Natural Jammers: When Non-Robust Features Become Antagonistic

Jiaming Zhang[1,*] **Fuyao Zhang**[1] **Pengjun Xie**[2] **Wei Yang Bryan Lim**[1,*]
[1]Nanyang Technological University [2]Alibaba Group
Project page: `https://jiamingzhang.netlify.app/naturaljammers`

Abstract

Deep neural networks are known to rely on non-robust features that align poorly with human perception. Prior work has largely framed these features as opportunistic shortcuts that help standard models on clean data. We show instead that non-robust cues can be antagonistic: they can act as **Natural Jammers** that actively mislead standard models on clean, unmodified images. Using disagreement between a standard model and robust models as an analysis lens, we identify a Jammed Set on which robust models outperform standard models. We provide causal evidence for *Spectral–spatial Locking*: jamming arises only when high-frequency components are precisely aligned with specific spatial structures. Disrupting this lock through spectral filtering or small geometric shifts can restore correct predictions. Crucially, we uncover a *Substitution Paradox*: while individual jammers are fragile to geometric shifts, the aggregate prevalence of jamming remains approximately stable and is redistributed rather than eliminated under such transformations. Our results reframe natural non-robustness as a failure field over the empirical model–data geometry: one that redistributes rather than disappears under small transformations.

1 Introduction

Deep neural networks (DNNs) achieve remarkable success on visual tasks, yet their reasoning differs fundamentally from human perception. Rather than relying solely on semantic content, DNNs are highly sensitive to non-robust features—subtle, high-frequency patterns that predict labels but remain imperceptible to humans Ilyas et al. [2019], Gyevnar and Kasirzadeh [2025], Liu et al. [2025a], Sang et al. [2022], Zhang et al. [2023a]. The prevailing view treats these features as opportunistic shortcuts: brittle patterns that standard models exploit as a crutch to reach correct decisions Geirhos et al. [2020], Kauffmann et al. [2025], Eriksson et al. [2025]. This framing assumes non-robust signals and semantic labels work synergistically.

We challenge this synergistic view by uncovering natural feature antagonism. As Figure 1 illustrates, non-robust signals do not always aid classification—for a substantial fraction of natural images, they actively mislead standard inference. We term these signals **Natural Jammers**. This creates an empirical paradox: although the robustness-accuracy trade-off suggests standard models (f_{std}) should upper-bound clean performance Tsipras et al. [2018], Suzuki et al. [2023], Zhang et al. [2024], Kafali et al. [2025], Li and Li [2025], we identify a distinct **Jammed Set ($\mathcal{J}$)** where robustly trained models (f_{rob}) significantly outperform their standard counterparts on unmodified images. The standard model's reliance on non-robust features becomes a liability—it is misled by brittle cues that contradict the label, while the robust model often succeeds because it is less sensitive to such cues. We employ f_{rob} as a diagnostic lens to disentangle robust semantic cues from non-robust correlations.

*Corresponding authors: jiaming.zhang@ntu.edu.sg, bryan.limwy@ntu.edu.sg.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

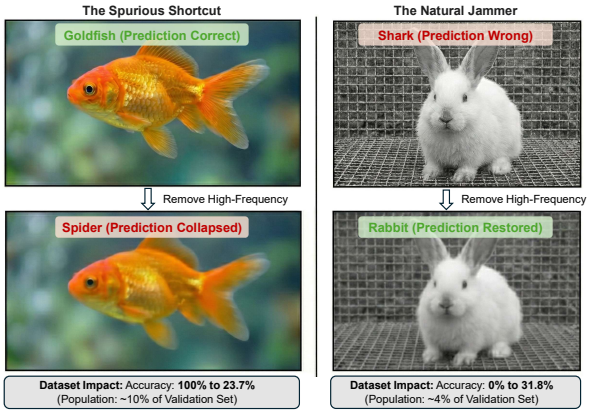

Figure 1: **The Duality of Non-Robust Features.** (Left) **Spurious Shortcut** ($\mathcal{S}$): prediction relies on fragile non-robust cues and collapses when these cues are suppressed. (Right) **Natural Jammer** ($\mathcal{J}$): non-robust cues interfere with semantics-aligned evidence, suppressing them partially restores correct predictions. This illustrates that non-robust cues can be antagonistic rather than merely supportive. Accuracies are measured with the primary lens (standard vs. robust ResNet-50, ℓ_2, $\epsilon = 1$).

To explain this interference, we propose the **Spectral-Spatial Locking** mechanism through controlled experiments. Natural jammers arise from high-frequency patterns, but their disruptive influence depends critically on alignment with specific spatial configurations. This phase-locked coupling explains a striking duality: while these features mislead standard inference, they are also fragile. Breaking the lock—through spectral filtering (the Cleansing Paradox, Section 3.1) or minor geometric shifts (the Sensitivity Inversion, Section 3.2)—renders the jammer inert, allowing the model's discriminative capacity to resurface.

This fragility raises a puzzle: if individual jammers are so easily disrupted, why does the model not achieve robust generalization? Our analysis reveals the **Substitution Paradox**: a 2% edge crop removes about one third of the images in the Jammed Set, yet its total size remains nearly unchanged. Natural non-robustness is not a defect of particular images but a persistent property of the model–data geometry. When one jammer's spatial-spectral lock breaks, new jammers emerge elsewhere, keeping the overall vulnerability level nearly constant. Non-robustness is redistributed rather than eliminated by simple augmentations.

Our contributions are:

- **Feature Antagonism:** We formalize the Jammed Set ($\mathcal{J}$), showing that a significant portion of standard model errors stem from non-robust signals that contradict semantic labels.
- **Spectral-Spatial Locking:** We provide causal evidence that non-robust features require precise alignment between high-frequency content and spatial structure to remain effective.
- **Substitution Paradox:** We demonstrate that jamming is a stable, systemic property—individual non-robust features are fragile and interchangeable, following a redistribution pattern under geometric transformations.

2 Landscape of Natural Non-Robustness

To move beyond a binary view of robust vs. non-robust features, we formalize a taxonomy based on the latent structure of the data distribution $\mathcal{D}$.

