# Adversarial attacks impact on physical environment

Andrei Vizureanu
*AI Multimedia Lab*
*National University of Science and Technology POLITEHNICA Bucharest*
Bucharest, Romania
andrei.vizureanu@gmail.com

Bogdan Ionescu
*AI Multimedia Lab*
*National University of Science and Technology POLITEHNICA Bucharest*
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—As deep learning continues to transform computer vision, camera-based smart systems are becoming integral to critical fields such as autonomous driving, surveillance, and biometric authentication. Unfortunately, this growing reliance on visual data exposes these systems to a serious category of adversarial threats: physical adversarial attacks. Unlike their digital counterparts, these attacks utilize real-world perturbations—like stickers, clothing, or projected light—to deceive deep neural networks in uncontrolled environments. This article explores the mechanics, challenges, and consequences of physical adversarial attacks, providing an overview of the latest trends and research over the past decade. It emphasizes the practical implications of these attacks, their stealth tactics, and the urgent need for effective defense strategies.**

***Index Terms*—AI threats, vulnerabilities, ML-DL, adversarial attacks, physical space**

## I. Introduction

While adversarial machine learning has gained attention primarily through digital attacks, recent research exposes a more concerning reality: attacks manifesting physically in our environment. These ”physical adversarial attacks” involve perturbing tangible objects in the real world, such as traffic signs, clothing, or facial accessories, with crafted patterns to deceive camera-fed models. As digital attacks imply that the adversary may have direct access to the data fed into the model, the physical ones do not know the digital representation of the data; the model is being directly fed with sensor inputs (e.g., images from video cameras and microphones). The danger lies in their deployability without any system access, making them both practical and threatening for black-box settings.

There are also concerns about the potential limitations of continual learning approaches against adversarial attacks, and their suitability for deployment in real-world settings [2].

## II. Feature visualization and adversarial examples

Feature visualization provides insight into how neural networks operate, showcasing their hierarchical extraction and representation of features—from low-level textures to high-level semantic concepts. Its effectiveness relies heavily on appropriate regularization, an effective optimization strategy, and careful architectural considerations. [3], [4], [5]

Neural networks do not just memorize—they learn structured, reusable, and interpretable features. With the right tools, we can begin to decode what those features are, why they matter, and how to make networks more transparent.

New approaches emphasize how adversarial examples can be integrated into explainable machine learning scenarios. This allows humans to gain insight not only into the input and the output classification but also into the reasoning behind the model's decision. This information can be further utilized to generate and understand adversarial perturbations [6].

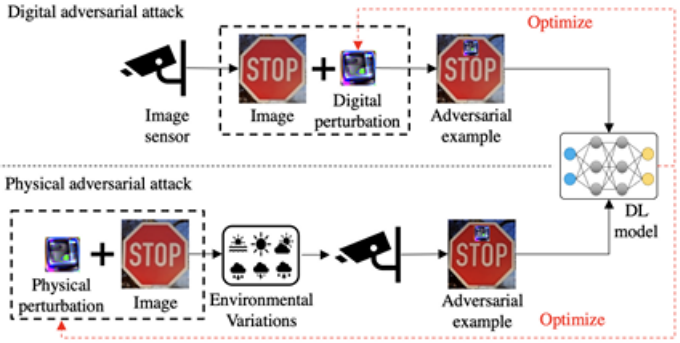


Fig. 1: Digital vs physical adversarial attacks [1].

## III. Attack implications in physical space

### A. Attack Modalities and Real-World Forms

- Patch- and Sticker-Based Attacks: Strategically placed perturbations on objects or persons. These patches can cause misclassification (e.g., turning a stop sign into a speed-limit sign) [7], [8].
- Camouflaged Clothing: T-shirts or cloaks printed with adversarial textures that evade person detectors. Such attacks gain stealth by mimicking natural patterns or graffiti [9], [10].
- Wearable Accessories: Adversarial glasses, hats, and masks have proven effective in impersonating identities or evading facial recognition systems.
- Light-Based Perturbations: Laser beams or projectors can inject malicious patterns into scenes without touching the object, offering transient and highly stealthy attack mechanisms [11].



*2nd Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 15 July, 2025*

- Device Manipulation: Some attacks modify the imaging process itself, e.g., placing stickers on camera lenses or exploiting rolling shutter effects [12].

### B. Practical Challenges

- Viewpoint Variation: Unlike digital attacks, physical ones must remain effective from diverse angles and distances.
- Environmental Lighting: Shadows, reflections, and lighting shifts can degrade or reveal perturbations.
- Fabrication Constraints: Perturbations must be printable and physically realizable with accurate color reproduction and minimal fabrication error.
- Non-Rigid Deformation: Especially with wearables like shirts, the surface continuously deforms due to body movement.

### C. Evaluation metrics

Although there are no standard evaluation metrics, different survey papers propose some comprehensive evaluation metrics for stealth, robustness, and transferability [1].

