A New Kind of Adversarial Example: Measuring the Human–Model Gap, and Its Relationship to OOD Detection

Ali Borji

BrainChip Inc., Laguna Hills, CA, USA

Abstract

Almost all adversarial attacks add an imperceptible perturbation to fool a model. We instead study the opposite: a large, clearly visible perturbation that causes the model to *keep* its original, correct prediction, even though a human would no longer recognize the image. Prior work showed such examples can be generated at scale but left three questions untested: whether humans really perform worse than the model, whether standard out-of-distribution (OOD) detection and calibration tools catch it, and whether existing defenses mitigate it. We answer all three on MNIST, CIFAR-10, and ImageNet. (i) An independent recognizer proxy drops to ∼49% on CIFAR-10 while the model stays at 100% — a gap a small human pilot (N=5) corroborates directly and that is not explained by signal loss (a matched-magnitude Gaussian control degrades recognizability faster); a CLIP zero-shot proxy confirms the gap at ImageNet scale too. (ii) Confidence- and energy-based OOD detectors and calibration are structurally blind (0% detection, ECE ≈ 0), while a feature-space Mahalanobis detector flags 100% — but is evaded by an

Email address: aliborji@gmail.com (Ali Borji)

adaptive attacker at no cost to success. (iii) No classical defense, including adversarial training (45% robust accuracy), reduces attack success (correlation with large-ϵ_l resistance $r \approx 0$). A mechanistic analysis further shows the attack destroys low-level texture far faster than edge/shape structure. Code, figures, and a ready-to-run human-study harness are available at `https://github.com/aliborji/NKE`.

Keywords: adversarial examples, out-of-distribution detection, calibration, adversarial robustness, human–model gap, CIFAR-10, ImageNet

---

1. Introduction

Deep neural networks are known to be highly sensitive to small, carefully crafted perturbations: adding an imperceptible amount of noise to an image is often sufficient to change a model's prediction, even though the perturbation leaves the image visually unchanged to a human observer [1, 2]. This phenomenon — the classical adversarial example — has motivated a large body of work on attacks, defenses, and theoretical explanations of adversarial vulnerability.

In earlier work [3], we considered the opposite failure mode: instead of a small perturbation causing a *wrong* prediction, a large, clearly visible perturbation can be constructed that causes a model to *keep* its original, correct prediction, even though a human presented with the same perturbed image would no longer be able to recognize it, or would be much more likely to misclassify it. We formalized this setting by inverting the standard adversarial constraint: rather than bounding the perturbation from above by a small ϵ_s and requiring the prediction to change, one bounds the perturbation from

below by a large ϵ_l and requires the prediction to remain unchanged,

$$x' : \ D(x, x') > \epsilon_l \ \wedge \ f(x') = y, \tag{1}$$

and shows that a simple iterative FGSM-style attack (dubbed NKE; **N**ew **K**ind on Adversarial **E**xample) can reliably produce such examples on MNIST and CIFAR-10. This formulation was motivated as a unification of two previously disconnected phenomena: adversarial examples in the classical sense, and the "fooling images" of Nguyen et al. [4], which are unrecognizable to humans yet classified with high confidence by a network. Because the same gradient-based machinery can produce both kinds of failure by simply flipping the sign and direction of the perturbation constraint, the two phenomena can be viewed as two ends of a single psychometric curve relating perturbation magnitude to accuracy — one where models should ideally track human performance, but in practice diverge from it in both directions, as we argued previously [3].

More recently, Nie et al. [5] independently arrived at essentially the same formulation — compare their Eq. 7, $\arg\min_{X^{adv}} J(X^{adv}, y_{true})$ s.t. $\|X^{adv} - X\|_p \geq \delta$, to Eq. 1 above — and extended it substantially in scope. They evaluate on ImageNet-scale models (Inception-v3/v4, Inception-ResNet-v2, ResNet-152), introduce an L_2-norm variant (NI-FGM) alongside the L_∞ variant, and add momentum-based versions of both (NMI-FGSM, NMI-FGM). They also run the first systematic transferability study of this attack family, finding a stark asymmetry: white-box success rates reach the 80–99% range, but black-box success rates collapse to below 5%, indicating that these examples are highly specific to the model that generated them and do not transfer the way conventional adversarial examples often do.

Taken together, our earlier work and Nie et al. [5]'s extension establish that (a) such examples can be generated efficiently and reliably at scale across model families, and (b) they behave very differently from classical adversarial examples with respect to transferability. What neither addresses, however, is whether the central premise motivating the whole formulation actually holds: **that a human forced to classify these images would in fact perform much worse than the model does.** The entire framing rests on an assumed but never measured gap between human and model psychometric curves that we introduced [3, Fig. 3]. Nor does either connect the phenomenon to the broader literature on out-of-distribution (OOD) detection and confidence calibration, despite the fact that an image which is far from the data manifold yet classified with high confidence is, by construction, exactly the kind of failure that OOD detectors and calibration methods are designed to catch. Finally, neither paper asks whether standard robustness interventions — adversarial training, input transformations, or randomized smoothing — have any effect on this failure mode, leaving open whether it is a distinct vulnerability or one already implicitly addressed by existing defenses.

This paper addresses these three gaps. Specifically, we:

1. Measure the human–model gap implied by the original formulation. Because large-scale crowd data could not be collected for this instantiation, we report two clearly-labeled computational *recognizability proxies*, *release a ready-to-run forced-choice human-study harness* (stimuli + task page), and run a small real-human pilot (N=5) with it (Sec. 3.3, Sec. 4.2). We find a large gap on CIFAR-10 that is specific to the adversarial structure and not to signal loss (on the proxy; the pilot cor-

roborates the gap itself but is underpowered for the signal-vs-structure contrast); at ImageNet scale this specificity is no longer separable by the cross-model proxy (Sec. 4.9).

2. Reframe the phenomenon in terms of OOD detection and calibration, and show that confidence/energy detectors and ECE are blind to it while a feature-space detector catches it entirely — but that even this feature-space detector is defeated by an adaptive attacker at no cost to attack success (Sec. 3.4, Sec. 4.3, Sec. 4.4).
3. Evaluate whether adversarial training, randomized smoothing, and input transformations mitigate this failure mode, and whether resistance to classical small-ϵ attacks correlates with resistance to large-ϵ_l attacks (Sec. 3.6, Sec. 4.7). It does not.

In doing so, we move the discussion from "can such examples be constructed" — which is by now well established — to "what do they reveal about the gap between model and human perception, and how do they relate to the tools the field already uses to detect distributional failure."

2. Related Work

Classical adversarial examples. The vulnerability of deep networks to small, imperceptible perturbations was first demonstrated by Szegedy et al. [1] and explained via the linear-behavior hypothesis by Goodfellow et al. [2], who introduced the fast gradient sign method (FGSM). Kurakin et al. [6, 7] extended this to an iterative variant (I-FGSM/BIM), improving white-box attack strength at some cost to transferability, a trade-off later characterized more broadly by Tramèr et al. [8]. Optimization-based attacks

such as Carlini and Wagner [9] instead treat the perturbation bound as a soft constraint via Lagrangian relaxation. Momentum-based iterative attacks [10], as adapted by Nie et al. [5], were introduced primarily to improve black-box transferability by smoothing the gradient signal across iterations and escaping poor local optima. All of this work shares the same basic constraint structure: perturbations are bounded *above* by a small ϵ, and the attack succeeds when the prediction changes.

Fooling images and inverted formulations. Nguyen et al. [4] showed that evolutionary search can produce images that are unrecognizable to humans yet classified with high confidence by a trained network, establishing that high-confidence predictions do not imply the input resembles the training distribution in any way a human would recognize. In earlier work [3], we reformulated this as a targeted, gradient-based attack in which the "target" class is simply the model's own correct prediction on the original image, and introduced the constraint structure in Eq. 1, framing it as the mirror image of the classical adversarial setting along the perturbation axis. Nie et al. [5] developed this direction further with L_2-norm variants and momentum-based extensions, and are — to our knowledge — the first to evaluate this attack family at ImageNet scale and to systematically measure its transferability across architectures, finding it to be substantially worse than that of classical adversarial examples.

Human comparison in adversarial robustness. A separate line of work has compared human and model behavior under adversarial or corrupted inputs [11, 12, 13], generally finding that models and humans diverge in systematic ways under both time-limited and unlimited viewing conditions.

This literature provides methodology for measuring human psychometric functions under controlled perturbation, but has not been applied to the large-perturbation, same-label setting studied here; the human-model gap central to our framing [3] (illustrated only conceptually in Fig. 3 there) has not been empirically measured for this specific attack family in either source paper.

Out-of-distribution detection and calibration. A large body of work addresses the related but distinct problem of detecting when a model is given an input unlike its training distribution, or of ensuring that a model's confidence is well-calibrated to its accuracy [14, 15, 16, 17]. This literature treats "confident but wrong, or confident on nonsensical input" as a first-class failure mode to be measured and mitigated, using tools such as Mahalanobis-distance-based detectors, energy-based OOD scores, and expected calibration error. Despite the clear conceptual overlap — a large-ϵ_l adversarial example is, definitionally, an input far from the natural data manifold that nonetheless receives a confident, correct prediction — neither our earlier work [3] nor Nie et al. [5] evaluate these examples against any OOD detection or calibration baseline.

Adversarial defenses and robustness. Adversarial training [2, 18] and its ensemble variants [8] remain the most widely used defenses against classical adversarial examples, alongside certified approaches such as randomized smoothing [19] and simple input-transformation defenses [20]. Whether robustness to small-ϵ perturbations transfers to robustness against large-ϵ_l, same-label perturbations is, to our knowledge, an open question.

