

Alexander Matyasko, PhD

Researcher in Adversarial Machine Learning, Robustness, and AI Safety



matyasko.net



Google Scholar: 821+ citations, h-index 7



GitHub



LinkedIn



alexander.matyasko@gmail.com



+65 8468 4575



Work Authorization: Singapore Permanent Resident

Research Vision

My research develops principled methods for evaluating and improving machine-learning robustness under adversarial attacks, distribution shifts, and deployment constraints. I focus on perception systems, including semantic segmentation and autonomous-driving pipelines, where reliability failures have direct safety implications. My work combines robust optimization, adversarial evaluation, and empirically grounded stress testing to expose failure modes missed by standard benchmarks. I am extending this agenda to foundation and multimodal models, with emphasis on robustness evaluation, inference-time defenses, and security failures in safety-critical settings.

Education

2020

Ph.D. in Machine Learning, Nanyang Technological University

Thesis: *Towards Deep Neural Networks Robust to Adversarial Examples*

- Generalized SVM margin maximization objective to improve the robustness of DNNs.
- Formulated a problem of learning a robust classifier in the framework of generative adversarial networks (GAN).
- Proposed a primal-dual proximal gradient attack to estimate the model's robustness accurately for all ℓ_p -norms.

Thesis advisor: Prof. Lap-Pui Chau

2013

Dip.Eng. in Informatics ¹, Belarusian State University of Informatics and Radioelectronics

Major: Applied Mathematics and Computer Science

Research Experience

My research experience centers on adversarial machine learning, with a focus on developing robustness evaluation methods, adversarial attacks, and defenses across safety-critical perception, robust optimization, federated learning, physics-informed learning, and multimodal model evaluation.

Singapore Institute of Technology

May 2024 – Present

Research Fellow

Robust AI Grand Challenge on Safety and Robustness for Autonomous Vehicles

- Introduced **TsallisPGD**, a semantic-segmentation attack that outperforms PGD-style baselines on Cityscapes, PASCAL VOC, and ADE20K.
- Developed **physically grounded adversarial attacks** for systematic robustness testing of AV perception systems.
- Optimized Jetson Orin deployment, reducing inference latency by **30%** under a 500 ms budget through model and data-loading optimizations.
- Co-developed **moving-target defenses** to reduce adversarial transferability and model-extraction risk in deployed AV perception systems.

¹Diploma of Engineer is a five-year program in Belarus that meets the expectations of a Master's degree.

Research Experience (continued)

Agency for Science, Technology and Research (A*STAR)

Mar. 2023 – Mar. 2024  **Senior Scientist** – Centre for Frontier AI Research

- Introduced a novel threat model for multi-norm robustness evaluation, exposing failure modes of ℓ_p -trained defenses missed by union-of-norms robustness evaluations.
- Co-authored research on **multimodal reasoning**, establishing theoretical results on modality-fusion failure and designing simulation studies for VideoQA transformer models (ICML 2024).
- Contributed to physics-informed neural networks research for adaptive sensor placement (IJCNN 2024).

National University of Singapore

Feb. 2022 – Feb. 2023  **Research Fellow** – Deep Learning in Theory and Practice Lab

- Co-developed **Interpolated Adversarial Training**, reducing the clean-accuracy drop caused by standard adversarial training while maintaining robustness to adversarial attacks.
- Analyzed the robustness of physics-informed neural networks to adversarial input perturbations and data poisoning attacks.
- Delivered a graduate ML seminar on physics-informed neural networks for CS6216 to 50+ students.

Imperial College London

Feb. 2021 – Dec. 2021  **Research Associate** – Resilient Information Systems Security Group

Horizon 2020 Musketeer: Machine Learning to Augment Shared Knowledge in Federated Privacy-Preserving Scenarios

- Developed robust aggregation defenses for federated learning, improving model performance against malicious clients by up to **11%**.
- Built the **MUSKETEE** ML library for evaluating federated-learning robustness in an EU Horizon 2020 multi-partner consortium.
- Analyzed the robustness of federated averaging to poisoning attacks and translated findings into platform-level defense recommendations.

Nanyang Technological University

Nov. 2017 – Dec. 2020  **Research Associate**

Computer vision applications for industry partners.

- Introduced **PDPGD**, a primal-dual proximal adversarial attack for efficient norm minimization, outperforming prior ℓ_p -minimization attacks across multiple norms.
- Developed end-to-end computer-vision pipelines for industry-partner projects in rain removal, fire and smoke detection, and 3D point-cloud tracking.
- Implemented GPU-accelerated rain removal for near-real-time video enhancement.

Publications

Selected Publications

- 1 **Matyasko, A.**, Lou, X., Atmosukarto, I., & Zhang, W. (2026). Tsallispgd: Adaptive gradient weighting for adversarial attacks on semantic segmentation. *2026 International Joint Conference on Neural Networks (IJCNN)*, 1–8.

