SPECIOUS - Spectral Perturbation Engine for Contrastive Inference Over Universal Surrogates

Dhruv Kumar   
*Department of Artificial Intelligence and Machine Learning*   
*Guru Gobind Singh Indraprastha University, Delhi 110032*  
pvt.dhruvkumar@gmail.com

*Abstract*— Generative AI models trained on large uncurated image corpora often appropriate artists’ copyrighted work without consent. We introduce SPECIOUS (Spectral Perturbation Engine for Contrastive Inference Over Universal Surrogates), a universal, multi-model defensive technique that embeds imperceptible high-frequency perturbations into the luminance (Y) channel of YCbCr representations. At inference time, these perturbations remain invisible to humans but systematically degrade feature embeddings across multiple surrogate models (ResNet-50, CLIP ViT-B/32, etc.), preventing downstream generative or classification tasks from reproducing the protected style. Our U-Net-based generator works in the Fourier domain, with a learnable high-pass mask, while our novel SpeciousLoss simultaneously minimizes LPIPS perceptual distance and maximizes the surrogate model’s feature distortion under a strict perceptual threshold. We demonstrate that S.P.E.C.I.O.U.S. effectively fools zero-shot predictions and classifications, generalizes across resolutions, and preserves visual fidelity.

Keywords—Universal Adversarial Perturbations, Frequency-Domain Defense, YCbCr Luminance Perturbation, Learned Perceptual Image Patch Similarity (LPIPS), U-Net-Based Fourier Perturbation

# Introduction

The recent “Ghiblification” phenomenon, more like a social media trend—in which users transform images into a Studio Ghibli style via ChatGPT’s image model [1][2][3]—has reignited urgent debates about copyright and artist rights in AI-generated art. Studio Ghibli itself has weighed legal action against OpenAI [4], but trademark and copyright statutes offer only partial recourse because “visual style” per se often falls outside traditional infringement claims. Meanwhile, AI companies continue to scrape public and private image repositories—often including copyrighted works—without artists’ consent [5][6], allowing models like DALL·E 2, Midjourney, Stable Diffusion, etc, to reproduce distinctive styles in seconds. Authors and illustrators worldwide report moral injury, as their life’s work is co-opted without credit or compensation [7].

Adversarial perturbations offer a pathway for self-defence [8]: small, model-imperceptible modifications that disrupt downstream inference. Empirical studies demonstrate that such perturbations concentrate in high-frequency components (sharp edges and textures)—that deep networks exploit to build feature representations [9][10], while leaving low-frequency content (soft edges and textures) intact. Furthermore, perturbations applied in the YCbCr color space show that changes in chrominance (Cb, Cr) are less salient to humans, but the most effective adversarial distortions actually arise in the luminance (Y) channel [11].

## Limitations of Existing Defences

Although tools like Glaze and Nightshade represent important steps toward protecting artists, they exhibit several key shortcomings that hinder their broader applicability:

### Model-Specific Targeting: Both Glaze and Nightshade focus narrowly on Stable Diffusion–style pipelines rather than offering a general defence. Glaze’s perturbations are designed specifically to mislead diffusion-based generators (e.g., Stable Diffusion or Midjourney) and may not transfer to other architectures or tasks [12][13]. Nightshade likewise tailors its poisoning attack to prompt-specific vulnerabilities in text-to-image diffusion models, requiring intimate knowledge of a single target model’s training data and concept sparsity [14]. In contrast, SPECIOUS can incorporate any number of surrogate models—from ResNet-50 classifiers to CLIP encoders—via its dynamic loss function, making it universal across architectures [15].

### Dependence on a Target Label: Nightshade’s prompt-specific poisoning hinges on choosing exactly which concept or class to corrupt [16]. You must supply a target label that directs the model’s latent representation into an unwanted state. Glaze similarly bakes in perturbations aimed at derailing the model’s style mimicry for a predefined artist. SPECIOUS, however, functions as a label-agnostic perturbation engine [17]: it crafts universal changes that degrade feature embeddings broadly, without needing to specify any particular downstream label or prompt.