Assumption 2.1 (Manifold Decomposition)**.** Following the latent manifold hypothesis Ilyas et al. [2019], we assume the input space $\mathcal{X}$ maps to a latent representation $\mathcal{Z}$ that can be decomposed into two subspaces, $\mathcal{Z} = \mathcal{Z}_r \times \mathcal{Z}_{nr}$, where $\mathcal{Z}_r$ and $\mathcal{Z}_{nr}$ denote the robust and non-robust manifolds, respectively.

The interaction between these manifolds determines whether a model's reliance on non-robust signals yields an opportunistic shortcut or a predictive failure. We utilize a standard classifier f_{std} and an adversarially trained counterpart f_{rob} to analyze these divergent behaviors.

Table 1: **Distribution of Feature Pathologies on ImageNet.** Data is reported for $\epsilon = 1.0$ constraints. The proportions of $\mathcal{J}$ and $\mathcal{S}$ remain relatively stable across ResNet and WideResNet architectures, suggesting that jamming is not confined to a single architecture.

MODEL	$\mathcal{J}$ (JAMMED)	$\mathcal{S}$ (SPURIOUS)	$\mathcal{C}$ (COMMON)	$\mathcal{G}$ (UNINFORMATIVE)
RESNET-18 (ℓ_2)	2,448 (4.90%)	6,168 (12.34%)	28,712 (57.42%)	12,672 (25.34%)
RESNET-50 (ℓ_2)	2,140 (4.28%)	5,002 (10.00%)	33,071 (66.14%)	9,787 (19.57%)
WRN-50-2 (ℓ_2)	2,167 (4.33%)	4,698 (9.40%)	34,543 (69.09%)	8,592 (17.18%)
RESNET-18 (ℓ_∞)	2,502 (5.00%)	5,659 (11.32%)	29,221 (58.44%)	12,618 (25.24%)
RESNET-50 (ℓ_∞)	2,252 (4.50%)	4,306 (8.61%)	33,767 (67.53%)	9,675 (19.35%)
WRN-50-2 (ℓ_∞)	2,232 (4.46%)	4,145 (8.29%)	35,096 (70.19%)	8,527 (17.05%)

2.1 Formal Taxonomy: The Quad-Set Partition

Given a sample $(x, y) \in \mathcal{D}$ and a chosen reference pair $(f_{\text{std}}, f_{\text{rob}})$, we partition samples into four disjoint sets based on their predictions on the original clean input x:

- **Common Set ($\mathcal{C}$):** $\{x \mid f_{\text{std}}(x) = y \wedge f_{\text{rob}}(x) = y\}$. Samples where features in $\mathcal{Z}_r$ and $\mathcal{Z}_{nr}$ are synergistic.
- **Spurious Set ($\mathcal{S}$):** $\{x \mid f_{\text{std}}(x) = y \wedge f_{\text{rob}}(x) \neq y\}$. The classic shortcut regime where f_{std} succeeds via brittle patterns in $\mathcal{Z}_{nr}$.
- **Jammed Set ($\mathcal{J}$):** $\{x \mid f_{\text{std}}(x) \neq y \wedge f_{\text{rob}}(x) = y\}$. Samples where f_{std} is misguided by antagonistic signals in $\mathcal{Z}_{nr}$.
- **Uninformative Set ($\mathcal{G}$):** $\{x \mid f_{\text{std}}(x) \neq y \wedge f_{\text{rob}}(x) \neq y\}$. Samples where signals in both manifolds are insufficient for classification.

While $\mathcal{S}$ represents the well-documented reliance on human-irrelevant shortcuts, $\mathcal{J}$ exposes a more pathological interaction where non-robust features actively contradict the ground-truth logic:

Definition 2.2 (Natural Jammer)**.** A **Natural Jammer** is a non-robust signal $\mathbf{z}_{nr} \in \mathcal{Z}_{nr}$ that exhibits feature antagonism, such that its predictive direction contradicts the semantic label y.

Operationally, for a standard model f_{std}, an input x is said to be *jammed* if there exists a semantics-preserving transformation $T \in \mathcal{T}$ (e.g., spectral filtering or geometric cropping) such that:

$$f_{\text{std}}(x) \neq y \quad \text{and} \quad f_{\text{std}}(T(x)) = y. \tag{1}$$

In this context, T acts as a de-locking operator that suppresses the antagonistic non-robust signal, allowing the model's latent discriminative capacity to resurface.

2.2 Empirical Prevalence and Stability

We conduct a representative audit on the ImageNet validation set Russakovsky et al. [2015] ($N = 50,000$) to evaluate the prevalence of these sets across different architectures. The membership of the set is defined on the original inputs and, therefore, on the unmodified images $\text{Acc}_{f_{\text{std}}}(\mathcal{S}) = 100\%$ and $\text{Acc}_{f_{\text{std}}}(\mathcal{J}) = 0\%$ by construction (and analogously for f_{rob}). As shown in Table 1, the Jammed Set $\mathcal{J}$ consistently accounts for a significant portion of the error manifold.

The consistency of these proportions (4.3%–5.0% for $\mathcal{J}$) across architectures suggests that natural jamming is not confined to a single evaluated model. However, this statistical stability does not necessarily imply that the "Jammed" status is an intrinsic property of the image. If the effect stems from universal data pathologies, the samples in $\mathcal{J}$ should overlap significantly across different models. Conversely, if it is a manifestation of specific inductive biases, these sets should diverge. We investigate this question by dissecting the spectral and spatial dependencies of these features in the next section.

