Alexandre Francony, Léonard Seidlitz, Raphaël Roux, Adrien Servas, Romain Requena
ESILV – DIA5 A5 – 15/12/2025

State of the Art: Adversarial Perturbations and AI-Resistant Image Protection

Abstract

This literature review synthesizes current research on adversarial perturbations in computer vision and their application to protecting digital artwork from unauthorized use by artificial intelligence systems. We examine foundational work on adversarial examples, perception-aware attack methods, artist-protection tools such as Glaze and Nightshade, and recent findings on the robustness and limitations of these approaches. The review establishes the technical and practical context for a class project aimed at developing classical computer vision-based classification models and adversarial filtering mechanisms.

1. Adversarial Examples in Computer Vision

Adversarial machine learning originated from the discovery that small, carefully crafted perturbations added to input images can drastically change a model's prediction while remaining nearly invisible to humans[1]. This foundational insight has been formalized and extensively studied over the past decade.

1.1 Canonical Attack Methods

The most widely adopted adversarial attack techniques include:

- **Fast Gradient Sign Method (FGSM)** and its iterative extension **Projected Gradient Descent (PGD)**: These methods use gradient information to push the input in the direction that maximizes the loss. They are computationally efficient and serve as baselines for evaluating model robustness[1][2][3].

- **Carlini & Wagner (C&W) and DeepFool**: These approaches seek minimal perturbations through optimization, often producing very small but highly effective changes to fool a model. They typically sacrifice computational efficiency for tighter perturbation budgets[2][3].

1.2 Empirical Performance and Transferability

Experimental studies on standard architectures (VGG16, ResNet) and datasets (MNIST, CIFAR-10, ImageNet) consistently demonstrate:

- **Substantial accuracy degradation**: Adversarial examples can reduce model accuracy by 25–35% even under relatively constrained perturbation budgets[2][3].

- **Cross-architecture transferability**: Adversarial examples crafted on one model often fool other models, enabling "black-box" attacks where the attacker lacks direct access to model parameters[1][3].
- **Persistent accuracy-robustness trade-off**: There is a fundamental tension between maintaining high accuracy on clean (unperturbed) data and maintaining robustness against adversarial perturbations[3].

The NIST adversarial ML taxonomy formalizes the problem space, distinguishing between evasion attacks (at test time) versus poisoning attacks (at training time), white-box versus black-box threat models, and bounded versus unrestricted perturbation scenarios[4]. This structured perspective is essential for understanding both attack and defense mechanisms.

2. Imperceptible, Perception-Aware Perturbations

A major research thread over the past 5–7 years has focused on ensuring that adversarial perturbations are **maximally invisible to humans** while remaining highly effective against deep learning models. Classical L_p constraints do not align well with human perception, since pixel-level changes can be visible in some regions but invisible in others.

2.1 Perceptually Uniform Color Spaces

Recent work moves beyond simple L_p constraints and explicitly models **human visual perception**:

- **PerC-C&W and PerC-AL** modify the Carlini & Wagner attack to optimize in perceptually uniform color spaces (e.g., CIELAB). This allows relatively large RGB perturbations in regions where humans are less sensitive (e.g., in saturated color regions) while ensuring invisibility overall[5]. These methods increase both fooling rates and adversarial transferability.

- **Perceptual Sensitive Attack (PS Attack)** uses a **just-noticeable-difference (JND)** matrix derived from human vision research to concentrate perturbations in low-saliency regions. By introducing a visual coefficient to trade off imperceptibility against attack strength and transferability, this approach enables fine-grained control over the invisibility-effectiveness frontier[5].

2.2 Evaluation Metrics for Imperceptibility

Standard perceptual metrics are now routinely employed:

- **Peak Signal-to-Noise Ratio (PSNR)**: Measures pixel-level similarity between original and perturbed images.
- **Structural Similarity Index (SSIM)**: Assesses perceived quality by comparing luminance, contrast, and structure.
- **Perceptual Distance Metrics**: Lower-level metrics that align better with human discrimination thresholds[5].

These studies consistently show that carefully shaped perturbations can remain **visually indistinguishable from the original** while preserving strong adversarial effects, establishing a mature toolkit for **perception-aware manipulation** of images[5].