### Single-Metric Penalty: Both prior defences apply only an LPIPS penalty [18] to control perceptual visibility, so they minimize human-noticeable changes but do not actively push feature representations apart. SPECIOUS’s SpeciousLoss, by contrast, is a bi-objective function: it simultaneously minimizes LPIPS and maximizes distortion in multiple surrogate feature spaces. This balanced approach guarantees both stealth and efficacy, rather than sacrificing one for the other.

### Neglect of Y-Channel Perturbations: Glaze and Nightshade operate solely in the RGB pixel domain [19], dispersing small cloaks uniformly across red, green, and blue channels. Yet perceptual and adversarial research shows that the luminance (Y) channel in YCbCr carries the most potent perturbations—human vision is less sensitive to chrominance, but deep networks leverage brightness variations heavily. SPECIOUS exploits this by concentrating all changes in the Y channel, preserving color fidelity while attacking the model-relevant signal.

### Overlooking Frequency-Domain Structure: Neither Glaze nor Nightshade uses frequency-domain filtering [20], neglecting the fact that adversarial deltas predominantly occupy high-frequency bands—edges and textures—that classification and generative models rely on. SPECIOUS embeds a learnable high-pass mask in the Fourier domain, adaptively isolating precisely those spectral regions where perturbations inflict maximum damage, ensuring minimal perceptual impact.

Taken together, these limitations underscore why a universal, multi-model defence—one that works at inference time, requires no specific label, balances perceptual and feature distortion, targets only the Y channel, and leverages frequency analysis—is essential for protecting creative content in the age of generative AI.

## Our Approach: SPECIOUS

We propose **SPECIOUS** (Spectral Perturbation Engine for Contrastive Inference Over Universal Surrogates), which integrates four key innovations:

### **Frequency-Domain Perturbation on Y Channel:** We transform only the Y channel into the Fourier domain [11][21], apply a learnable high-pass mask to isolate sharp edges and textures, then invert back to the spatial domain. This focuses perturbations on features that both classification (ResNet-50) and zero-shot (CLIP ViT-B/32) models rely on.

### **U-Net Generator with FFT/IFFT Blocks:** A bespoke U-Net architecture [22] ingests the high-frequency component and outputs a single-channel perturbation, ensuring the capacity to learn complex spatial patterns while maintaining low inference cost.

### **Specious Loss: Joint Perceptual and Feature Distortion Objective:** To train this, we designed SpeciousLoss, which balances two goals:

#### Perceptual fidelity, by minimizing LPIPS (a learned measure of human-perceived similarity) under a strict threshold.

#### Model disruption, by maximizing the difference in feature representations extracted by several pre-trained networks.

An exponential formulation keeps the loss positive and smoothly balances these pressures, while a penalty term ensures we never exceed our perceptual budget.

### **Universal Multi-Model Defence:** By training against multiple surrogate models simultaneously, SPECIOUS generates perturbations that generalize across architectures and tasks, unlike prior single-model attacks [15][24].

We demonstrate that SPECIOUS preserves visual fidelity while achieving high fooling rates (> 70%) on both classification and zero-shot generative benchmarks.

# Related Work

1. **Style-Cloaking and Dataset-Poisoning Attacks**

**Glaze’s Style Cloaks**  
Glaze crafts tiny RGB perturbations (“style cloaks”) that artists can apply before sharing images, so that diffusion models trained on these cloaked versions learn a skewed style representation rather than the true artistic style. In a large‐scale user study involving over 1,000 professional illustrators[12][13], surveyed artists found the visual changes acceptable, and automated CLIP‐based metrics confirmed **over 92% disruption** of style mimicry under normal conditions and still **above 85%** under adaptive counter-attacks. However, Glaze’s effectiveness hinges on models fine-tuned on cloaked data and does not extend to inference-time scenarios.

**Nightshade’s Prompt-Specific Poisoning**  
Nightshade takes a data-poisoning route[14] by injecting carefully crafted image–text pairs into the Stable Diffusion training set, such that fewer than **100 poisoned samples** can hijack a model’s response to a specific prompt (e.g., making “a cat” generate flowers instead). These poison images are visually indistinguishable from authentic samples and can even “bleed through” to semantically related prompts. Yet because Nightshade requires **access to and modification of the training pipeline**, it cannot defend images shared against black-box, inference-only attacks.

