# Synthesizing the Light That Defeats a Palette: Adversarial Illuminant Search Over Physically Realizable Spectra

Satyajeeth Suresh Kannan

*Innovation Academy · Mary's Lab*

github.com/JeetuSK0808/maryslab

Preprint · August 2026 · Mary's Lab technical report TR-09

**ABSTRACT**

Testing a palette under a list of illuminants finds the failures on that list. We instead search the space of physically realizable illuminants for the one that minimises a palette's worst pairwise separation. The objective is a minimum over pairs and therefore non-smooth — the identifying pair changes as the search moves — so we optimise with a (1+1) evolution strategy, which compares objective values and never differentiates them. Constraints keep the result physical: non-negative, bounded in total power, so the search cannot trivially darken the scene. On a three-surface palette the attack reduces the minimum separation from 0.351 to 0.1111 $\Delta E_{OK}$, a reduction of 0.684, and identifies the collapsing pair. Repeating the search under simulated deuteranopia yields a different and more damaging attack, which we argue makes observer-specific adversarial testing necessary rather than optional. Against a palette of measured USGS materials the relative damage is smaller (0.233) but the absolute outcome is worse — the attacked palette retains 0.0933 against 0.1111 for the designed one, and under deuteranopia it collapses to 0.0041. Percentage reduction is therefore the wrong way to report robustness; post-attack separation is the number that decides whether the reader can still do the task.

*Keywords:* adversarial search; illuminant design; palette robustness; evolution strategies; accessibility

## 1. Introduction

Machine learning made a methodological argument that generalises well beyond it: evaluating a system on a sampled test set understates its fragility, and searching for the input that defeats it reveals failure modes no sample contains. The argument applies directly to palettes under lighting.

A palette validated under three named illuminants has been validated under three illuminants. The question a designer actually needs answered is whether a light exists that breaks it — and if so, what that light looks like, since its plausibility determines whether the failure matters.

## 2. Method

The optimiser is a (1+1) evolution strategy: perturb the incumbent illuminant, keep the perturbation if the palette's minimum pairwise separation decreases, adapt the step size from the recent success rate. The choice is deliberate rather than default. The objective is a minimum over pairs, so as the search moves the identifying pair changes and the objective is non-smooth with kinks exactly where the argmin switches. Gradient methods handle that badly; a comparison-only method does not notice.

The illuminant is parameterised by a bounded basis expansion and constrained non-negative with bounded total power. Without the power constraint the search discovers the degenerate attack of turning the lights off, which collapses all separations and teaches nothing.

The search is seeded, so a reported attack is reproducible. It finds *an* adversarial illuminant, not provably the worst one.

## 3. Evaluation

**Inputs are synthetic.** The palette is three analytic Gaussian reflectances at 455, 550 and 640 nm, the same set used in TR-01, which allows the certified bound there and the attack here to be read together.

**Table 1.** Adversarial illuminant search against a three-surface palette, baseline D65. The observer is part of the threat model: repeating the search under simulated deuteranopia finds a different collapsing pair.

| observer | baseline min ΔE | attacked min ΔE | reduction | collapsed pair |
|---|---|---|---|---|
| normal trichromat | 0.351 | 0.1111 | 0.684 | 1-2 |
| deutan, full severity | 0.1723 | 0.0016 | 0.991 | 1-2 |

Two findings. First, a palette comfortably separated under daylight can be driven substantially closer by a light the search constructs, confirming that illuminant-list validation understates exposure. Second, the observer belongs in the threat model: the dichromatic search starts from a lower baseline — the palette is already weaker for that observer — and the attack it finds is not the same attack. A designer who hardens a palette against the trichromatic attack alone has not hardened it for the readers most at risk.

Read alongside TR-01, the two instruments bound the problem from both sides: the certificate states a guarantee that holds over a declared hull, and the attack demonstrates what can happen outside it. A palette that passes both is one whose behaviour is understood.

### *Reproducing these numbers*

Clone `github.com/JeetuSK0808/maryslab` and run `node papers/build/experiments.mjs`. Every figure in this report is written by that script into `papers/build/results.json`; the instrument itself is exercised by `synthesize_adversarial_spectrum`. The engines carry a 401-vector conformance suite shared by a TypeScript reference implementation, a C++20 core, and a WebAssembly build, so a reported number is a property of the specification rather than of one implementation.