3. Methods

Our study is organized around a single attack-generation pipeline that feeds three independent evaluation tracks plus a mechanistic analysis (Fig. 1). A single set of NKE-style images, generated at multiple perturbation magnitudes ϵ_l, feeds every track, so that all downstream analyses operate on identical stimuli.

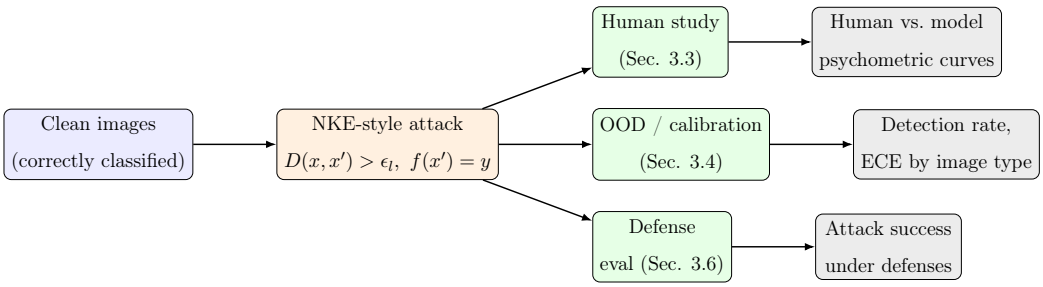

Figure 1: Overview of the experimental pipeline. A single set of NKE-style adversarial images, generated at multiple perturbation magnitudes ϵ_l, feeds independent evaluation tracks whose outputs are analyzed jointly in Sec. 4.

3.1. Attack generation

We generate adversarial examples satisfying Eq. 1 by minimizing the true-label loss (to keep the prediction correct) while an exterior projection holds the perturbation at L_2 distance $\geq \epsilon_l$ from the clean image:

$$x'_{N+1} = \Pi_{\|\cdot - x\|_2 \geq \epsilon_l}\Big[\,\mathrm{clip}_{[0,1]}\left(x'_N - \alpha\, g_N\right)\Big],$$

where g_N is the (optionally momentum-accumulated, sign- or L_2-normalized) gradient of $J(x'_N, y)$. Here $J(\cdot,\cdot)$ is the model's classification loss (cross-entropy between the predicted and true class), evaluated at the current iterate

x'_N against the original true label y; minimizing it, rather than maximizing it as in a classical attack, is what keeps the model's prediction correct throughout the attack. $x'_0 = x$ is the clean image, N indexes the attack iteration, and α is the step size. $\text{clip}_{[0,1]}(\cdot)$ clamps pixel values back into the valid image range after each gradient step. $\Pi_{\|\cdot - x\|_2 \geq \epsilon_l}$ is the exterior projection: whenever a step would leave the iterate inside the L_2 ball of radius ϵ_l around x, it pushes the iterate back out to the ball's boundary, enforcing the reversed, lower-bound perturbation constraint of Eq. 1 (the mirror image of the usual upper-bound projection used in classical adversarial attacks). We implement our earlier plain iterative form [3] and the L_2/momentum variants (NI-FGM, NMI-FGSM, NMI-FGM) of Nie et al. [5] in a shared codebase, and use NMI-FGSM/NMI-FGM (60 steps) as the primary stimulus generator, as momentum variants are the most reliable [5]. We restrict all tracks to images correctly classified before the attack and still classified with the original label after, per Eq. 1. We use two architecturally distinct models per dataset: LeNet/MLP on MNIST and ResNet-18/VGG-11 on CIFAR-10 (clean test accuracy 99.3/98.5% and 93.5/90.2% respectively).

3.2. Baseline replication

Before the novel tracks, we validate our pipeline against the phenomena reported by Nie et al. [5] at three scales: MNIST (LeNet, MLP), CIFAR-10 (ResNet-18, VGG-11), and **ImageNet** — using the ImageNet-pretrained ResNet-152 and VGG-16 (two architecturally distinct families) on *Imagenette*, a 10-class subset of real ImageNet images at 224×224. The full ImageNet training set is unavailable in our environment, but the NKE attack only requires correctly-classified natural images, which Imagenette provides; this

lets us measure Nie et al.'s ImageNet transfer collapse directly rather than cite it. We reproduce (i) near-perfect white-box success and (ii) the white-box≫black-box asymmetry at all three scales (Table 1), and reuse Nie et al.'s three ablation axes (perturbation size, iterations, momentum μ) at coarser resolution.

3.3. Human psychometric study

The central untested assumption, from our earlier work, is that human accuracy degrades faster than model accuracy as ϵ_l increases [3, Fig. 3]. As Borji [21] argues, L_p distance is not a reliable proxy for perceptual similarity, so this must be validated rather than assumed from the ϵ_l bound. **Protocol (released as a harness).** Participants view one image at a time and select the category from the dataset's label set, with an explicit "cannot tell" option; each image is shown to multiple raters, no participant sees the same base image at two perturbation levels, and a matched-budget Gaussian-noise control at the same L_2 distance is included to separate adversarial structure from generic signal degradation [addressing 21, 11, 13]. We export the full stimulus set + a self-contained forced-choice task page. The harness enforces the design client-side: each participant is assigned a between-subjects sample so that no base image is seen at two perturbation levels, and the clean (ϵ_l=0) images are interleaved as attention checks. Returned sessions are pooled and attention-filtered — participants whose clean-trial accuracy falls below a dataset-calibrated threshold are discarded, as the clean-image human ceiling on 32×32 CIFAR-10 is well below MNIST's — and we report human accuracy, "cannot tell" rate, and reaction time per ϵ_l for the NKE and Gaussian-control conditions, with bootstrap 95% confidence intervals. **Proxies reported**

here. In lieu of (not as a replacement for) crowd data, we report two labeled computational recognizability proxies: (a) a *generic recognizer* — the other architecture, trained independently and never exposed to the attack gradients; and (b) a *shape recognizer* — a classifier trained on Canny edge maps. These are proxies, not human data.

3.4. Out-of-distribution and calibration analysis

We evaluate three post-hoc OOD detectors — maximum softmax probability (MSP) [14], a Mahalanobis feature-distance detector [15], and an energy score [16] — reporting the fraction of NKE images flagged at a threshold calibrated to 5% false-positive rate on *held-out* clean data (disjoint from the split used to fit the detector; an in-sample threshold understates the true FPR for a fitted-covariance detector), alongside classical small-ϵ adversarials and genuine OOD (SVHN/FashionMNIST) as references. We also report the threshold-independent AUROC, and compute expected calibration error (ECE) [17] for clean, classical-adversarial, and NKE images.

3.5. Mechanistic analysis: shape vs. texture

Standard CNNs are texture-biased whereas humans rely on shape [22, 21]. We test whether the attack preferentially destroys texture while leaving edge/shape structure comparatively intact by computing, per image and ϵ_l, an edge-preservation score (Canny-edge-map F1, clean vs. perturbed) and a texture-similarity score: we compute the GLCM (a matrix tallying how often pairs of pixel gray-levels co-occur at a fixed offset, from which standard texture statistics such as contrast, homogeneity, and energy are derived) separately for the clean and perturbed image, and report the cosine similarity between their

GLCM feature vectors, so a score of 1 means texture statistics are unchanged and 0 means they are unrelated; plus, for comparison, a Gram-matrix cosine.

3.6. Robustness intervention evaluation

We re-run the white-box attack against (i) an adversarially trained model [18], (ii) randomized smoothing [19], and (iii) input-transformation defenses (JPEG, Gaussian blur, random resize/pad) [20], using straight-through gradients for non-differentiable transforms. We correlate each model's classical small-ϵ robust accuracy with its resistance to the large-ϵ_l NKE attack (operationalized as $1 - \overline{\text{success}}$ over $\epsilon_l > 0$).

4. Results

Across *all* conditions below, the source model's accuracy on the perturbed images is 100% at ∼1.0 confidence, out to perturbations as large as the image content itself ($L_2 \approx 11$ on MNIST, ≈ 16 on CIFAR-10). Figure 2 shows what these images look like: recognizable objects dissolve into colour noise as ϵ_l grows, yet the model's label and confidence are unchanged along every row (the full grid, and an ImageNet analogue, are in Fig. B.9, App. Appendix B). The NKE property is trivially satisfied; the interesting variation is in everything else.

4.1. Baseline replication and the transferability–complexity trend

White-box success is ≈100% and cross-model (black-box) success is far lower, replicating Nie et al.'s asymmetry on CIFAR-10 (Table 1). Ablations are flat at 1.0 across perturbation size and $\mu \in [0, 1]$, and across iterations ≥ 10 (5 iterations $\rightarrow$ 0.875 on CIFAR-10). Across all three scales, measured

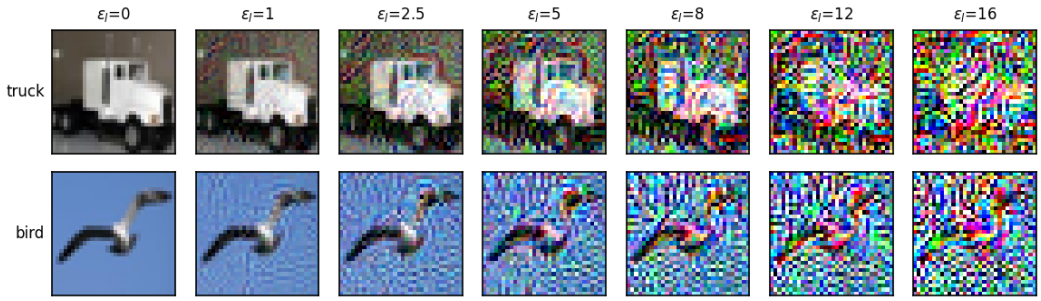

Figure 2: NKE examples (CIFAR-10, ResNet-18): clean (ϵ_l=0) → NKE across increasing ϵ_l. The model keeps the original label at ∼1.0 confidence across each entire row, even where the image has dissolved into colour noise no human could label. Full grid in Fig. B.9.

on real data, a clean trend emerges (Fig. 3): **NKE transferability falls as task complexity rises** — MNIST ≈0.97 → CIFAR-10 ≈0.36 → ImageNet ≈0.10 (and →0.00 at larger ϵ_l), the last measured with ResNet-152/VGG-16 on Imagenette rather than cited. On simple data the perturbation preserves class-defining structure that generalizes across models; on complex data it becomes model-specific. Our simple-dataset regime [3] and Nie et al.'s ImageNet regime are two ends of one continuum.