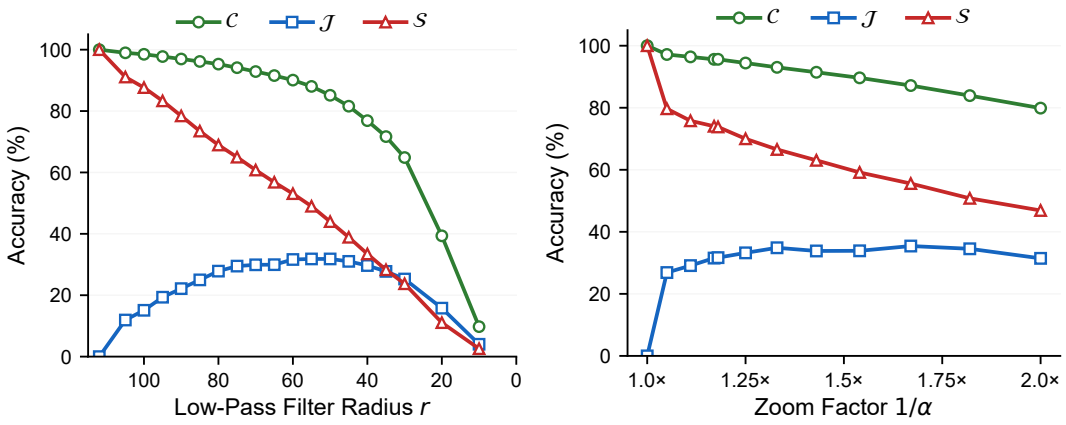

Figure 2: **The Duality of Sensitivity.** (a) In the spectral domain, a *Cleansing Paradox* is observed: removing high-frequencies restores accuracy for the Jammed Set ($\mathcal{J}$) while destroying it for the Spurious Set ($\mathcal{S}$). (b) In the spatial domain, zooming induces an *Sensitivity Inversion* transition, showing the reversal of dominance between $\mathcal{J}$ and $\mathcal{S}$ features. All curves use the primary lens.

3 Spectral-Spatial Locking: The Core Evidence

The divergent behaviors of f_{std} and f_{rob} on the Jammed and Spurious sets suggest that these models prioritize different structural components within the same image. In this section, we conduct a series of diagnostic experiments to isolate whether these non-robust features are defined by their frequency content, their spatial location, or a combination of both. Our results lead to the discovery of a **Spectral-Spatial Locking** mechanism.

Evaluation convention. All sets $\mathcal{C}, \mathcal{S}, \mathcal{J}, \mathcal{G}$ are defined using predictions on the original clean inputs. We then apply transformations and evaluate predictions on the same fixed subset of samples, without re-computing set membership after transformation. Unless stated otherwise, all experiments in this section use the primary lens (standard vs. robust ResNet-50, ℓ_2, $\epsilon = 1$; Appendix A.1).

3.1 The Cleansing Paradox: Spectral Sensitivity

We begin by testing whether the interference in the Jammed Set is an intrinsic property of high-frequency (HF) patterns. We apply a low-pass filter (LPF) to progressively remove HF components and observe the accuracy of the standard model f_{std}.

Protocol. For each image x, we compute its 2D Fourier transform $\mathcal{F}(x)$, apply a circular hard-cutoff mask M_r with varying radius $r \in [10, 112]$ (on a 224×224 spectrum), and reconstruct the filtered image $\tilde{x}_r = \mathcal{F}^{-1}(M_r \cdot \mathcal{F}(x))$. A smaller r corresponds to more aggressive high-frequency removal.

Results. As shown in Figure 2(a), removing HF content yields opposite outcomes for the two sets: **Spurious Set ($\mathcal{S}$):** Accuracy collapses from 100.0% to 23.7% at $r = 30$. This confirms that these predictions rely entirely on high-frequency shortcuts. **Jammed Set ($\mathcal{J}$):** Accuracy *recovers* from 0.0% to 31.8% at $r = 50$. **Common Set ($\mathcal{C}$):** Accuracy degrades gracefully from 100.0% to 85.1% at $r = 50$, demonstrating that robust semantic features are largely preserved in the low-frequency skeleton. These findings are diagnostic evidence of the *spectral fragility* of non-robust features: their predictive utility is tied to specific frequency supports (HF patterns).

3.2 The Sensitivity Inversion: Spatial Localization

To determine where these jammers are located, we examine model sensitivity to spatial transformations.

Protocol. We apply a center-crop zoom T_α, where the image is cropped to $\alpha \times 224$ pixels and resized back to 224×224 with bilinear interpolation. We sweep the ratio α from 1.0 (original) to 0.5 ($2\times$ zoom).

Table 2: **Frequency Grafting: Toxicity Transfer Test.** We graft the HF of Jammed images onto Common images. *Native Toxicity* represents the accuracy gain when HF is removed from the source host. *Transferred Toxicity* measures the rate at which the wrong label of the jammer infects a new semantic host.

TRANSFER SCENARIO	NATIVE TOXICITY (0% → X%)	TRANSFERRED TOXICITY (ON NEW HOST)
$\mathcal{J} \to \mathcal{C}$	31.8% (0%→31.8%)	**0.17%**

Results. Figure 2(b) reveals a distinct **Sensitivity Inversion**: as we zoom in, accuracy on **Jammed Set ($\mathcal{J}$)** increases while accuracy on **Spurious Set ($\mathcal{S}$)** decreases. Since zooming removes peripheral regions, this pattern is consistent with the hypothesis that many interfering cues are spatially localized near image boundaries or object context, where they can conflict with central semantic evidence. The accuracy on **Common Set ($\mathcal{C}$)** remains relatively stable, gracefully decreasing from 100% to 79.9%. Robust semantic features exhibit geometric invariance. This sensitivity highlights the *spatial fragility* of the underlying non-robust signals. This recovery is not specific to the resampling method: it appears with all six interpolation kernels we tested and when a 2% border is painted over instead of cropped (Appendix D).

3.3 Frequency Grafting: A Causality Test

The preceding experiments suggest a strong coupling between spectral content and spatial structure. To provide causal evidence, we test whether the interference in $\mathcal{J}$ is an intrinsic HF signal (analogous to a universal adversarial perturbation) or if its efficacy is strictly dependent on the host image's geometry. We perform a frequency grafting experiment that transplants the HF components of a Jammed image onto the LF skeleton of a Common image.

Protocol. Let $x_j \in \mathcal{J}$ be a Jammed image and $x_c \in \mathcal{C}$ be a Common image. We construct a hybrid image x_{hybrid} by combining their spectral components as follows:

$$x_{\text{hybrid}} = \mathcal{F}^{-1}\left(M_r \odot \mathcal{F}(x_c) + (1 - M_r) \odot \mathcal{F}(x_j)\right) \tag{2}$$

where $\mathcal{F}$ denotes the 2D Discrete Fourier Transform, $\odot$ represents element-wise multiplication, and M_r is a circular low-pass binary mask with radius $r = 50$. This operation preserves the semantic skeleton of the Common image while grafting onto it the exact high-frequency textures of the Jammed image. We define *transferred toxicity* as the fraction of random pairs (x_j, x_c) for which the hybrid inherits the source jammer's predicted class, i.e., $f_{\text{std}}(x_{\text{hybrid}}) = f_{\text{std}}(x_j)$.

