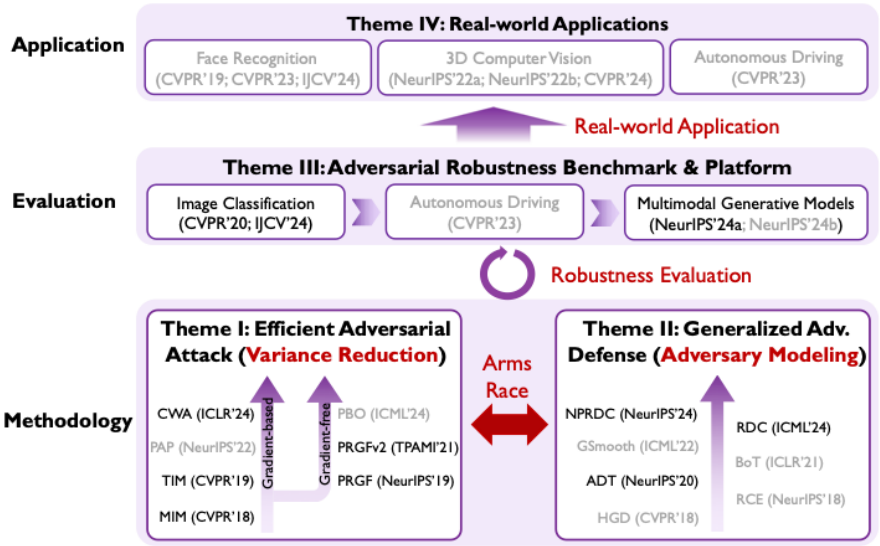
Tsinghua University

https://dongyp13.github.io/

Bio: Dr. Yinpeng Dong is an incoming Assitant Professor at College of AI, Tsinghua University. He is now a Postdoc at the Department of Computer Science, Tsinghua University, where he also obtained his B.E. and Ph.D. degrees. His research interests are primarily on adversarial machine learning and AI safety. He has published over 50 papers in the prestigious conferences and journals, including TPAMI, IJCV, NeurIPS, ICML, CVPR. These papers have amassed more than 10000 citations (Google Scholar). He has organized several adversarial ML workshops at ICML'21, AAAI'22, ICCV'23, etc. He served as an Area Chair for ICLR'25 and ICML'25. He received CCF Outstanding Doctoral Dissertation Award, Tsinghua Oustanding Postdoctoral Researcher, Microsoft Research Asia Fellowship, Baidu Fellowship, ByteDance Scholarship, etc. He was also been listed among World's 2% Scientists (2022-2024) by Stanford/Elsevier.

Contribution:

Machine learning and deep learning models have achieved remarkable success, often surpassing human-level performance in various tasks. However, their significant vulnerability to adversarial examples—inputs crafted with subtle perturbations that are nearly indistinguishable from normal data but induce erroneous predictions—highlights a stark contrast with human intelligence and raises critical security and safety concerns in real-world applications. My research focuses on enhancing the robustness and safety of these models in adversarial environments, with an emphasis on adversarial machine learning (AdvML).



I have developed efficient adversarial attack algorithms, including the Momentum Iterative Method, Translation Invariant Method, and Prior-Guided Random Gradient-Free Method, to generate adversarial examples under black-box settings, facilitating understanding of model vulnerabilities. Additionally, I have proposed generalizable adversarial defenses such as Adversarial Distributional

Training and the Robust Diffusion Classifier, which enhance model robustness under a broad spectrum of unseen attacks. To address the ongoing “arms race” between attacks and defenses, I have also contributed to the development of robustness evaluation platforms and benchmarks, such as ARES (https://github.com/thu-ml/ares) and MultiTrust (https://multi-trust.github.io), which evaluate the robustness and trustworthiness of deep learning and multimodal foundation models. Furthermore, my research extends to real-world applications, identifying and mitigating practical security threats in deep learning systems, bridging the gap between theoretical advancements and their deployment in critical, safety-sensitive domains.

Dr. Salem Lahlou

Mohamed bin Zayed University of Artificial Intelligence

https://saleml.github.io/

Bio: Dr. Salem Lahlou is an Assistant Professor in the Machine Learning department at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). Previously, he was a Senior Researcher at TII in 2024 and completed his PhD in 2023 from Mila and Université de Montréal under the supervision of Yoshua Bengio. He also worked as a researcher at Google Brain at Paris and IBM Research at Singapore. His research interests span multiple areas of machine learning, with a particular focus on understanding intelligence (both animal and artificial), uncertainty estimation, reinforcement learning sample complexity, and GFlowNets. He has published several influential papers in top-tier AI conferences, including "DEUP: Direct Epistemic Uncertainty Prediction," "GFlowNet Foundations," and "BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning." He received his education from prestigious French institutions, including École Polytechnique (applied mathematics) and École Normale Supérieure Paris-Saclay (statistical learning). In terms of academic service, he has served as a reviewer for top conferences in the field.

Contributions:

Understanding and improving AI systems' reasoning and reliability is crucial as these systems become increasingly deployed in complex tasks and real-world applications. The ability to quantify uncertainty, perform efficient search and sampling, and ensure reliable reasoning are fundamental challenges that need to be addressed for the safe and effective deployment of AI systems. To address these challenges, Dr. Lahlou has made significant contributions through the development of Generative Flow Networks (GFlowNets), a novel framework that bridges reinforcement learning and probabilistic modeling. His work on GFlowNets has established their theoretical foundations, demonstrated their effectiveness in Bayesian inference of discrete and continuous structures, and shown their advantages in handling multimodal distributions compared to traditional variational methods. He also released torchgfn, an open-source software framework for GFlowNets that enables both discrete and general applications. Beyond GFlowNets, he has made important contributions to uncertainty estimation through DEUP (Direct Epistemic Uncertainty Prediction), a method that addresses model misspecification in interactive learning settings by maintaining a secondary predictor trained on the main model's errors. This approach provides more reliable uncertainty

estimates for decision-making in downstream tasks. Currently, his research focuses on scaling GFlowNets to more complex problems, leveraging them to enhance Large Language Model reasoning, and developing better uncertainty estimation methods. These contributions collectively aim to improve the reliability, efficiency, and reasoning capabilities of AI systems, particularly in scenarios requiring robust decision-making under uncertainty.