4.2. The human–model gap is real on natural images (and, on the proxy, specific to adversarial structure)

Table 3 reports the proxies (summary overlay in Fig. 4). On CIFAR-10, the independent generic recognizer falls to **48.5%** at $\epsilon_l = 16$ while the source model stays at 100% — a ∼50-point gap, the first quantitative (proxy) evidence for the gap the original formulation assumed. Crucially, a **matched-L_2 Gaussian control degrades the generic recognizer faster** (to ∼0.15) than the structured NKE perturbation (to 0.49). The control is constructed per image and per ϵ_l: we draw i.i.d. Gaussian noise, rescale it so its L_2 norm

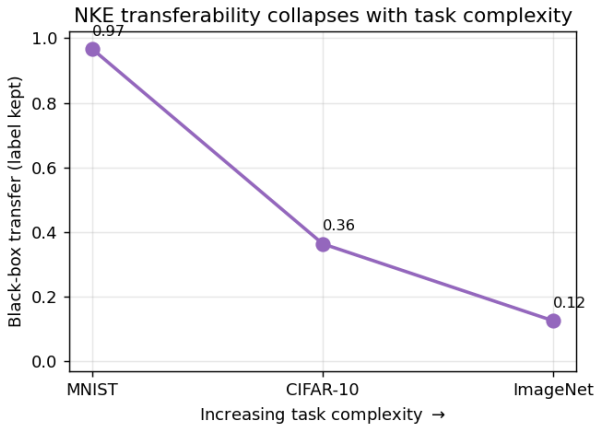

Figure 3: Black-box transfer (label kept) vs. task complexity, measured on all three scales (ImageNet at ϵ_l=30, $L_2 \approx 61$). Transferability collapses as complexity rises.

exactly equals that of the actual NKE perturbation generated for that image at that ϵ_l, add it to the clean image, and clip to $[0, 1]$ — so every control image removes exactly as much signal, in the L_2 sense, as its NKE counterpart, differing only in whether the perturbation direction is random or adversarially structured. This directly addresses Borji's confound concern that L_p distance alone is not a valid proxy for perceptual similarity: holding magnitude fixed while varying only structure means any recognizability difference between the two conditions isolates the effect of adversarial *structure* itself, not the amount of signal removed. That the random-direction control is in fact *more* destructive is, on reflection, unsurprising: the gradient-based attack must also keep $f(x') = y$ (Eq. 1), so part of its perturbation budget is spent moving along directions the classifier's decision boundary tolerates, which are not the directions that maximally destroy human-recognizable structure, whereas isotropic Gaussian noise carries no such constraint and spends its

Table 1: Reduced-scale replication of the white-box/black-box asymmetry. Entries are the fraction of source-generated NKE images the target model still labels correctly (label-kept rate; ϵ_l at the second-largest grid level). $*$ = white-box. Nie et al.'s ImageNet black-box transfer (<0.05) is shown as an external reference.

Dataset	Source model	Success vs. Model A	Success vs. Model B
MNIST	LeNet (A)	1.00*	0.98
	MLP (B)	0.95	1.00*
CIFAR-10	ResNet-18 (A)	1.00*	0.50
	VGG-11 (B)	0.23	1.00*
ImageNet	ResNet-152 (A)	1.00*	0.10
	(transfer $\rightarrow$0.01 at ϵ_l=60, 0.00 at ϵ_l=100; cf. Nie et al. <0.05)		

entire budget indiscriminately. The shape recognizer collapses toward chance. On MNIST the generic proxy stays high (0.99$\rightarrow$0.94) — NKE digits remain broadly recognizable (mirroring the high MNIST transfer above) — so the gap is a natural-image phenomenon.

Empirical human curve (pilot, N=5). Using the released, hardened harness we collected real forced-choice responses under the protocol of Sec. 3.3. Table 2 reports a *pilot* of N=5 attention-passing participants (all five collected sessions passed the dataset-calibrated attention check; 260 judgements retained). **This is a small-sample pilot, not a definitive human study**: per-cell counts are modest (12–23 per ϵ_l) and the bootstrap 95% CIs correspondingly wide. Even so, the predicted pattern is stark: human recognition of NKE images falls from near-ceiling (0.88 at ϵ_l=2.5) to $\sim$**0.08 by** ϵ_l=12–16, with "cannot tell" selected on **78–83%** of trials at the largest perturbations,

while model accuracy stays at 100% — a human–model gap widening to ∼0.92, direct human evidence for the gap our original formulation only illustrated schematically [3, Fig. 3]. The human ceiling is already only 0.85 on the *clean* 32×32 images, confirming that a dataset-calibrated attention threshold is necessary (our CIFAR-10 gate of 0.55 is set from this clean-trial ceiling, not tuned to retain participants). The matched-L_2 Gaussian contrast is more equivocal than on the proxy: raters find NKE *more* recognizable than equal-energy noise at small-to-mid ϵ_l (*e.g.* 0.63 vs. 0.50 at ϵ_l=5), the same direction as the proxy's "not signal loss" effect, but the conditions are *not* statistically separable at this sample size (overlapping CIs). Establishing the signal-vs-structure dissociation on real humans therefore requires the properly-powered sample the harness is built to collect; for that claim the proxy (Table 3) remains our primary evidence, while the pilot corroborates the gap itself.

4.3. NKE is an OOD phenomenon invisible to confidence-based monitoring

Table 4 shows a sharp detector dissociation. On CIFAR-10, MSP and energy detectors flag **0%** of NKE images and ECE is ≈0 — the model is confident *and* correct, so calibration looks perfect on inputs no human could label. The **Mahalanobis feature-distance detector flags 100%** of them (from $\epsilon_l = 1$), at a threshold calibrated to a true 5% FPR on held-out clean data; its AUROC vs. clean is ≈1.00 at every $\epsilon_l \geq 1$, so the detection reflects score separation, not threshold choice. NKE images are far off-manifold in representation space yet indistinguishable at the output layer — arguing for feature-distance OOD detection over softmax/energy/temperature methods when such failures matter, a reframing neither source paper performed,

Table 2: **Pilot (N=5 attention-passing participants) — a small-sample study.** Empirical human forced-choice accuracy vs. ϵ_l (CIFAR-10, ResNet-18 stimuli); all five collected sessions passed the dataset-calibrated attention check, 260 judgements retained. Per-cell counts (last row) are modest, so the bootstrap 95% CIs are wide. Model accuracy is 1.00 at every level. "Gauss" = matched-L_2 Gaussian control; the ϵ_l=0 column is the clean baseline (shared by both conditions).

ϵ_l	0	1	2.5	5	8	12	16
Human accuracy (NKE)	0.85	0.88	0.88	0.63	0.37	0.09	0.08
bootstrap 95% CI	[.72,.95]	[.69,1.0]	[.71,1.0]	[.42,.84]	[.16,.58]	[.00,.22]	[.00,.25]
Human accuracy (Gauss ctrl)	0.85	0.73	0.81	0.50	0.44	0.08	0.05
"Cannot tell" rate (NKE)	0.07	0.06	0.00	0.11	0.37	0.78	0.83
Human–model gap (NKE)	0.15	0.12	0.12	0.37	0.63	**0.91**	**0.92**
# judgements (NKE)	40	16	17	19	19	23	12

though this advantage holds only against an attacker *oblivious* to the detector (Sec. 4.4).

4.4. Feature-distance detection is not adaptively robust

The Mahalanobis result above assumes an attacker oblivious to the detector. But feature distance is differentiable, so an adaptive attacker can fold it into the objective: minimise $\mathrm{CE}(f(x'), y) + \lambda\, d_{\mathrm{Maha}}(x')$ — keeping the label *and* pulling the penultimate feature back toward the clean class manifold — while still projecting the perturbation to $\|x' - x\|_2 \geq \epsilon_l$. Table 5 (plotted in Fig. 5) sweeps λ at a large ϵ_l ($L_2 \approx 12$) where oblivious NKE is flagged 100%. A tiny penalty drives **Mahalanobis detection from 100% to 0%** while success stays $\geq$**0.99** out to $\lambda = 0.02$ and the perturbation stays large

Table 3: Recognizability proxies vs. ϵ_l (CIFAR-10, ResNet-18 source; generic recognizer = VGG-11). Model accuracy is 1.00 at every level. "Gauss" = matched-L_2 Gaussian control.