Results. As shown in Table 2, the results demonstrate a near-complete collapse of predictive influence upon transplantation. While suppressing high-frequency components yields a 31.8 percentage-point recovery on the native jammed inputs (Section 3.1), the transferred toxicity to a foreign host is only 0.17% (averaged over five random pairings), whereas reassembling a jammer from its own spectral components preserves its jammed prediction in 99.9% of cases (Appendix D).

This result confirms that Natural Jammers are not independently "toxic" patterns. Their capacity to mislead the model is not an intrinsic property of the signal but emerges from a *fragile resonance* with the host image's spatial structure. Once the spectral-spatial lock is broken by moving the HF signal to other host images, the feature becomes inefficient. This causal evidence shifts the understanding of non-robustness from a feature-level defect to a systemic interaction between spectrum and space. A margin-matched control for the alternative that jammed images simply lie near the decision boundary is reported in Appendix D.

3.4 Formal Definition: Spectral–Spatial Locking

The empirical evidence in the preceding sections reveals that natural non-robustness is neither a purely spectral defect nor a static spatial anomaly. Instead, it manifests as a structural dependency where the predictive utility of non-robust features is conditioned on the alignment of frequency and space. We formalize this mechanism as follows:

Definition 3.1 (Spectral–Spatial Locking)**.** Let f be a classifier and $x \in \mathcal{X}$ denote an input sample. A non-robust predictive signal $\mathbf{z}$ is said to be *spectral–spatially locked* if its contribution to the

Table 3: **Complete Cross-Model Landscape.** Analysis on ImageNet validation set ($N = 50,000$). Within each architecture block, f_{std} is fixed to that architecture's standard model; each row uses a different robust model f_{rob} to define the quad-set partition on clean inputs.

Configuration	Jammed ($\mathcal{J}$)	Spurious ($\mathcal{S}$)	Common ($\mathcal{C}$)	Uninformative ($\mathcal{G}$)	Rob. Acc.
ResNet-18		*Standard Accuracy: 69.76%*			
$\ell_2, \epsilon = 0.5$	2,639 (5.28%)	4,786 (9.57%)	30,094 (60.19%)	12,481 (24.96%)	65.5%
$\ell_2, \epsilon = 1.0$	2,448 (4.90%)	6,168 (12.34%)	28,712 (57.42%)	12,672 (25.34%)	62.3%
$\ell_2, \epsilon = 3.0$	1,924 (3.85%)	10,240 (20.48%)	24,640 (49.28%)	13,196 (26.39%)	53.1%
$\ell_\infty, \epsilon = 1.0$	2,502 (5.00%)	5,659 (11.32%)	29,221 (58.44%)	12,618 (25.24%)	63.5%
$\ell_\infty, \epsilon = 4.0$	1,867 (3.73%)	10,497 (21.00%)	24,383 (48.77%)	13,253 (26.51%)	52.5%
ResNet-50		*Standard Accuracy: 76.15%*			
$\ell_2, \epsilon = 0.5$	2,317 (4.63%)	3,803 (7.61%)	34,270 (68.54%)	9,610 (19.22%)	73.2%
$\ell_2, \epsilon = 1.0$	2,140 (4.28%)	5,002 (10.00%)	33,071 (66.14%)	9,787 (19.57%)	70.4%
$\ell_2, \epsilon = 3.0$	1,807 (3.61%)	8,463 (16.93%)	29,610 (59.22%)	10,120 (20.24%)	62.8%
$\ell_\infty, \epsilon = 1.0$	2,252 (4.50%)	4,306 (8.61%)	33,767 (67.53%)	9,675 (19.35%)	72.0%
$\ell_\infty, \epsilon = 4.0$	1,812 (3.62%)	7,948 (15.90%)	30,125 (60.25%)	10,115 (20.23%)	63.9%
WideResNet-50-2		*Standard Accuracy: 78.48%*			
$\ell_2, \epsilon = 0.5$	2,202 (4.40%)	3,887 (7.77%)	35,354 (70.71%)	8,557 (17.11%)	75.1%
$\ell_2, \epsilon = 1.0$	2,167 (4.33%)	4,698 (9.40%)	34,543 (69.09%)	8,592 (17.18%)	73.4%
$\ell_2, \epsilon = 3.0$	1,910 (3.82%)	7,701 (15.40%)	31,540 (63.08%)	8,849 (17.70%)	66.9%
$\ell_\infty, \epsilon = 1.0$	2,232 (4.46%)	4,145 (8.29%)	35,096 (70.19%)	8,527 (17.05%)	74.7%
$\ell_\infty, \epsilon = 4.0$	2,000 (4.00%)	7,037 (14.07%)	32,204 (64.41%)	8,759 (17.52%)	68.4%
ViT-B/16		*Standard Accuracy: 81.07%*			
$\ell_\infty, \epsilon = 1.0$	1,569 (3.14%)	5,467 (10.93%)	35,067 (70.13%)	7,897 (15.79%)	73.3%
ViT-B/32		*Standard Accuracy: 75.91%*			
$\ell_\infty, \epsilon = 1.0$	1,828 (3.66%)	6,511 (13.02%)	31,443 (62.89%)	10,218 (20.44%)	66.5%

model's inference depends jointly on its frequency support and its spatial configuration. Formally, $\mathbf{z}$ is characterized by the following properties:

1. **Spectral sensitivity.** There exists a low-pass operator $\mathcal{L}_\sigma$ that suppresses frequencies above σ such that removing these components changes the prediction, i.e., $f(\mathcal{L}_\sigma(x)) \neq f(x)$ for a non-trivial fraction of inputs exhibiting the signal.

2. **Spatial Fragility.** Let $\mathcal{T}$ be a group of semantics-preserving geometric transformations. There exists $T \in \mathcal{T}$ such that $f(\tau(x)) \neq f(x)$, demonstrating that the efficacy of $\mathbf{z}$ is sensitive to precise spatial alignment.