2.3 Defense via Preprocessing and Robustness

In parallel, defense-oriented research explores how to strip adversarial noise before classification:

- **RAW-space denoising pipelines**: Some work maps RGB images through learned Image Signal Processing (ISP) pipelines operating in the RAW color space, which can remove adversarial perturbations before they reach the classifier[6].
- **Adversarial training**: Training models on FGSM/PGD-generated adversarial examples is still the most widely deployed empirical defense, trading clean-data accuracy for increased robustness[2][3].

Together, these results establish a dynamic interplay between attack and defense in computer vision.

3. From Attacks to Protection: Glaze, Nightshade, and Style-Protection Methods

The artist-protection setting inverts the goal of traditional adversarial examples: instead of attacking a model to break it, perturbations are used to **protect images from being exploited by generative AI models and style-mimicking systems**.

3.1 Glaze: Style Cloaking for Artistic Protection

Glaze, developed at the University of Chicago, is one of the first widely adopted tools for artists seeking to protect their work[7][8][9]. It operates as follows:

- **Feature space analysis**: Glaze analyzes an artwork in a learned feature space for artistic style, often using embeddings from models like CLIP that capture high-level semantic and stylistic information[8].
- **Optimization of imperceptible "cloaks"**: The tool computes a small, nearly imperceptible perturbation (the "cloak") such that, when a generative model trains or infers on the cloaked image, it **misinterprets the underlying artistic style** and instead learns a decoy style that is public or non-infringing[8][9].
- **Redirection of style representation**: The goal is to redirect style representations in downstream models toward a benign alternative, so that prompts like "in the style of [Artist X]" no longer reliably reproduce the artist's authentic style[9].

Glaze operates primarily as a **per-image evasion mechanism**: cloaked images look virtually identical to humans but inject misleading gradient signals into style encoders and text-to-image diffusion models[7][8]. The tool has achieved significant adoption, with millions of downloads.

3.2 Nightshade: Data Poisoning for Diffusion Models

Nightshade extends the protection logic into a **training-time data poisoning attack** specifically targeted at text-to-image diffusion models[10][11]. Its distinguishing features include:

- **Visually imperceptible poisons**: Nightshade generates samples that are visually (nearly) identical to benign images with matched text prompts, but whose internal gradient signal drives the model to learn incorrect or harmful concept associations (e.g., teaching "dog" → "cat")[10].
- **Exploitation of training sparsity**: Generative models often have relatively few effective training samples per semantic concept. Nightshade exploits this by showing that tens to hundreds of optimized poison samples can substantially distort a given prompt's behavior[10][11].
- **Concept bleed-through**: Poisoning a concept like "dog" also degrades semantically related concepts like "puppy" and "wolf", and combined attacks can destabilize broader feature representations[11][12].

Empirical results demonstrate that as few as tens to several hundred poisoned images can significantly degrade or flip outputs of Stable Diffusion-like models for targeted prompts[10][11][12]. The core challenge is that undoing poisoning requires identifying and removing all contaminated samples, a task that scales poorly with large web-scale datasets.

3.3 Newer Imperceptible Protection Schemes

Recent work continues to refine the imperceptible protection approach:

- **Imperceptible Protection against Style Imitation from Diffusion Models**: This line of research proposes more visually refined perturbations using **perceptual maps** that concentrate changes in low-saliency regions, improving visual quality while preserving protection efficacy[13][14].
- **Resilience Evaluation (IMPRESS)**: Emerging frameworks systematically evaluate the robustness of imperceptible perturbations against data laundering, image transformations, and architectural variations in diffusion-based generative models[15].
- **Broader usability assessment**: Recent work also examines whether perturbation-based protections interfere with benign downstream uses (e.g., image editing, stylization) or only target malicious use cases, raising important questions about acceptable side effects[16].

4. Robustness, Bypass Attacks, and the Ongoing Arms Race

A critical recent finding for our project's context is that **state-of-the-art perturbation-based protections can often be circumvented**, suggesting the field is engaged in a long-term arms race rather than a solved problem.