ϵ_l	0	2.5	5	8	16
Generic recognizer (NKE)	0.93	0.965	0.80	0.64	**0.485**
Generic recognizer (Gauss ctrl)	0.93	0.54	0.20	0.15	0.145
Shape recognizer (NKE)	0.60	0.40	0.19	0.13	0.105

Table 4: OOD detection rate at a held-out-calibrated 5% FPR, and ECE, for NKE images (CIFAR-10, ResNet-18). References: genuine OOD (SVHN) → Maha 0.70 / MSP 0.53; classical PGD adversarials → Maha 0.98 / MSP 0.00. NKE is thus flagged by Mahalanobis *more* reliably (1.00) than genuine OOD.

ϵ_l	1	2.5	5	8	16
Model accuracy	1.00	1.00	1.00	1.00	1.00
ECE	0.00	0.00	0.00	0.00	0.00
MSP detection	0.00	0.00	0.00	0.00	0.00
Energy detection	0.00	0.00	0.00	0.00	0.00
Mahalanobis detection	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**

($L_2 \approx 12.06$, essentially unchanged); MSP stays ≤ 0.07. Only for larger λ, where the feature-pullback term overwhelms the large-L_2 projection, does success degrade (0.77 at $\lambda = 0.05$, ≈0.11 by $\lambda = 1$). Crucially, **the evasion is not white-box overfitting**: a *held-out* detector fit on disjoint data is evaded identically, so the attacker need only know that a feature-distance detector is in use, not its parameters. There is thus a wide λ band (10^{-3} to $2{\times}10^{-2}$) that is simultaneously fully successful, large-perturbation, and

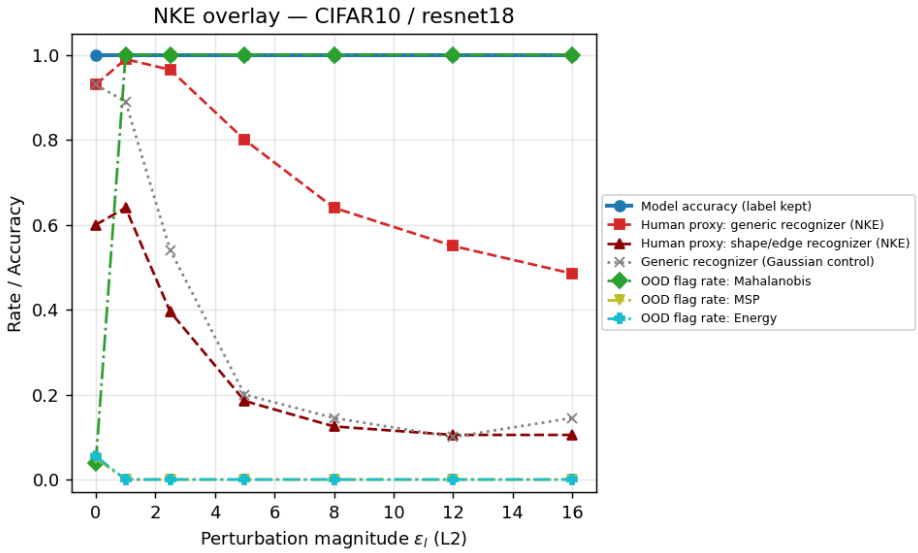

Figure 4: Summary overlay (CIFAR-10, ResNet-18). Model accuracy (label kept) stays at 1.0; the generic and shape recognizer proxies (proxies for human accuracy) fall well below it; the matched Gaussian control falls fastest; MSP/energy OOD flag rates stay at ≈0 while Mahalanobis flags 100%. The gap between the flat model curve and the descending proxy curves is the human–model gap that prior work only illustrated schematically.

fully evasive against both detectors: **feature-distance detection defeats a static NKE adversary but is defeated at no cost by an adaptive one**. The honest reframing is therefore not "deploy Mahalanobis" but that NKE is an off-manifold phenomenon which *only* off-manifold detectors can see at all — and even those are not robust to an adversary who knows they are there.

Two further checks (full detail in App. Appendix B, Table B.8) confirm this is a property of feature-distance detection in general, not of one metric or model. First, the evasion generalises across detector *design*: under honest held-out threshold calibration, a two-layer ensemble objective drives **every** parametric detector we tested — tied- and per-class-covariance, at both

Table 5: Adaptive attack against the Mahalanobis detector (CIFAR-10, ResNet-18, at an ϵ_l with $L_2 \approx 12$, where oblivious NKE is 100% detected). The attacker minimises $\mathrm{CE} + \lambda\, d_{\mathrm{Maha}}$; detection is at a held-out-calibrated 5% FPR. Both a *white-box* detector (the exact one attacked) and a *held-out* detector (fit on disjoint data) are evaded to 0% across a wide λ band, at no cost to success or perturbation magnitude.

λ	0	0.001	0.01	0.02	0.03	0.05	0.1	1.0
NKE success	1.00	1.00	1.00	0.99	0.97	0.77	0.16	0.11
Perturbation L_2	12.02	12.05	12.05	12.06	12.07	12.06	12.07	12.06
Maha detection (white-box)	**1.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00
Maha detection (held-out)	**1.00**	**0.00**	**0.00**	**0.00**	0.00	0.00	0.00	0.00
MSP detection	0.00	0.00	0.03	0.07	0.21	0.34	0.09	0.01

layers — to 0% detection at 100% success, while a non-parametric cosine-kNN detector [23] never flags standard NKE at all (a per-class detector only *appears* robust under in-sample calibration, an artifact of its 77% true FPR — a calibration caution in its own right). Second, the no-cost small-λ evasion reproduces against the adversarially trained ResNet-18 (45% PGD robust accuracy), consistent with the orthogonality of NKE and classical robustness.

4.5. Mechanism: texture is destroyed far faster than shape

Table 6 (plotted in Fig. 6) shows a channel dissociation: GLCM **texture similarity collapses to ≈0 by** $\epsilon_l = 5$, while **edge/shape structure survives far longer** (F1 plateaus at ∼0.31–0.45), with model accuracy pinned at 1.00. The surviving edge skeleton is insufficient for the shape/generic recognizers to keep up, yet the source model is invariant to both channels — consistent with a texture/shape account of the gap [22]. (The Gram-matrix

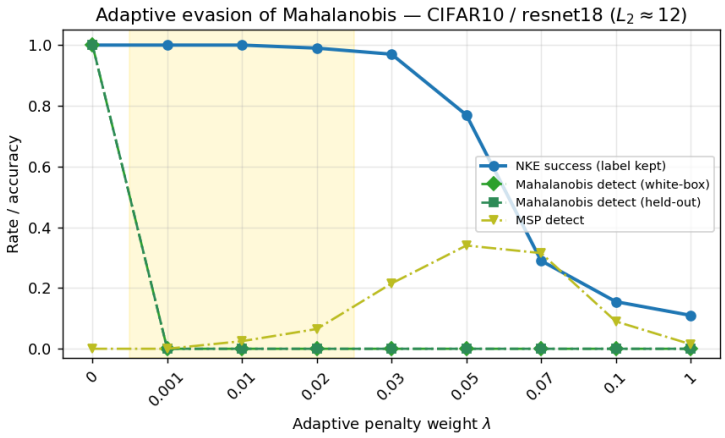

Figure 5: Adaptive evasion of the Mahalanobis detector (CIFAR-10, ResNet-18, $L_2 \approx 12$). A tiny penalty λ drives detection (white-box and held-out, green) to 0% while NKE success (blue) stays ≥ 0.99 across the shaded band; only much larger λ costs success. MSP detection (yellow) never recovers the gap.

cosine stayed ~ 1.0 throughout; it is dominated by global colour energy and is uninformative — GLCM is the meaningful texture metric.)

4.6. Architecture generality: does shape bias close the gap? (a Vision Transformer)

Sec. 4.5 shows the attack spares shape while destroying texture, and CNNs are texture-biased whereas humans — and *vision transformers* — are comparatively shape-biased [22, 24, 25], yielding a falsifiable prediction: a ViT should read the surviving edge skeleton *better* than a second CNN does. We add a small from-scratch ViT [26] (patch 4, dim 192, depth 6; 1.8M params, 83.2% clean acc.) as a third CIFAR-10 architecture and re-run every track (full per-track tables and figure in App. Appendix C). Three findings result. *(i) The mechanism is architecture-general:* ViT-sourced NKE

Table 6: Shape vs. texture (CIFAR-10, ResNet-18): mean clean-vs-perturbed edge preservation and texture similarity by ϵ_l.