1. **Frequency-Domain Adversarial Perturbations**

Early work revealed that gradient-based adversarial deltas concentrate in **high-frequency bands** [9][25]—the fine edges and textures that convolutional networks heavily weight. Building on this, researchers have embedded **spectral filtering layers** inside defences to emphasize low-frequency content for robust recognition, effectively “damping” model sensitivity to perturbations. Other defences apply **input transformations in the Fourier domain**, such as mixing or removing select frequencies, to neutralize adversarial noise before classification. These techniques underscore the power of frequency analysis, but most target single architectures or rely on fixed, hand-crafted filters rather than learnable, adaptive masks.

1. **Perceptual Metrics in Adversarial Optimization**

To achieve “stealthy” attacks, perceptual fidelity is critical. The **Learned Perceptual Image Patch Similarity (LPIPS)** metric[18] aligns deep feature distances with human judgments, mapping differences onto a **0–1 scale**. LPIPS has been used both as an evaluation metric and as a **loss term** during adversarial example generation, ensuring perturbations remain below human-noticeable thresholds. The **Perceptual Sensitive Attack** further integrates LPIPS constraints to maximize downstream task disruption under tight perceptual budgets, illustrating how perceptual losses can guide more effective black-box attacks. However, existing methods seldom **jointly optimize** LPIPS and multi-model feature distortion under a strict threshold.

1. **Y-Channel Specific Attacks**

Pestana et al. demonstrated that adversarial perturbations “prevail” in the **luminance (Y) channel** [11] of YCbCr space, achieving higher fooling rates for the same perturbation budget compared to RGB-space attacks. Their **ResUpNet** defence uses a Y-channel de-noising network to remove adversarial noise (from FGSM, PGD, DDN) while leaving chrominance untouched, thereby preserving color fidelity. This work highlights how human sensitivity to chrominance changes is lower, making Y-focused perturbations more “efficient”—yet most RGB-based defences ignore this channel specificity.

1. This is a figure caption. It appears directly underneath the figure.

appropriate style is still applied to each section, reapplying styles if necessary.

#### This is another level 4 heading: It’s also possible to add bullet points when appropriate, using the “bullet list” style:

* Treat the word “data” as plural, not singular.
* For example, “the data indicate that …”

## This Is Another Level 2 Heading

Place any figures or tables you use at the top or bottom of a column. Don’t place them in the middle of a column. If particularly wide, a table or figure can span across both columns. Insert a table or figure after the point where it is first cited in the text.

When inserting a figure, such as a photograph or infographic, use 8 pt. Times New Roman for any labeling text within the image and for the figure caption. You can see an example of a figure caption in Fig. 1, above. Refer to figures like that, using the abbreviation “Fig.” and the figure’s number.

A table heading (using the “table head” style) appears above a table. This will automatically number the table for you. Any footnotes appear below the table, using the “table footnote” style. Footnotes are indicated by superscript lowercase letters within the table. An example of a table can be seen in Table I, below.

# This Is Another Level 1 Heading

All the headings in the main body of your paper are numbered (automatically).

Another type of heading is the “component heading”, which is used for other components that aren’t part of the main text. These are usually your acknowledgments and your references, which you can see examples of below. These headings are not numbered. The correct styling for them can be applied using the “Heading 5” style, which is the same as the “Heading 1” style but without numbering.

1. This Is the Heading for a Table
2. This is a table footnote.

You can cite your references in text by including the corresponding number, in square brackets [1]. If you need to cite a specific part of the source, you can include a page number [2, p. 13] or range [3, pp. 41–56].

##### Acknowledgments

“Acknowledgment(s)” is spelled without an “e” after the “g” in American English.

As you can see, the formatting ensures that the text ends in two equal-sized columns rather than only displaying one column on the last page.

This template was adapted from those provided by the IEEE on their own website.

##### References

1. D. V. Lindberg and H. K. H. Lee, “Optimization under constraints by applying an asymmetric entropy measure,” *J. Comput. Graph. Statist.*, vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.
2. B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.
3. I. Boglaev, “A numerical method for solving nonlinear integro-differential equations of Fredholm type,” *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.

**Make sure to remove all placeholder and explanatory text from the template when you add your own text. This text should not be here in the final version!**