3. **Non-Transferability.** Let $x_A, x_B \in \mathcal{X}$ be two distinct samples. Grafting the spectral components of x_A onto the spatial structure of x_B fails to reproduce the original prediction bias, as the non-robust signal becomes functionally inert when the original lock is broken.

Definition 3.1 establishes that individual jammers are fragile configurations that collapse when their spectral–spatial lock is disrupted. However, this fragility presents a paradox: if specific locks are so easily broken, why does the aggregate size of the Jammed Set remain approximately stable across models and augmentations? This indicates a redistribution of non-robustness across the data manifold. We will explore this in the next section.

4 Substitution Paradox: Systemic Redistribution of Natural Non-Robustness

Section 3 established that many clean-image failures of standard models are consistent with *spectral–spatial locking*: interfering cues are fragile under mild spectral or geometric perturbations. This

fragility raises a natural question: if individual jammers are so easily disrupted, why does non-robust vulnerability persist at scale?

This section provides system-level evidence that natural non-robustness is not confined to a small set of "bad images" but is a persistent outcome of the model–data interaction. We focus on three questions: **Q1:** Is the prevalence of $\mathcal{J}$ and $\mathcal{S}$ stable across architectures and robustness recipes? **Q2:** Are jammed/spurious identities intrinsic to images, or model-dependent? **Q3:** Under tiny semantics-preserving transformations, do these pathologies disappear or get redistributed?

Assumption 4.1 (Robust Models as Analysis Lenses)**.** Robust models are used as analysis lenses to induce the quad-set partition via disagreement with a standard model. We do not treat any single robust model as a ground-truth oracle; instead, we test whether conclusions persist across multiple robust oracles spanning architectures and robustness recipes.

Experimental Setup. All analyses are conducted on ImageNet validation ($N = 50{,}000$). For each architecture family, we use its corresponding standard model as f_{std} and a collection of robust models as candidate lenses f_{rob} that vary in threat model (ℓ_2 or ℓ_∞) and budget ϵ. *Within each architecture block*, f_{std} is fixed to that architecture's standard model, and each robust model defines a quad-set partition $\mathcal{C}, \mathcal{S}, \mathcal{J}, \mathcal{G}$ according to Section 2 on the original clean inputs. Implementation details (model sources/checkpoints, preprocessing, and full configuration list) are provided in the appendix.

4.1 Prevalence Across Architectures

We first establish a baseline fact that will serve as a foundation for the remainder of the section: natural jamming is not rare, nor specific to a single robustness recipe. Table 3 reports the quad-set partition induced by 17 robust lenses spanning CNNs and ViTs.

Across a wide range of architectures and robustness recipes, the Jammed Set consistently occupies a few percent of ImageNet validation set , and the Spurious Set occupies a larger but still substantial portion (roughly $\sim$7–20%, depending on ϵ). Natural jamming is thus not an isolated artifact of a particular model or robustness configuration; across these lenses, $|\mathcal{J}|$ is relatively stable compared to $|\mathcal{S}|$.

4.2 Model-Specific Vulnerabilities

We next test whether jammed is an intrinsic property of the image or depends on the model used to define the partition. For two induced Jammed sets $\mathcal{J}_A$ and $\mathcal{J}_B$, we quantify agreement via Intersection-over-Union (IoU):

$$\text{IoU}(\mathcal{J}_A, \mathcal{J}_B) = \frac{|\mathcal{J}_A \cap \mathcal{J}_B|}{|\mathcal{J}_A \cup \mathcal{J}_B|},$$

and analogously for $\mathcal{S}$.

Agreement is modest within architecture and low across architectures. Figure 3 shows that even within the same architecture, changing only the robustness budget yields moderate agreement (IoU ≈ 0.40–0.47). Across architectures, IoU collapses (typically < 0.15), and can be even smaller between CNNs and ViTs. These results indicate that *which* images are jammed/spurious depends strongly on model inductive bias and robustness recipe, rather than being image-intrinsic.

While IoU captures *identity*, confidence helps assess *risk*: jammed errors are often high-confidence for standard models. Table 4 reports mean softmax confidence of f_{std} and the corresponding robust lens f_{rob} on the induced sets. Across CNNs, f_{std} is typically *more confident* than f_{rob} on $\mathcal{J}$, consistent with a brittle but dominant cue overwhelming semantic evidence. ViTs show a reversal (robust confidence $>$ standard confidence), which may reflect architecture-specific interfering cues; however, the ViT and CNN lenses also differ in training recipe, so this comparison cannot separate the two factors.

4.3 The Substitution Paradox

Section 3 demonstrated that individual jammed behaviors can be neutralized by mild spectral filtering or small geometric shifts, suggesting that jamming depends on fragile spectral–spatial alignment.

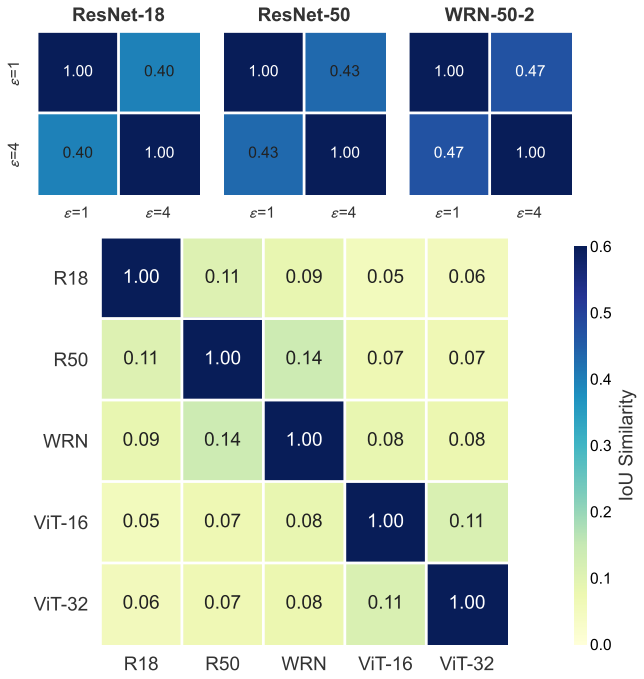

Figure 3: **Jammed Set IoU Analysis (ℓ_∞).** *Top row*: intra-architecture agreement when changing only robustness budget ($\epsilon = 1$ vs $\epsilon = 4$). *Bottom*: cross-architecture IoU matrix. Agreement is modest within the same architecture and collapses across architectures, indicating that jammed identity is model-dependent.