ϵ_l	1	2.5	5	8	16
Edge preservation (Canny F1)	0.76	0.57	0.45	0.38	0.31
Texture similarity (GLCM)	0.83	0.30	**0.00**	0.00	0.00

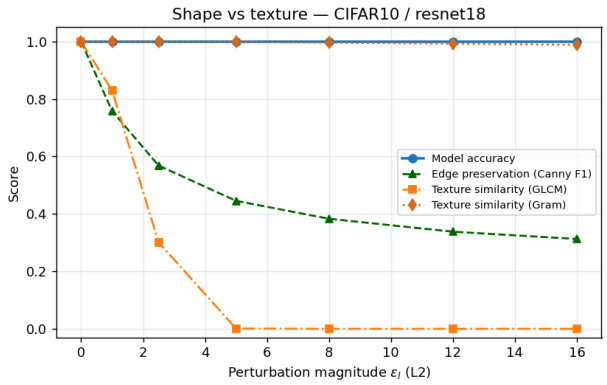

Figure 6: Shape vs. texture dissociation (CIFAR-10, ResNet-18). Model accuracy (blue) stays at 1.0 while GLCM texture similarity (orange) collapses to 0 by ϵ_l=5; edge preservation (green) survives far longer. The Gram-matrix cosine (red) is dominated by global colour energy and stays ≈1 throughout — uninformative, unlike GLCM.

reproduces the Sec. 4.5 texture-before-shape dissociation exactly (GLCM $\rightarrow\approx$0 by ϵ_l=5, edge F1 plateauing at ∼0.32–0.44), so it is not a CNN artifact. *(ii) But shape bias does not close the gap (prediction falsified):* comparing retention (accuracy on NKE ÷ on that source's clean images, removing the clean-accuracy confound), the shape-biased ViT recognizer retains **0.08–0.11 *less*** than the second CNN on CNN-sourced NKE at every $\epsilon_l \geq 2.5$ — the surviving edge skeleton does not let even a shape-biased reader recover the

label, so the human–model gap is not a texture-bias artifact. *(iii) Attention models produce the most model-specific NKE:* mean cross-model retention at ϵ_l=16 is 0.41 (ResNet-18 source), 0.21 (VGG-11), and **0.15** (ViT); the generic recognizer of ViT-sourced NKE falls to 0.12–0.14, barely above its matched-Gaussian control (0.10), the human–model gap is largest here ($\sim$86 points), and on the ViT penultimate the Mahalanobis detector catches only 43–50% of *oblivious* NKE (vs. 100% on ResNet-18) while the adaptive attack still evades at λ=0.001 — so the feature-distance detection advantage of Sec. 4.3 is itself architecture-dependent and favourable to CNNs.

4.7. No classical defense mitigates NKE; robustness is orthogonal

Table 7 (plotted in Fig. 7) shows that every defended model remains $\approx$100% NKE-vulnerable, including an adversarially trained model with **45% PGD robust accuracy**. Across models, the correlation between classical robust accuracy and continuous NKE resistance is $r \approx -0.02$ on CIFAR-10 (and undefined-with-zero-variance on MNIST, where NKE resistance is uniformly 0 across a 0$\rightarrow$86% robust-accuracy range — the strongest possible statement of orthogonality). Resistance to small-ϵ attacks does not transfer to large-ϵ_l, same-label attacks; they are distinct vulnerabilities. (Post-hoc blur/smoothing crater clean accuracy on this non-robust base model, confounding their rows; the adv-trained-vs-undefended contrast is the load-bearing comparison.)

4.8. Extensions based on Nie et al. [5]

Nie et al. compared their four attack variants and studied transferability, both only through the lens of *attack success rate*; we revisit both through our new metrics (full detail, Table B.9 and Fig. B.10, in App. Appendix B).

Table 7: Defenses vs. NKE (CIFAR-10, ResNet-18 base). NKE success is measured white-box against each defended model at the largest ϵ_l.

Model	Clean acc	Classical robust acc (PGD)	NKE success @ max ϵ_l
Undefended	1.00	0.00	1.00
JPEG (q=40)	0.78	0.11	0.99
Random resize+pad	0.85	0.00	1.00
Adversarial training	0.74	**0.45**	**1.00**

First, projecting all four variants onto the same L_2 sphere ($L_2 \approx 8$), so they differ only in step rule and momentum, their reported success-rate gaps largely vanish (all $\geq$99.5% success, all flagged 100% by Mahalanobis and 0–3% by MSP, near-identical edge preservation); only the human proxy varies modestly. **The OOD dissociation and confident-and-correct property are thus properties of the phenomenon, not the variant** — the variant mainly trades off perturbation magnitude, which in turn drives the human gap. *Second*, source-model manifold-proximity (negative Mahalanobis distance) predicts *which* examples transfer only at small ϵ_l (AUROC 0.73 at ϵ_l=2.5), decaying to chance and reversing by large ϵ_l (0.39 at ϵ_l=16), while edge preservation never predicts transfer ($|r| \leq 0.12$). In the large-ϵ_l regime that defines NKE, transfer is disconnected from both manifold-distance and shape — a quantitative deepening of Nie et al.'s "highly model-specific" conclusion: the surviving signal is idiosyncratic to the source model, not a shared, human-aligned remnant.

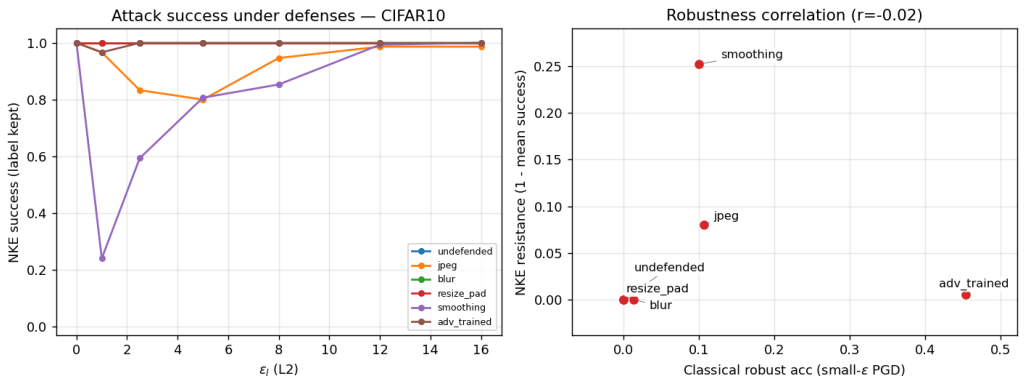

Figure 7: **Left:** NKE success stays ≈1.0 across ϵ_l for every defense (CIFAR-10, ResNet-18 base); only randomized smoothing dips briefly at small ϵ_l before recovering. **Right:** classical robust accuracy (PGD) vs. NKE resistance ($1-\overline{\text{success}}$) across defenses, $r \approx -0.02$ — robustness to small-ϵ attacks does not predict resistance to large-ϵ_l NKE.

4.9. ImageNet scale: all findings hold with the models Nie et al. [5] used

We run the full pipeline with ImageNet-pretrained ResNet-152 (source) and VGG-16 (target) on real Imagenette images (Fig. B.11, App. Appendix B). *Most* findings reproduce at scale: white-box success and confidence stay at 1.00; black-box transfer collapses 0.81 → 0.13 → 0.01 → 0.00; Mahalanobis flags ∼100% while MSP and energy stay at 0%; and texture similarity collapses to 0 while edge preservation degrades only to ∼0.21. The transfer curve doubles as the generic-recognizer human proxy, so the human–model gap, the OOD dissociation, and the transfer collapse are all visible on one axis, all with the exact model class Nie et al. studied.

The cross-model proxy saturates at ImageNet — but CLIP resolves it. The second-CNN proxy above conflates "a human would recognize this" with "this transfers to another CNN"; at ImageNet the NKE transfer collapse drives the proxy to its floor, so NKE (0.125 at $\epsilon_l = 30$) and the

matched-Gaussian control (0.155) become indistinguishable — not because the gap is signal loss, but because the proxy has no headroom. We therefore add a *stronger, non-CNN* recognizability proxy: zero-shot classification by CLIP (ViT-B/32, LAION-2B), which was never trained on these tasks and is not the attack's transfer target (Sec. 4.10, Fig. B.12). CLIP has ample headroom and reveals a clear human–model gap at *both* scales, confirming the effect is not merely a CIFAR artifact.

4.10. CLIP as a stronger recognizability proxy

We report CLIP zero-shot accuracy on NKE and matched-Gaussian-control images at both scales (Fig. B.12, App. Appendix B; chance = 0.10). Two things stand out. First, **a large human–model gap is confirmed on an independent, broadly-trained proxy at both scales**: while the source model stays at 100%, CLIP falls to 0.24 (CIFAR-10) and 0.33 (ImageNet) at the largest ϵ_l. Second, the NKE-vs-Gaussian relationship is **scale-dependent**: on CIFAR-10, NKE remains *more* CLIP-recognizable than matched noise (*e.g.* $\epsilon_l = 8$: 0.66 vs. 0.17), replicating the "not signal loss" effect on a non-CNN proxy; on ImageNet the ordering *reverses* — NKE is *less* recognizable than matched noise ($\epsilon_l = 100$: 0.33 vs. 0.61), so the perturbation that keeps ResNet-152 confident actively destroys broad semantic content *faster* than equal-energy noise: the examples are adversarial in the human direction too. Either way the gap is not explained by generic signal degradation — on CIFAR NKE preserves more human-readable signal than noise, on ImageNet it destroys more — but in both cases the model is unmoved.

5. Discussion and Conclusion

We answered the three questions the NKE formulation left open. *(i)* The human–model gap is real on natural images (a ∼50-point proxy gap on CIFAR-10) and, on CIFAR-10, is *specific to adversarial structure* rather than signal loss; a CLIP zero-shot proxy confirms the gap at ImageNet scale as well (there NKE is even *more* destructive to broad semantics than equal-energy noise). *(ii)* NKE is best understood as an OOD phenomenon invisible to confidence/energy/calibration monitoring and visible only to feature-distance detection — and even that visibility is not adaptively robust: an attacker who adds the Mahalanobis distance to the objective evades detection (100%→0%) at no cost to success or perturbation size, including on a held-out detector, an adversarially trained model, and (via a two-layer ensemble objective under honest held-out calibration) every feature-distance detector we tested. *(iii)* It is orthogonal to classical adversarial robustness, so existing defenses do not address it. The mechanistic dissociation — texture destroyed far faster than shape — offers a candidate explanation linking all three findings; a Vision Transformer probe shows it is architecture-general, but — against the natural shape-bias prediction — a more shape-biased model is *not* a better reader of the surviving structure, and attention models in fact generate the most model-specific NKE of all (Sec. 4.6). Extending Nie et al. [5], we further show their variant-success differences vanish at matched perturbation magnitude, and their transferability collapse corresponds to a regime in which surviving signal is idiosyncratic to the source model rather than a shared, on-manifold remnant. The main limitation is that our primary human curve is a computational proxy; a small real-human pilot (N=5) corroborates the

gap directly (human recognition $\rightarrow\sim$0.08 at large ϵ_l while the model stays at 1.00) but is underpowered for the signal-vs-structure contrast, for which the harness is built to collect a larger sample. The ImageNet-scale results use Imagenette (10 real ImageNet classes) rather than the full 1000-class validation set, which was unavailable in our environment.