- 2 Lamb, A., Verma, V., Kawaguchi, K., **Matyasko, A.**, Khosla, S., Kannala, J., & Bengio, Y. (2022). Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy. *Neural Networks*, 154, 218–233.
- 3 **Matyasko, A.**, & Chau, L. (2021). Pdpd: Primal-dual proximal gradient descent adversarial attack. *arXiv e-prints*, Article arXiv:2106.01538.
- 4 **Matyasko, A.**, & Chau, L. (2018). Improved network robustness with adversary critic. In *Advances in neural information processing systems* 31 (pp. 10578–10587). Curran Associates, Inc.
- 5 **Matyasko, A.**, & Chau, L. (2017). Margin maximization for robust classification using deep learning. *2017 International Joint Conference on Neural Networks (IJCNN)*, 300–307.

Manuscripts Under Review

- 6 **Matyasko, A.**, Lou, X., & Fernando, B. (2026). Mind the hull: C-apgd for evaluating multinorm robustness. *Under review*.
- 7 Mo, Y., Lou, X., **Matyasko, A.**, Zhang, W., Atmosukarto, I., & Sun, X. (2026). Adversarially robust semantic segmentation for autonomous driving via dynamic moving target defense and synthetic data augmentation. *Under review*.

Other Publications

- 8 Ee, Y. K., Zhang, H., **Matyasko, A.**, & Fernando, B. (2025). Deduce and select evidences with language models for training-free video goal inference. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- 9 Rawal, I. S., **Matyasko, A.**, Jaiswal, S., Fernando, B., & Tan, C. (2024). Dissecting multimodality in VideoQA transformer models by impairing modality fusion. *Proceedings of the 41st International Conference on Machine Learning*, 235, 42213–42244.
- 10 Zhou, C., Chen, J., Yang, Z., **Matyasko, A.**, & Png, C. E. (2024). Enhanced physics-informed neural networks with optimized sensor placement via multi-criteria adaptive sampling. *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8.
- 11 Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A., **Matyasko, A.**, Behzadan, V., Hambardzumyan, K., Zhang, Z., Juang, Y.-L., Li, Z., Sheatsley, R., Garg, A., Uesato, J., ... McDaniel, P. (2016). Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv e-prints*, Article arXiv:1610.00768.
- 12 Chen, J., **Matyasko, A.**, & Chau, L. (2014). *A light field sparse representation structure and its fast coding technique*.

Technical Skills

Safety & Robustness	Adversarial robustness, robustness evaluation, stress testing, and security evaluation of ML systems.
Models & Perception	Semantic segmentation, object detection, 3D perception, multi-sensor fusion, transformers, multimodal (vision–language) models, foundation models.
Frameworks	PyTorch (DDP, torch.compile), TensorFlow, JAX, MMDetection/MMCV, Foolbox, CleverHans, mixed-precision training.
Systems	GPU-accelerated training and inference, distributed training (DDP), mixed-precision training, large-scale experimentation.
Languages & Tools	Python, C/C++, CUDA, Bash, Linux, Git, Docker, and $\LaTeX$.

Teaching Experience

Teaching Interests  Machine Learning; Trustworthy & Robust AI; Optimization for ML; AI Security and Safety; Advanced Topics in Deep Learning.

National University of Singapore

Aug. 2022 – Dec. 2022  **Lecturer** for CS6216 Advanced Topics in Machine Learning seminar on Physics-Informed Deep Learning to masters/PhD students.

- Designed and delivered a graduate-level seminar on physics-informed deep learning, emphasizing robustness, inductive bias, and optimization stability.
- Mentored student project development and assessment in theory-grounded model design.

Nanyang Technological University

Aug. 2019 – Jun. 2020  **Research Mentor** for URECA undergraduate research program

- Supervised two undergraduate researchers from project scoping to final research delivery.

Aug. 2016 – Mar. 2017  **Lab Instructor** for undergraduate course “EE2008 Data Structures & Algorithms”

- Facilitated the instruction and assisted students during practical labs.

Service

Organizer  Volunteer Coordinator at AAAI’26.

Reviewer  NeurIPS (2019–2026), ICLR (2020–2024), ICML (2020–2023), ICME, DSP.

Contributor  Open-source contributions to Theano (#5159, #5198, #5323, #5778, #5914), TensorFlow (#6664), PyTorch (#1373, #1481), CleverHans (#125), Foolbox (#559).

Awards

2017  DeepMind Travel Award, ICML 2017 Workshop on Implicit Models.

2013  A*STAR Graduate Scholarship for doctoral study at Nanyang Technological University.

References

Available upon request.