Table 4: **Confidence Comparison: Standard vs Robust Models.** Each cell reports f_{std} / f_{rob} mean confidence (max softmax on predicted class). ViT entries use ℓ_∞, ϵ=1.0 only.

		JAMMED ($\mathcal{J}$)			SPURIOUS ($\mathcal{S}$)		
CONFIG	ϵ	RN-18	RN-50	WRN-50-2	RN-18	RN-50	WRN-50-2
ℓ_2	0.5	.457 / .440	.511 / .509	.545 / .556	.571 / .349	.624 / .412	.684 / .459
ℓ_2	1.0	.470 / .402	.517 / .464	.555 / .531	.601 / .303	.661 / .368	.710 / .427
ℓ_2	3.0	.491 / .318	.544 / .384	.574 / .449	**.678** / .217	**.726** / .267	**.770** / .323
ℓ_∞	1.0	.468 / .418	.513 / .485	.554 / .550	.588 / .316	.643 / .382	.698 / .445
ℓ_∞	4.0	.490 / .319	.538 / .394	.577 / .464	**.680** / .206	**.717** / .270	**.760** / .321
VIT-B/16 (ℓ_∞=1.0)			.504 / **.701**			.640 / .584	
VIT-B/32 (ℓ_∞=1.0)			.447 / **.670**			.591 / .541	

We now formalize and test a stronger claim: *tiny transformations can break many specific jammed identities, yet the overall prevalence of jamming remains approximately conserved.*

Definition 4.2 (Transformed Partition and Substitution Statistics)**.** Fix an architecture and its standard model f_{std}, together with a robust lens f_{rob} that induces the original sets $\mathcal{C}, \mathcal{S}, \mathcal{J}, \mathcal{G}$ on clean inputs.

For a transformation T, define the transformed Jammed set by evaluating both models on $T(x)$:

$$\mathcal{J}_T := \{x : f_{\text{std}}(T(x)) \neq y \ \wedge \ f_{\text{rob}}(T(x)) = y\},$$

and analogously $\mathcal{S}_T, \mathcal{C}_T, \mathcal{G}_T$. We quantify (i) *identity stability* via $\text{IoU}(\mathcal{J}, \mathcal{J}_T)$, and (ii) *prevalence stability* via the transformed set sizes $|\mathcal{J}_T|$ and $|\mathcal{S}_T|$.

Proposition 4.3 (Substitution Paradox (Empirical))**.** *For small semantics-preserving geometric transformations T (e.g., $\leq 8\%$ edge crop with resize), we observe simultaneously:*

1. ***Identity instability:*** $\text{IoU}(\mathcal{J}, \mathcal{J}_T)$ *(and* $\text{IoU}(\mathcal{S}, \mathcal{S}_T)$*) is low.*

Table 5: **Structural Fragility + Substitution Paradox.** Minimal edge crops substantially change the *identity* of jammed/spurious samples (IoU drops sharply), yet the *prevalence* (set size) remains approximately stable.

	SET SIZE		IOU VS. ORIGINAL	
CROP	$\lvert\mathcal{J}_T\rvert$	$\lvert\mathcal{S}_T\rvert$	$\mathcal{J}$	$\mathcal{S}$
0% (ORIGINAL)	2,140	5,002	1.00	1.00
2%	2,289	5,193	**0.46**	0.58
4%	2,268	5,208	0.42	0.54
6%	2,276	5,127	0.38	0.51
8%	2,291	5,159	0.35	0.49

2. ***Prevalence stability:*** *the set sizes $|\mathcal{J}_T|$ and $|\mathcal{S}_T|$ remain approximately unchanged relative to $|\mathcal{J}|$ and $|\mathcal{S}|$.*

*Thus, small geometric changes tend to **redistribute** jammed/spurious identities across samples rather than eliminate non-robust vulnerability.*

We apply minimal center crops removing only 2%, 4%, 6%, or 8% of edge pixels and resize back to 224×224. For each crop ratio, we compute the transformed partition (Definition 4.2) and report both set sizes and IoU against the original partition.

Identity instability (structural fragility): Even a 2% crop yields $\mathrm{IoU}(\mathcal{J}, \mathcal{J}_T) = 0.46$, implying that about one third of the original jammed images leave $\mathcal{J}$, and a similar number of new ones enter, under an almost imperceptible geometric perturbation. This supports the view that the underlying cues depend on fragile spatial alignment, consistent with spectral–spatial locking.

Prevalence stability (substitution): Despite the large change in identity, the transformed set sizes remain stable: $|\mathcal{J}_T| \approx 2.2$k and $|\mathcal{S}_T| \approx 5.1$k across crop ratios. This stability is approximate rather than exact: in a re-run that tracks individual images, a 2% crop moves 738 of the 2,140 jammers out of $\mathcal{J}$ and 873 new images in. Thus, breaking many specific locks does not reduce overall vulnerability; instead, new jammed/spurious identities emerge elsewhere. This resolves the tension raised by Section 3: *individual* jammers are fragile, yet *aggregate* non-robustness persists as a systemic property and is redistributed under small transformations. Similar turnover appears under mild low-pass filtering and on six further datasets (Appendix E); our claim is limited to these tested mild transformations (Appendix F). In this view, natural non-robustness is a systemic **"tax"** that the model must pay to achieve high training accuracy.

We formalized and verified the *Substitution Paradox*: tiny semantics-preserving geometric changes drastically alter jammed/spurious identities while leaving their prevalence approximately unchanged.

5 Related Work

5.1 Adversarial Robustness and Non-robust Features

The vulnerability of deep neural networks to adversarial examples Szegedy et al. [2013], Goodfellow et al. [2014], Madry et al. [2017] has been increasingly understood through the lens of data-driven features. Ilyas et al. [2019] famously posited that adversarial examples arise from *non-robust features*—patterns that are highly predictive yet imperceptible to humans—and that these features are actively *useful* for standard classification. This framework has led to a prevailing view of non-robust features as "spurious shortcuts" Geirhos et al. [2020], Shah et al. [2020] that models exploit to achieve high standard accuracy, often resulting in an inherent trade-off between robustness and clean performance Tsipras et al. [2018], Zhang et al. [2019], Rebuffi et al. [2021].