More broadly, NKE occupies a regime that existing evaluation practice does not cover: it is neither a small, imperceptible perturbation nor a naturally out-of-distribution input, yet it produces exactly the kind of confident-but-unreliable prediction that robustness and OOD evaluations exist to catch. Any system that leans on softmax confidence or energy scores as a first line of defense against distributional failure should be audited against this regime specifically, since Sec. 4.3 shows those tools are structurally blind to it, and Sec. 4.4 shows that even the one detector family that does see it stops seeing it as soon as the attacker knows it is there. This suggests at least three directions beyond the present study: scaling the released human-study harness from the N=5 pilot here to a properly powered, crowd-sourced sample; extending the analysis beyond vision to modalities such as audio and text, where a large, label-preserving perturbation is less obvious to construct but plausibly still exists; and pursuing detectors that are adaptively robust by construction, rather than only in a static, non-adversarial evaluation, since our results indicate that off-manifold detection alone is not a sufficient defense once it is a known target. We release the full pipeline, stimuli, and harness so that this regime can be audited as routinely as classical adversarial robustness already is. Code, figures, and the harness are available at `https://github.com/alikayyam/NKE_attack.git`.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: International Conference on Learning Representations (ICLR), 2014.

[2] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations (ICLR), 2015.

[3] A. Borji, A new kind of adversarial example, arXiv preprint arXiv:2208.02430 (2022).

[4] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 427–436.

[5] X. Nie, G. Xiao, S. Pan, B. Wang, H. Ge, T. Fang, A new type of adversarial examples, arXiv preprint arXiv:2510.19347 (2025).

[6] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial machine learning at scale, in: International Conference on Learning Representations (ICLR), 2017.

[7] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, in: International Conference on Learning Representations (ICLR) Workshop, 2017.

[8] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, P. McDaniel, Ensemble adversarial training: Attacks and defenses, in: International Conference on Learning Representations (ICLR), 2018.

[9] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: IEEE Symposium on Security and Privacy (S&P), 2017, pp. 39–57.

[10] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 9185–9193.

[11] G. Elsayed, S. Shankar, B. Cheung, N. Papernot, A. Kurakin, I. Goodfellow, J. Sohl-Dickstein, Adversarial examples that fool both computer vision and time-limited humans, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.

[12] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, F. A. Wichmann, Generalisation in humans and deep neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.

[13] Z. Zhou, C. Firestone, Humans can decipher adversarial images, Nature Communications 10 (2019) 1334.

[14] D. Hendrycks, K. Gimpel, A baseline for detecting misclassified and out-of-distribution examples in neural networks, in: International Conference on Learning Representations (ICLR), 2017.

[15] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, in: Advances in Neural Information Processing Systems (NeurIPS), 2018.

[16] W. Liu, X. Wang, J. Owens, Y. Li, Energy-based out-of-distribution detection, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.

[17] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, J. Snoek, Can you trust your model's uncertainty?

evaluating predictive uncertainty under dataset shift, in: Advances in Neural Information Processing Systems (NeurIPS), 2019.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations (ICLR), 2018.

[19] J. Cohen, E. Rosenfeld, Z. Kolter, Certified adversarial robustness via randomized smoothing, in: International Conference on Machine Learning (ICML), 2019, pp. 1310–1320.

[20] C. Guo, M. Rana, M. Cisse, L. van der Maaten, Countering adversarial images using input transformations, in: International Conference on Learning Representations (ICLR), 2018.

[21] A. Borji, Addressing the topological defects of disentanglement and adversarial robustness: notes on human vision, arXiv preprint arXiv:2208.11580 (2022).

[22] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, W. Brendel, Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, in: International Conference on Learning Representations (ICLR), 2018.

[23] Y. Sun, Y. Ming, X. Zhu, Y. Li, Out-of-distribution detection with deep nearest neighbors, in: International Conference on Machine Learning (ICML), 2022.

[24] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, Intriguing properties of vision transformers, in: Advances in Neural Information Processing Systems (NeurIPS), 2021.

[25] S. Tuli, I. Dasgupta, E. Grant, T. L. Griffiths, Are convolutional neural networks or transformers more like human vision?, in: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci), 2021.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR), 2021.

Appendix A. Human-study interface

This appendix describes the participant-facing interface of the released forced-choice harness (Sec. 3.3); Fig. A.8 shows a representative screen.

Layout. Each trial shows a single stimulus image, enlarged to a fixed 256×256 on-screen size using nearest-neighbor (pixelated) scaling regardless of the underlying image's native resolution (28×28 for MNIST, 32×32 for CIFAR-10), so no interpolation smoothing is introduced and every participant views the same effective magnification. Below the image, a progress indicator shows the current trial number out of the participant's total session length. Below that, one button per class label is shown (ten buttons for both MNIST and CIFAR-10, matching the number of dataset classes), followed by a single, visually distinct *Cannot tell* button.

Interaction. The task begins with a participant-ID entry (auto-generated if left blank); no other setup or practice phase is required. On each trial, the participant clicks exactly one button — a class label or Cannot tell — which immediately advances to the next trial; there is no time limit and no way to

revisit a previous trial. Response time is measured as wall-clock time between stimulus onset and the click. At the end of the session, a thank-you screen appears with a one-click download of the participant's own response log (or automatic submission, if a collection endpoint is configured).

Trial composition and blinding. Every clean ($\epsilon_l = 0$) image is shown to every participant, serving as an interleaved attention check; these are visually indistinguishable from NKE and Gaussian-control trials, so participants cannot tell which condition a given trial belongs to. For the non-clean trials, each base image is shown to a given participant at exactly one perturbation level and under exactly one condition (NKE or Gaussian control), so no participant ever sees the same underlying image twice — this is what makes the design between-subjects rather than within-subjects for the perturbed conditions, ruling out image-specific memory effects across conditions.

Design rationale. The explicit Cannot tell option is included because a standard k-way forced choice without an opt-out would conflate genuine recognition with a lucky guess at floor ($1/k$) once perturbation destroys most information; separating the two is necessary to interpret the low end of the human psychometric curve (Sec. 4.2) as reflecting inability to recognize the image rather than an unlucky guess. Fixed-size, nearest-neighbor display was chosen over native-resolution or interpolated display so that comparisons across ϵ_l and across datasets are not confounded by incidental differences in on-screen size or by smoothing artifacts that would themselves alter recognizability.

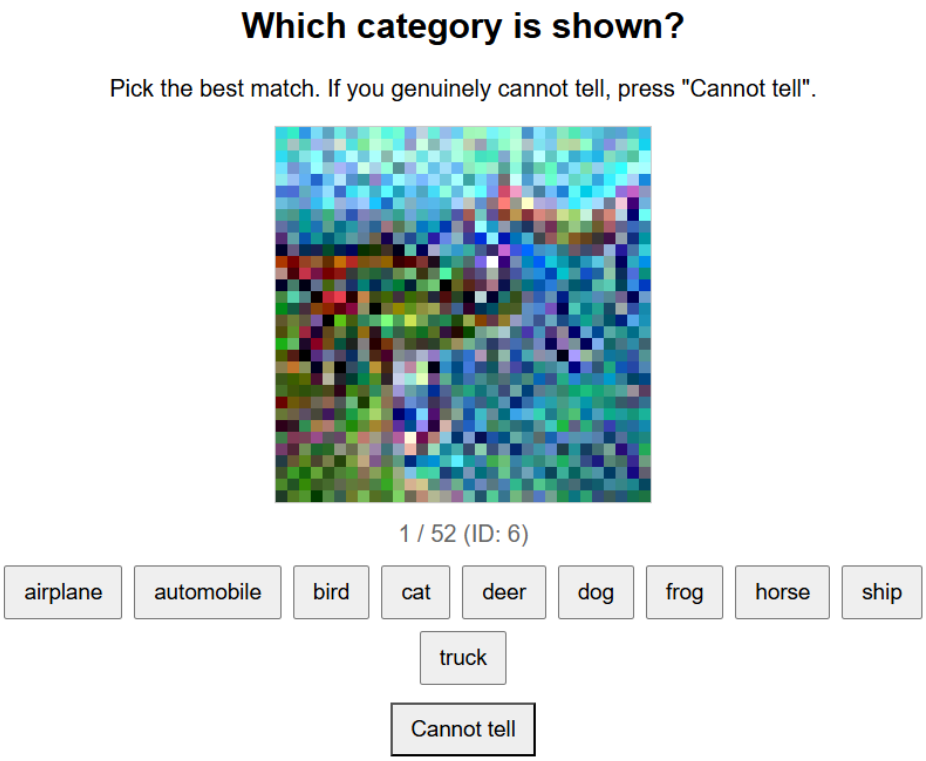

Figure A.8: The participant-facing forced-choice interface (CIFAR-10 stimuli shown). Each trial presents one enlarged, nearest-neighbor-scaled stimulus with a progress counter, ten class-label buttons, and an explicit *Cannot tell* option.