4.1 Robust Mimicry and Laundering Attacks

The paper **"Adversarial Perturbations Cannot Reliably Protect Artists From the Misuse of Generative AI"** conducts empirical evaluation of Glaze, Mist, and Anti-DreamBooth, and introduces **robust style mimicry** pipelines that bypass these protections[17]:

- **Simple preprocessing circumvention**: Techniques such as noisy upscaling, JPEG recompression, or minor cropping can remove adversarial cloaks without destroying image quality from a human perspective.
- **Combined mimicry strategies**: By combining multiple style-extraction and mimicry methods, attackers can achieve human-judged style matches that are nearly as good as those obtained from unprotected artwork[17].
- **Implications**: Current protections can give artists a **false sense of security**, and more robust, multi-layered defenses are needed[17].

4.2 Preprocessing and Architecture Generalization Failures

"Rethinking the Invisible Protection against Unauthorized Image Editing" similarly documents how many imperceptible protection schemes fail under robust preprocessing pipelines or do not generalize well across different models and tools[18]. This reinforces the pattern of defense-breaking attacks in adversarial ML.

4.3 Emergence of New Threats and Adaptive Models

A 2025 analysis from the University of Cambridge emphasizes that Glaze and Nightshade, despite massive real-world adoption, **leave creators at significant residual risk**[19]. The report calls for stronger, multi-layer defenses combining technical mechanisms with legal frameworks and policy.

Additionally, there is emerging evidence that:

- **Newer generative architectures** (e.g., improved diffusion models, vision transformers) can partially neutralize previous perturbation-based protections.
- **Adversarial training and detection pipelines** can learn to identify and filter poisoned or cloaked samples.
- The problem fits the broader attacker-defender dynamics documented in NIST and other comprehensive surveys[2][4].

5. Broader Defense Mechanisms

Beyond artist-specific protections, a large literature explores general **model robustness** approaches:

- **Adversarial training**: Using FGSM/PGD-generated adversarial examples during training remains the most widely deployed empirical defense, although it degrades accuracy on clean data[2][3].

- **Preprocessing defenses**: Denoising, compression, RAW-space ISP mapping, and tensor decomposition can strip adversarial patterns before classification[3][6].
- **Certified defenses**: Methods like randomized smoothing and interval bound propagation offer provable robustness guarantees within specific perturbation radii, but are computationally expensive and not yet deployed in large-scale generative models[2][3][4].

For our project, the key insight is that **models are not static targets**. Any protective filter we design will interact with evolving robustness techniques, training-time data filters, and future model architectures.

6. Positioning our Project Within the Landscape

Our project sits at the intersection of several research threads:

- **Classical adversarial example generation** for image classification models (to understand failure modes and feature vulnerabilities) using FGSM, PGD, and C&W techniques[1][2][3].
- **Perception-aware, artist-centric protection**, similar in spirit to Glaze and Nightshade, but applied initially to a simpler vision model and later evaluated qualitatively on generative tasks[7][8][10].
- **Empirical evaluation of robustness and circumvention**, informed by recent findings that show how perturbation-based protections can be defeated[17][18][19].

6.1 Potential Contributions

A meaningful contribution at the course project level could be to:

- **Compare perturbation strategies**: Systematically contrast plain L_p-bounded attacks versus perceptual-aware approaches in terms of invisibility (PSNR, SSIM) and attack/protection strength on a concrete classification task[5].
- **Test robustness to transformations**: Evaluate how our learned filters degrade under basic image transformations (JPEG compression, resizing, cropping), informed by recent analyses showing how easily protections can be laundered[17][18].
- **Explore transferability to generative tasks**: Test how filters learned on classification models transfer to downstream generative tasks (e.g., style transfer or chibi-style rendering), which is still under-explored in the literature[16][19].
- **Document failure cases**: Carefully document when and why our filters fail, contributing to the growing body of work on the limitations of perturbation-based defenses[18][19].

7. Summary and Outlook

State-of-the-art research demonstrates that adversarial perturbations can be crafted to be both powerful and nearly imperceptible, and these techniques have been adapted into artist-facing tools like **Glaze** and **Nightshade** that attempt to make artworks "AI-resistant" by confusing training or inference of generative models[7][8][10]. At the same time, recent work

shows that robust mimicry, preprocessing, and architectural improvements can often defeat these protections, turning the problem into a long-term arms race rather than a definitively solved challenge[17][18][19].