- **Stealth**: Some attacks mimic real-world textures (graffiti, logos, clothing styles) to avoid raising suspicion from humans or surveillance monitors.
- **Transferability**: Robust attacks aim to deceive multiple models, across architectures and training datasets, enhancing their practical threat level.
- Advanced generation techniques (e.g., Expectation Over Transformation, Thin Plate Spline deformation, and naturalistic patch synthesis using GANs) enhance **robustness** under dynamic conditions and viewing transformations.

### D. Implications for Safety-Critical Systems

With life-threatening implications, especially for AI-based solutions that are being used in Safety-Critical Systems and infrastructures, physical adversarial attacks pose a significant practical threat as they deceive deep learning systems by producing prominent and maliciously designed physical perturbations.

- Autonomous Vehicles: Adversarial traffic signs can cause misinterpretation, leading to navigation failures or accidents.
- Surveillance and Law Enforcement: Attackers can evade detection or impersonate identities, undermining public safety.
- Biometric Access Systems: Adversarial masks or glasses can bypass facial authentication systems, granting unauthorized access [13].

As these attacks become increasingly sophisticated, they erode trust in AI-powered security infrastructure.

## IV. Conclusions

Physical adversarial attacks mark a dangerous evolution of adversarial machine learning—stepping out of simulated environments and into the physical world. Their stealth, accessibility, and growing efficacy demand immediate attention from the security and AI communities. Securing perception models in real-world settings is no longer optional—it is a critical necessity for the future of safe, trustworthy AI.

Defending against physical adversarial threats remains an open challenge. Promising avenues include:

- Real-world-aware training with adversarial augmentation;
- Cross-modal verification (e.g., combining visual and IR data); [14].
- On-device anomaly detection to flag improbable visual patterns;
- Using the feature visualisation and explainable AI in realation to adversarial samples.

Moreover, standardizing benchmarks for physical adversarial robustness and integrating simulation-to-reality pipelines during model training will be crucial.

## References


[1] Amira Guesmi, Muhammad Abdullah Hanif, B. Ouni, and Muhammed Shafique, "Physical Adversarial Attacks for camera based smart Systems," arXiv.org, vol. 11, 2023, doi: 10.48550/arxiv.2308.06173.

[2] Hikmat Khan, Pir Masoom Shah, Syed Farhan Alam Zaidi, and S. M. Fakhrul Islam, "Susceptibility of Continual Learning Against Adversarial Attacks," ArXiv, 2022, doi: 10.48550/arxiv.2207.05225.

[3] C. Olah, A. Mordvintsev, and L. Schubert, "Feature Visualization," Distill, 2017. [Online]. Available: https://distill.pub/2017/feature-visualization/

[4] Nguyen, Anh and Yosinski, Jason and Clune, Jeff. (2019). Understanding Neural Networks via Feature Visualization: A survey. 10.48550/arXiv.1904.08939.

[5] Fel, Thomas, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley et al. "Unlocking feature visualization for deep network with magnitude constrained optimization." Advances in Neural Information Processing Systems 36 (2023): 37813-37826.

[6] Vadillo, J., Santana, R. and Lozano, J.A. (2025), Adversarial Attacks in Explainable Machine Learning: A Survey of Threats Against Models and Humans. WIREs Data Mining Knowl Discov, 15: e1567. https://doi.org/10.1002/widm.1567

[7] D. Karmon, Daniel Zoran, and Yoav Goldberg, "LaVAN: Localized and Visible Adversarial Noise," International Conference on Machine Learning, vol. abs/1801.02608, 2018.

[8] Heemin Kim et al., "RAPID: Robust multi-pAtch masker using channel-wise Pooled varIance with two-stage patch Detection," Journal of King Saud University: Computer and Information Sciences, vol. 36, 2024, doi: 10.1016/j.jksuci.2024.102188.

[9] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in European conference on computer vision. Springer, 2020, pp. 665–681.

[10] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, "Making an invisibility cloak: Real world adversarial attacks on object detectors," in European Conference on Computer Vision. Springer, 2020, pp. 1–17.

[11] R. Daimo and S. Ono, "Projection-based physical adversarial attack for monocular depth estimation," IEICE TRANSACTIONS on Information and Systems, vol. 106, no. 1, pp. 31–35, 2023

[12] Hui Wei, Hanxun Yu, Kewei Zhang, Zhixiang Wang, Jianke Zhu, and Zheng Wang, "Moiré Backdoor Attack (MBA): A Novel Trigger for Pedestrian Detectors in the Physical World," Proceedings of the 31st ACM International Conference on Multimedia, 2023, doi: 10.1145/3581783.3611910.

[13] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in ICPR. IEEE, 2021, pp. 819– 826.

[14] Taeheon Kim, Hong Joo Lee, and Yong Man Ro, "Map: Multispectral Adversarial Patch to Attack Person Detection," IEEE International Conference on Acoustics, Speech, and Signal Processing, 2022, doi:10.1109/icassp43922.2022.9747896.