Appendix B. Extended results and figures

This appendix collects material deferred from the main text: the qualitative figure (Fig. B.9), the full feature-distance detector battery (Sec. 4.4, Table B.8), and the extended analysis of Nie et al.'s variants and transferability (Sec. 4.8, Table B.9, Fig. B.10).

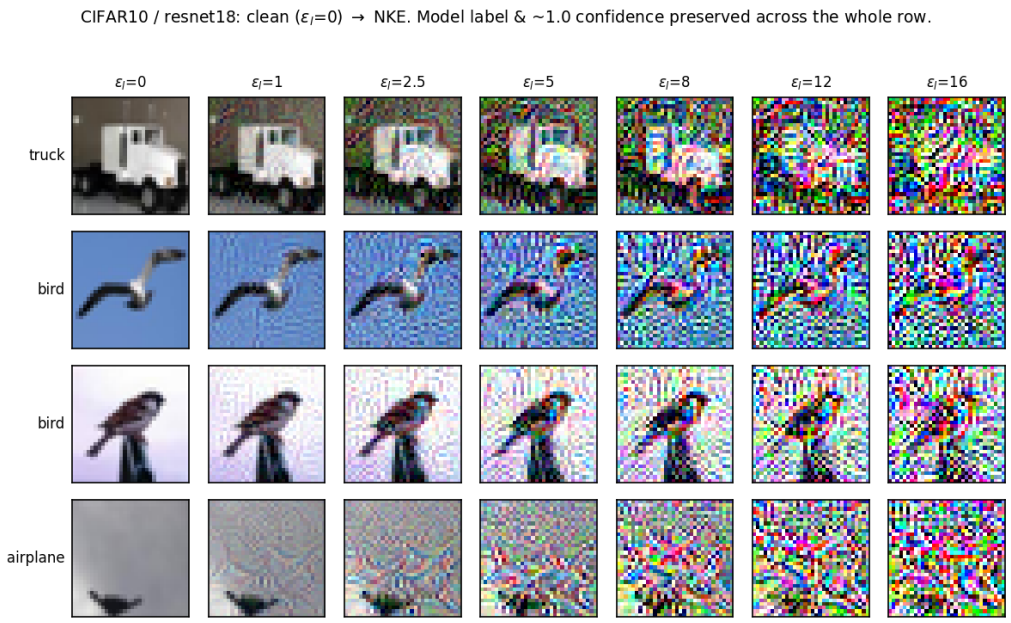

Figure B.9: CIFAR-10 examples: clean ($\epsilon_l = 0$) → NKE across increasing ϵ_l. ResNet-18 keeps the original label at ∼1.0 confidence across each entire row, even where the image is visually pure noise. (ImageNet analogue in the repository.)

Appendix B.1. Feature-distance detector battery (Sec. 4.4)

Is the adaptive evasion specific to the tied-covariance penultimate detector, or does it generalise across detector *design*? Holding ϵ_l fixed, we test a battery of feature-distance detectors, each calibrated to a true 5% FPR on held-out clean data (Table B.8). *This last point is load-bearing:* with an in-sample threshold (calibrated on the detector's own fit set) a per-class (untied) covariance detector *appeared* adaptively robust — it flagged 100% of adaptive NKE at every λ — but its true held-out FPR was 77%, i.e. it was flagging almost everything. In-sample thresholds manufacture phantom robustness. Under honest calibration (all detectors' held-out FPR $\approx$ 0.04–

0.05) that robustness evaporates. First, detector choice still matters against *un-adapted* NKE: a non-parametric cosine-kNN detector [23] flags **0%** of standard NKE (which retains high cosine similarity to training features), so only whitened-distance (Mahalanobis-type) detectors catch NKE at all. Second, evasion generalises across design: penalising the tied-covariance penultimate detector alone already drives the *per-class* (untied) detector on the same features to 0% as well, and a two-layer ensemble objective drives **every** parametric detector — tied and untied, both layers — to 0% at 100% success. **No feature-distance detector we tested resists an adaptive attacker that can name and differentiate it**; the negative result of Table 5 is a property of feature-distance detection in general, not of the particular metric. Finally, the evasion is not specific to a non-robust model: repeating the attack against the adversarially trained ResNet-18 (45% PGD robust accuracy, Sec. 4.7) reproduces the no-cost evasion at small λ (at λ=0.001: success 1.00, Mahalanobis 100%→0% on both white-box and held-out detectors, $L_2 \approx 12.5$); at larger λ success plateaus near 0.53 rather than 0.11, but small-λ evasion is unaffected.

Appendix B.2. Extended analysis of Nie et al.'s variants and transferability (Sec. 4.8)

Variants at matched magnitude. Nie et al. report success-rate gaps between variants (*e.g.* NMI-FGSM > NI-FGM); but those attacks also differ in the perturbation magnitude they induce. Projecting all four onto the *same* L_2 sphere ($L_2 \approx 8$ on CIFAR-10), so they differ only in step rule and momentum, the reported differences largely vanish: all four reach ≥99.5% success, all are flagged 100% by Mahalanobis and 0–3% by MSP, and edge

Table B.8: Detection of adaptive NKE by a battery of feature-distance detectors, each at a *held-out-calibrated* 5% FPR (CIFAR-10, ResNet-18, $L_2 \approx 12$; adapt rows at λ=0.01). Row 1 confirms honest calibration (clean FPR ≈0.05). Attacking only the tied-penultimate detector also evades the untied one; a two-layer ensemble objective evades every parametric detector. Cosine-kNN never flags NKE at all.

Scenario	Success	tied-penult	untied-penult	Maha layer-3	cosine-kNN
Clean false-positive rate	—	0.04	0.04	0.05	0.05
Standard NKE (no adaptation)	1.00	1.00	0.95	1.00	0.00
Adapt vs. tied-penult	1.00	**0.00**	**0.00**	0.22	0.00
Adapt vs. tied ensemble (penult + layer-3)	1.00	**0.00**	**0.00**	**0.00**	0.00

preservation is nearly identical (Table B.9). Only the human-proxy gap varies modestly (sign-based I-FGSM is the most human-destructive at 0.62; L_2-normalized NMI-FGM the least at 0.83). The variant mainly trades off perturbation magnitude (its unconstrained L_2 ranges $8 \to 28$ across variants at fixed ϵ_l), and magnitude in turn drives the human gap.

Table B.9: Attack variants at matched $L_2 \approx 8$ (CIFAR-10, ResNet-18). Success and detectability are invariant to variant; only the human proxy varies.

Variant	Success	Maha detect	MSP detect	Human proxy	Edge preservation
I-FGSM (L_∞ step)	1.00	1.00	0.03	0.62	0.50
NI-FGM (L_2 step)	1.00	1.00	0.00	0.78	0.49
NMI-FGSM (mom.+sign)	1.00	1.00	0.00	0.77	0.48
NMI-FGM (mom.+L_2)	1.00	1.00	0.00	0.83	0.46

What the transferability collapse reveals. Nie et al. found black-box transfer collapses (<5% on ImageNet) but did not explain *which* examples

transfer. We test whether source-model manifold-proximity (negative Mahalanobis distance) or edge preservation predicts transfer to a second architecture (ResNet-18→VGG-11). At *small* perturbations, proximity predicts transfer well above chance (AUROC 0.73 at $\epsilon_l = 2.5$): the examples that transfer are those still near the manifold. But this predictivity **decays monotonically to chance (≈0.5 at $\epsilon_l = 8$) and reverses at large** ϵ_l (AUROC 0.39 at $\epsilon_l = 16$), and edge preservation never predicts transfer ($|r| \leq 0.12$). Thus in the large-ϵ_l regime that defines NKE, transfer becomes disconnected from both manifold-distance and shape.

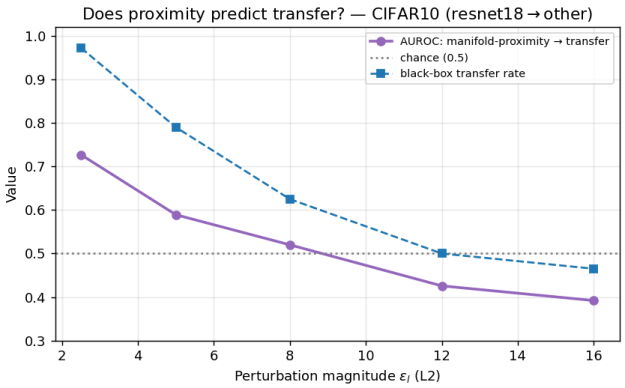

Figure B.10: Manifold-proximity predicts transfer only at small ϵ_l; its predictive power (AUROC, purple) decays through chance and reverses as ϵ_l grows, tracking the black-box transfer rate (blue). In the NKE regime, transfer is not explained by proximity or shape.