This view has since been actively debated. Kim et al. [2021] show via information bottleneck that non-robust features are decomposable, with only a subset driving adversarial predictions. Li et al. [2023] argue that non-robust features are paradigm-specific shortcuts useful only under supervised objectives, not across learning paradigms. Crucially, however, both works study features under *intentional adversarial perturbation*—inputs explicitly crafted to expose model vulnerabilities.

We ask a fundamentally different question: can non-robust features mislead standard models on *clean, unmodified natural images*, without any perturbation? We identify a Jammed Set ($\mathcal{J}$) where robust models outperform standard models on standard test images, and provide causal evidence for Spectral–Spatial Locking as the driving mechanism. The antagonism we document is not injected by an attacker but arises spontaneously from the interaction between standard training and the spectral-spatial structure of natural data. Our Substitution Paradox further reveals this to be a redistributing property of the model–data geometry: a model pair defines a failure field $j(x) = \mathbf{1}[f_{\text{std}}(x) \neq y \wedge f_{\text{rob}}(x) = y]$ over natural images, which are commonly assumed to lie near a low-dimensional data manifold Fefferman et al. [2016], Stutz et al. [2019], and small transformations move the support of j while approximately preserving its mass. This provides a complementary framework to studies of natural adversarial or hard examples Hendrycks et al. [2021], Recht et al. [2019], Li et al. [2024], Liu et al. [2025b], Zhang et al. [2025], which rely on curated atypical subsets rather than identifying a structural mechanism within standard uncurated data.

5.2 Spectral and Geometric Bias in Generalization

The inductive biases of CNNs are often characterized by their sensitivity to specific input frequencies and spatial structures. From a Fourier perspective, models frequently exhibit a *spectral bias*, relying on high-frequency components that are critical for generalization but lack robustness Yin et al. [2019], Wang et al. [2020], Fang and Xu [2024], Krizhevsky et al. [2012], Xia et al. [2024], Zhang et al. [2023b]. Simultaneously, the lack of geometric invariance—often caused by aliasing in downsampling layers Zhang [2019]—renders models fragile to even minor image transformations Azulay and Weiss [2019], Michaeli and Soudry [2025]. This fragility is often linked to a "texture bias," where models prioritize local patterns over global shapes Geirhos et al. [2018], Hermann et al. [2020].

Our work synthesizes these views through the mechanism of Spectral–spatial Locking, providing causal evidence that natural jamming is contingent on the precise alignment of frequency and space. Furthermore, we address the stability of these vulnerabilities, i.e., the Substitution Paradox where the aggregate rate of jamming remains stable across the data manifold. This observation relates to the broader problem of underspecification D'Amour et al. [2022], where models with identical training objectives exhibit divergent and unpredictable failure modes under distribution shifts Taori et al. [2020], Arjovsky et al. [2019], Sagawa et al. [2019].

6 Conclusion

We show that non-robust features on *clean* natural images can be antagonistic: on a Jammed Set, robust models outperform standard models without modifying inputs. Using a standard-vs-robust partition of ImageNet-Val, we characterize Jammed and Spurious regimes and provide evidence for *spectral–spatial locking* via frequency filtering, geometric perturbations, and frequency grafting.

At the system level, jammed identities are highly model-dependent while prevalence remains stable, yielding the *Substitution Paradox*: small semantics-preserving transformations reshuffle jammed membership without reducing its overall rate.

Acknowledgments and Disclosure of Funding

This research / project is supported by A*STAR under its Japan-Singapore Joint Call: JST-A*STAR 2024 (Project ID: R2416IR139). This research is also supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award MOE-T2EP20125-0005). This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). The authors declare no competing interests.

References

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

A. Azulay and Y. Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1–25, 2019.

L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461, 2014.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

M. Eriksson, E. Purificato, A. Noroozian, J. Vinagre, G. Chaslot, E. Gomez, and D. Fernandez-Llorca. Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation. *arXiv preprint arXiv:2502.06559*, 2025.

R. Fang and Y. Xu. Addressing spectral bias of deep neural networks by multi-grade deep learning. *Advances in Neural Information Processing Systems*, 37:114122–114146, 2024.

C. Fefferman, S. Mitter, and H. Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852.

R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

B. Gyevnar and A. Kasirzadeh. Ai safety for everyone. *Nature Machine Intelligence*, pages 1–12, 2025.

D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021.

K. Hermann, T. Chen, and S. Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in neural information processing systems*, 33:19000–19015, 2020.

A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

E. Kafali, T. Semertzidis, and P. Daras. Dynamic trade-offs in adversarial training: Exploring efficiency, robustness, forgetting, and interpretability. *Neural Processing Letters*, 57(3):47, 2025.

J. Kauffmann, J. Dippel, L. Ruff, W. Samek, K.-R. Müller, and G. Montavon. Explainable ai reveals clever hans effects in unsupervised learning models. *Nature Machine Intelligence*, pages 1–11, 2025.

J. Kim, B.-K. Lee, and Y. M. Ro. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems*, 34:17148–17159, 2021.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

A. Li, Y. Wang, Y. Guo, and Y. Wang. Adversarial examples are not real features. *Advances in Neural Information Processing Systems*, 36:17222–17237, 2023.

B. Li, Z. Lin, W. Peng, J. d. D. Nyandwi, D. Jiang, Z. Ma, S. Khanuja, R. Krishna, G. Neubig, and D. Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068, 2024.

J. Li and G. Li. Triangular trade-off between robustness, accuracy, and fairness in deep neural networks: A survey. *ACM Computing Surveys*, 57(6):1–40, 2025.

C. Liu, Y. Dong, W. Xiang, X. Yang, H. Su, J. Zhu, Y. Chen, Y. He, H. Xue, and S. Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *International Journal of Computer Vision*, 133(2):567–589, 2025a.