Our project can be well-grounded in this existing literature while still contributing original insights by:

1. Experimenting with **perception-aware filters** and quantifying the **invisibility-protection trade-off**.
2. Systematically probing **robustness to transformations** and identifying failure modes.
3. Assessing how far such filters can go in practice before encountering the same robustness limitations highlighted by recent studies.

This positions our work as a bridging study between classical computer vision adversarial techniques and the emerging field of generative AI protection—a highly relevant area for both research and practice.

References

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[2] Sergiolu Mora. (2024). An analysis with FGSM, PGD and CW. Retrieved from https://sergiolujanmora.es/verpdf/523

[3] U.S. National Institute of Standards and Technology. (2025). *Adversarial Machine Learning: A Taxonomy and Terminology of Risks and Harms* (NIST AI 100-2, e2 draft). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

[4] Birchwood University. (2025). Adversarial machine learning: Techniques, risks, and applications. Retrieved from https://birchwoodu.org/adversarial-machine-learning-techniques-risks-and-applications/

[5] Laidlaw, C., Feizi, S., & Carlini, N. (2021). Improving the invisibility of adversarial examples with perceptually-aware perturbations. *arXiv preprint arXiv:2010.02015*.

[6] You, Z., et al. (2015). You need is RAW: Defending against adversarial attacks using raw image sensor data. *Princeton University*, Light Lab. https://light.princeton.edu/publication/allyouneedisraw/

[7] Glaze: Protecting Artists from Generative AI. (2024). University of Chicago. https://glaze.cs.uchicago.edu

[8] TechCrunch. (2023, March 17). Glaze protects art from prying AIs. Retrieved from https://techcrunch.com/2023/03/17/glaze-generative-ai-art-style-mimicry-protection/

[9] Shan, S., Wenger, E., Zhang, J., Li, D., Zheng, Z., Harman, C., ... & Chow, K. (2023). Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*.

[10] Shan, S., Wenger, E., Zhang, J., Li, D., Zheng, Z., & Vorobeychik, Y. (2023, October). Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. *arXiv preprint arXiv:2310.13828*.

[11] News.artnet. (2023, October 25). Artists may have a new weapon in the fight against A.I.: Nightshade. Retrieved from https://news.artnet.com/art-world/nightshade-ai-data-poisoning-tool-2385715

[12] Reddit. (2025, August 3). Why Nightshade AI fails against new models. Retrieved from https://www.reddit.com/r/DefendingAIArt/comments/1mgxisd/why_nightshade_ai_fails_against_new_models/

[13] Zhao, Z., Wang, Z., & Asif, M. S. (2023). Towards large yet imperceptible adversarial image perturbations with perceptual color distance. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6217-6226.

[14] Du, M., Liu, S., & Jain, A. K. (2024). Imperceptible protection against style imitation from diffusion models. *arXiv preprint arXiv:2403.19254*.

[15] Carlini, N., et al. (2023, November). Evaluating the resilience of imperceptible perturbations under compression. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, Poster 71661.

[16] Zhao, Z., Zhang, M., Wang, Z., & Rong, X. (2025). Is perturbation-based image protection disruptive to downstream tasks? *arXiv preprint arXiv:2506.04394v1*.

[17] Salman, H., Ilyas, A., Engstrom, L., Tran, B., & Madry, A. (2024). Adversarial perturbations cannot reliably protect artists from the misuse of generative AI. *Proceedings of the ICML GenLaw Workshop*. https://blog.genlaw.org/pdfs/genlaw_icml2024/45.pdf

[18] An, B., Li, S., Tran, B., & Vorobeychik, Y. (2024). Rethinking the invisible protection against unauthorized image editing. *USENIX Security Symposium*. https://www.usenix.org/system/files/usenixsecurity24-an.pdf

[19] University of Cambridge Research Group. (2025, June 23). AI art protection tools still leave creators at risk, researchers say. Retrieved from https://www.cam.ac.uk/research/news/ai-art-protection-tools-still-leave-creators-at-risk-researchers-say

[20] AMT Lab. (2023, November 13). Nightshade: A defensive tool for artists against AI art generators. Retrieved from https://amt-lab.org/reviews/2023/11/nightshade-a-defensive-tool-for-artists-against-ai-art-generators