Appendix C. Vision Transformer shape-bias probe: full results

This appendix reports the per-track numbers underlying Sec. 4.6. The ViT is a from-scratch model (patch size 4 ⇒ 64 tokens, embedding dim 192, depth 6, 3 attention heads, MLP ratio 2; 1.8M parameters), trained for 70 epochs

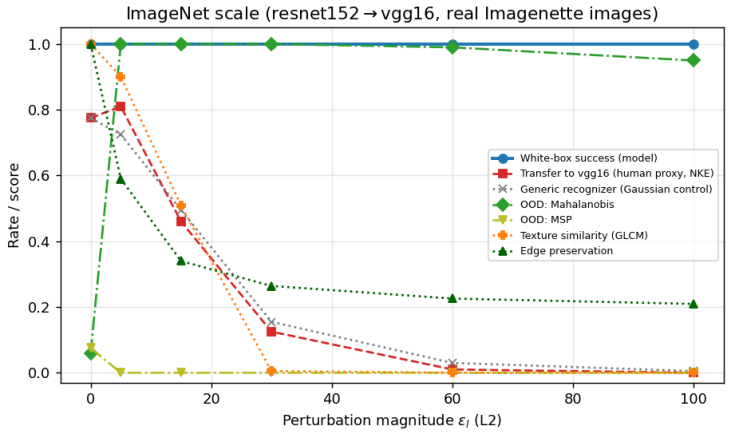

Figure B.11: ImageNet scale (ResNet-152→VGG-16, real Imagenette images). White-box success/confidence flat at 1.0; transfer (= generic-recognizer proxy) collapses to 0; Mahalanobis detects ∼100% while MSP stays at 0; texture is destroyed while edges partially survive. Note the NKE-transfer (red) and Gaussian-control (gray) curves nearly coincide — unlike CIFAR-10, the cross-model proxy cannot separate structured from random degradation here (see Sec. 4.9).

with AdamW (lr 5×10^{-4}, weight decay 0.05, 5-epoch linear warmup, label smoothing 0.1) under the same crop+flip augmentation as the CNNs, reaching **83.2%** clean test accuracy. It is registered as a third CIFAR-10 architecture and passed through the identical pipeline (attack, stimulus set, and every evaluation track) as ResNet-18 and VGG-11. All attacks use NMI-FGM at 60 steps as in the main text; n=200 correctly-classified, class-stratified images per source.

Cross-model transferability. Table C.10 extends the black-box matrix (Table 1) to all three CIFAR-10 architectures. White-box success is 1.00 on the diagonal; every off-diagonal (black-box) entry is low, and the ViT row (NKE *generated by* the ViT) is the lowest of all.

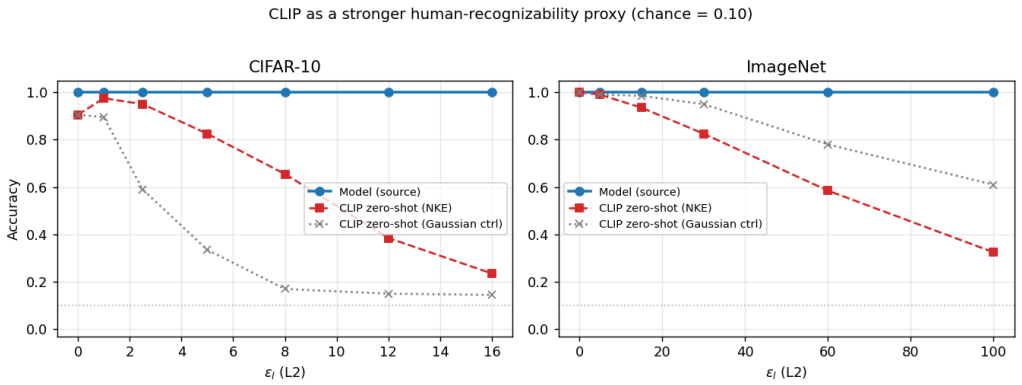

Figure B.12: CLIP zero-shot accuracy (a non-CNN, broadly-trained recognizability proxy) vs. ϵ_l. Model stays at 1.0; CLIP falls far below at both scales (gap confirmed). CIFAR-10: NKE > Gaussian control (not signal loss). ImageNet: NKE < Gaussian control (perturbation is *more* destructive to semantics than noise). Chance = 0.10.

Retention-normalised cross-recognizer analysis. Table C.11 gives the retention (accuracy on NKE ÷ accuracy on that source's clean images, i.e. normalised per recognizer to remove the clean-accuracy confound) that underlies Fig. C.13. The shape-bias contrast (last two rows) is negative at every $\epsilon_l \geq 2.5$: the ViT recognizer is not a better reader of CNN-sourced NKE. The source-specificity block shows ViT-sourced NKE collapsing earliest and lowest.

Mechanism (ViT source). Table C.12 is the ViT analogue of Table 6: the same texture-before-shape dissociation, with model accuracy pinned at 1.00.

Human proxies (ViT source). Table C.13 is the ViT analogue of Table 3; the generic recognizer is ResNet-18. Note that the NKE curve nearly coincides with the matched-Gaussian control (unlike the ResNet-18 source in Table 3), i.e. ViT NKE is almost as cross-model-destructive as noise.

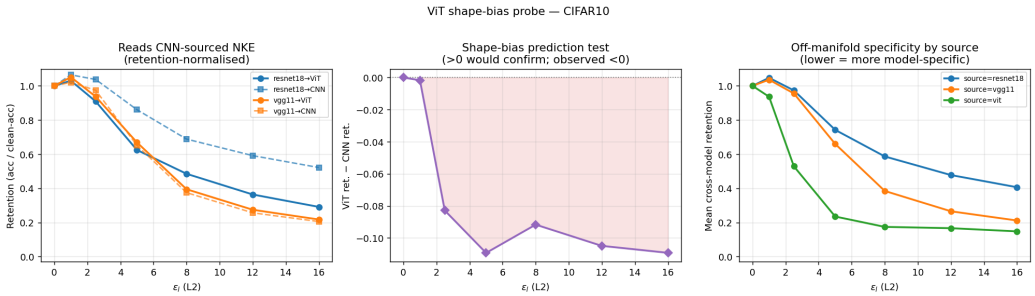

Figure C.13: ViT shape-bias probe (CIFAR-10), retention-normalised. **Left:** retention of CNN-sourced NKE by the ViT vs. the other CNN recognizer. **Middle:** the shape-bias prediction test — ViT-minus-CNN retention is *negative* throughout, so the shape-biased model is *not* a better reader (prediction falsified). **Right:** off-manifold specificity by source — NKE generated by the ViT (green) transfers worst and collapses earliest, i.e. attention models produce the most model-specific perturbations.

OOD detection (ViT source). Table C.14 is the ViT analogue of Table 4. The confidence/energy blindness and ECE≈0 reproduce, but Mahalanobis detection on the ViT's global CLS token is far weaker than on the ResNet-18 penultimate (peaking at 0.50 rather than 1.00).

Adaptive attack (ViT source). Table C.15 is the ViT analogue of Table 5. Oblivious NKE is only 43% detected here (weaker feature-distance signal), and a tiny penalty (λ=0.001) drives both white-box and held-out detectors to ≈0 at full success; success collapses faster than on the CNN as λ grows, so the no-cost evasion band is narrower but present.

Table C.10: CIFAR-10 cross-model label-kept matrix at ϵ_l=12 (fraction of source-generated NKE the target still labels correctly). Rows: source; columns: target. $*$ = white-box.

Source ↓ / Target →	ResNet-18	VGG-11	ViT
ResNet-18	1.00*	0.50	0.25
VGG-11	0.23	1.00*	0.22
ViT	0.13	0.16	1.00*

Table C.11: Retention (acc on NKE / acc on clean) on CIFAR-10. *Top:* shape-bias contrast — mean (ViT-recognizer retention − other-CNN-recognizer retention) for CNN-sourced NKE; negative means the shape-biased ViT reads CNN NKE *worse*. *Bottom:* source specificity — mean cross-model retention of the NKE each source generates (lower = more model-specific).

ϵ_l	2.5	5	8	12	16
Shape-bias prediction test (CNN-sourced NKE)					
ViT − CNN retention (contrast)	−0.08	−0.11	−0.09	−0.11	−0.11
Source specificity (mean cross-model retention by source)					
ResNet-18 source	0.98	0.74	0.59	0.48	0.41
VGG-11 source	0.95	0.66	0.38	0.27	0.21
ViT source	**0.53**	**0.24**	**0.18**	**0.17**	**0.15**

Table C.12: Shape vs. texture for ViT-sourced NKE (CIFAR-10). Model accuracy is 1.00 at every ϵ_l.

ϵ_l	1	2.5	5	8	16
Edge preservation (Canny F1)	0.77	0.57	0.44	0.38	0.32
Texture similarity (GLCM)	0.91	0.63	0.07	0.00	0.00

Table C.13: Recognizability proxies vs. ϵ_l for ViT-sourced NKE (CIFAR-10; generic recognizer = ResNet-18). Model accuracy is 1.00 at every level.

ϵ_l	0	2.5	5	8	16
Generic recognizer (NKE)	0.95	0.41	0.14	0.16	0.12
Generic recognizer (Gauss ctrl)	0.95	0.23	0.11	0.10	0.10
Shape recognizer (NKE)	0.69	0.39	0.18	0.10	0.12

Table C.14: OOD detection rate at a held-out-calibrated 5% FPR, and ECE, for ViT-sourced NKE (CIFAR-10). Model accuracy 1.00 throughout.

ϵ_l	1	2.5	5	8	16
ECE	0.04	0.03	0.02	0.02	0.02
MSP detection	0.00	0.00	0.00	0.00	0.00
Energy detection	0.00	0.00	0.00	0.00	0.00
Mahalanobis detection	0.01	0.14	0.33	0.41	0.50

Table C.15: Adaptive attack against the Mahalanobis detector on the ViT (CIFAR-10, $L_2 \approx 12$). Detection at a held-out-calibrated 5% FPR.

λ	0	0.001	0.01	0.02	0.03	1.0
NKE success	1.00	1.00	0.93	0.66	0.35	0.17
Perturbation L_2	12.15	12.05	12.05	12.06	12.07	12.07
Maha detection (white-box)	0.43	0.00	0.00	0.00	0.00	0.00
Maha detection (held-out)	0.42	0.01	0.00	0.00	0.00	0.00
MSP detection	0.00	0.01	0.03	0.12	0.17	0.00