## 4. Real-data evaluation: attacking a palette of measured materials

The palette attacked in §3 was built to be separable, which makes it the interesting case for an attack but not a representative one. We repeated the search against four measured reflectances drawn from snapshot `usgs-splib07a-1`, under the same budget and seed.

**Table 2.** Adversarial illuminant search against a palette of four measured USGS materials, baseline D65, with the synthetic palette of §3 for comparison.

| observer | baseline min ΔE | attacked min ΔE | reduction | synthetic reduction (§3) |
|---|---|---|---|---|
| normal trichromat | 0.1217 | 0.0933 | 0.233 | 0.684 |
| deutan, full severity | 0.1223 | 0.0041 | 0.966 | 0.991 |

Two things change and one does not. The baseline is much lower — 0.1217 against 0.351 for the designed palette — because four arbitrary minerals are simply not far apart, a result TR-03's real-data section quantifies directly. And the relative damage the attack does to a normal trichromat is correspondingly smaller, 0.233 against 0.684: there is less separation available to destroy, so the search has less room to work.

What does not change is the observer asymmetry, and it is sharper here. Against a dichromatic observer the attack drives the palette from 0.1223 to 0.0041 — a reduction of 0.966, which is essentially total collapse. The absolute attacked value is under 0.005 $\Delta E_{OK}$, meaning the four materials become, for that observer under that light, one color.

The comparison across the two tables makes a point neither makes alone. A designed palette is more attackable in relative terms and far safer in absolute ones: after the attack it retains 0.1111, while the material palette retains 0.0933. Percentage reduction is therefore a misleading way to report robustness, and the absolute post-attack separation is the number that decides whether a reader can still do the task. We report both for that reason.

## 5. Limitations

This section states what the result does not establish. It is written to be read before the abstract is quoted.

- **The search finds an attack, not the worst attack.** It is stochastic and seeded; a different seed may find a stronger one. Reporting the seed makes runs reproducible, not exhaustive.
- **"Physically realizable" is a constraint set, not a manufacturing survey.** A non-negative, power-bounded spectrum is not the same as a lamp somebody makes. The plausibility of a reported attack requires separate judgment.

- **Grid resolution under-models narrowband attacks.** At 10 nm the search cannot construct a laser line, so genuinely spiky adversaries are outside the representable set.
- **No claim of optimality or of coverage.** A palette surviving this search has survived this search.
- **Prior art was surveyed informally.** Adversarial example generation and illuminant optimisation are both established fields; a search found no tool searching illuminant space adversarially against palette discriminability with an observer in the threat model.

## 6. Acknowledgements and disclosure

The instruments described here were implemented in Mary's Lab, an open-source computational color science project. Software design, implementation, and the preparation of this report were carried out with substantial assistance from a large language model (Anthropic Claude), under the author's direction and review. All numerical results were produced by executing the released code; none were generated by a language model. The author is responsible for the claims made here.

## References

1. Szegedy, C., et al. (2014). Intriguing properties of neural networks. *Proceedings of ICLR*.
2. Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog.
3. Hansen, N. (2016). The CMA evolution strategy: a tutorial. *arXiv:1604.00772*.
4. Machado, G. M., Oliveira, M. M., and Fernandes, L. A. F. (2009). A physiologically-based model for simulation of color vision deficiency. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 1291–1298.
5. Ottosson, B. (2020). *A perceptual color space for image processing (Oklab)*. bottosson.github.io.
6. CIE (2004). *Colorimetry, 3rd edition*. CIE 15:2004. Commission Internationale de l'Éclairage, Vienna.
7. Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. *Proceedings of ICML*.
8. Wyszecki, G. and Stiles, W. S. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*, 2nd edition. Wiley.

Source, engines, and the script that produced every number in this report: `github.com/JeetuSK0808/maryslab`. Results regenerate with `node papers/build/experiments.mjs`. Inputs: synthetic where §3 says so, and measured USGS splib07a spectra (snapshot `usgs-splib07a-1`) where a real-data section says so. Prepared with AI assistance (see Acknowledgements).