L. Liu, Y. Liang, X. Yan, L. Huangfu, S. Samtani, Z. Yu, Y. Zhang, and D. D. Zeng. Hard sample mining: A new paradigm of efficient and robust model training. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

H. Michaeli and D. Soudry. Alias-free vit: Fractional shift invariance via linear attention. *arXiv preprint arXiv:2510.22673*, 2025.

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729, 2008.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

H. Salman, A. Ilyas, L. Engstrom, A. Kapoor, and A. Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

J. Sang, X. Zhao, J. Zhang, and Z. Lin. Benign adversarial attack: Tricking models for goodness. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6883–6889, 2022.

H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

D. Stutz, M. Hein, and B. Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

S. Suzuki, S. Yamaguchi, S. Takeda, S. Kanai, N. Makishima, A. Ando, and R. Masumura. Adversarial finetuning with latent representation constraint to mitigate accuracy-robustness tradeoff. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4367–4378. IEEE, 2023.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599, 2020.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

H. Wang, X. Wu, Z. Huang, and E. P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8684–8694, 2020.

J. Xia, Z. Yue, Y. Zhou, Z. Ling, Y. Shi, X. Wei, and M. Chen. Waveattack: Asymmetric frequency obfuscation-based backdoor attacks against deep neural networks. *Advances in Neural Information Processing Systems*, 37:43549–43570, 2024.

J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

J. Zhang, J. Sang, Q. Yi, Y. Yang, H. Dong, and J. Yu. Imagenet pre-training also transfers non-robustness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3436–3444, 2023a.

J. Zhang, Q. Yi, D. Lu, and J. Sang. Low-mid adversarial perturbation against unauthorized face recognition system. *Information Sciences*, 648:119566, 2023b.

J. Zhang, F. Liu, D. Zhou, J. Zhang, and T. Liu. Improving accuracy-robustness trade-off via pixel reweighted adversarial training. *arXiv preprint arXiv:2406.00685*, 2024.

R. Zhang. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pages 7324–7334. PMLR, 2019.

W. Zhang, P. L. K. Ding, and H. Liu. Beyond imagenet: Understanding cross-dataset robustness of lightweight vision models. *arXiv preprint arXiv:2511.00335*, 2025.

K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

A Appendix

A.1 Models, Robust Lenses, and Compute

Standard models. All standard CNN models use torchvision official ImageNet-1K pretrained weights. Standard ViT models (ViT-B/16, ViT-B/32) are initialized from torchvision ImageNet-1K pretrained weights.

Robust CNN lenses. All robust CNN checkpoints are taken from the public Madry-Lab ImageNet robust model release on HuggingFace (`https://huggingface.co/madrylab/robust-imagenet-models`), corresponding to Salman et al. [2020]. They trained all the models from scratch using SGD with batch size of 512, momentum of 0.9, and weight decay of 10^{-4}. They trained for 90 epochs with an initial learning rate of 0.1 that drops by a factor of 10 every 30 epochs with PGD-3. For ℓ_2 budgets, we follow the ϵ convention used in that release/paper. For ℓ_∞ budgets, we report ϵ in pixel units, i.e., $\epsilon \in \{1, 4\}$ corresponds to $\{1/255, 4/255\}$.

Robust ViT lenses. We perform adversarial training on ViT-B/16 and ViT-B/32 models, initialized with weights pre-trained on ImageNet-1k from torchvision. The models are fine-tuned for 100 epochs with batch size of 1024 using the AdamW optimizer with a base learning rate of 1×10^{-4} and weight decay of 0.05 using PGD-3. We employ a cosine annealing learning rate schedule with a 10-epoch linear warmup.

Primary robust lens. Unless otherwise stated, the **primary robust lens** used to define the quad-set partition in core experiments is a robust ResNet-50 with ℓ_2 budget $\epsilon = 1$ (MadryLab release), paired with the corresponding standard ResNet-50.

Compute Resources. The experiments in the paper are dominated by inference rather than training. All diagnostic evaluations, including the spectral/spatial transformations, quad-set recomputation, S^3P sweeps, and representation analyses, were run on a single NVIDIA A100 80GB GPU and each completed within tens of minutes. The only substantial training cost is the adversarial fine-tuning of the robust ViT lenses, which required on the order of tens of A100 GPU-hours.

A.2 Two Evaluation Protocols (Set Membership)

We use two complementary evaluation conventions.

Protocol P1 (fixed-membership evaluation). We first define the quad-set partition $(\mathcal{C}, \mathcal{S}, \mathcal{J}, \mathcal{G})$ using predictions on the original clean inputs, then apply a transformation T and evaluate model behavior *on the same fixed subset*, without recomputing set membership. This protocol is used for the sensitivity curves to isolate how a transformation affects a particular pathology.

Protocol P2 (recomputed partition under transformation). For a transformation T, we define a transformed partition $(\mathcal{C}_T, \mathcal{S}_T, \mathcal{J}_T, \mathcal{G}_T)$ by evaluating both models on $T(x)$ (Definition 4.2 in the main paper). This protocol is used for the Substitution Paradox (prevalence vs. identity stability under transformations).

A.3 Transform Operators Used in the Paper

We describe the three core transform families used throughout the paper. Each operator is applied to the image, and the resulting image is evaluated under the standard ImageNet evaluation pipeline before being fed to a model.

Low-pass filter (LPF). Given an image x of spatial resolution 224×224, let $\mathcal{F}$ denote the 2D discrete Fourier transform applied channel-wise. Let M_r be a circular binary mask in frequency space centered at the DC component, with radius r (in pixels of the frequency grid). The LPF operator is

$$T_f^{(r)}(x) \;=\; \mathcal{F}^{-1}(M_r \odot \mathcal{F}(x))\,. \tag{3}$$

Table 6: **Mini sweep for grafting cutoff radius selection (Jammed → Common).** All main grafting results use $r = 50$. This table is included only to justify the chosen operating point. The number of evaluated hybrids equals the number of Jammed samples under the primary lens ($n = 2140$).

Radius r	LPF Acc on $\mathcal{J}$ (%)	Infected (%)	Gap (%)
40	29.67	0.28	29.39
50	**31.82**	**0.00**	**31.82**
60	31.64	0.00	31.64

Zoom (center-crop with resize). For a retained crop ratio $\alpha \in (0, 1]$, define the zoom operator $T_g^{(\alpha)}$ by (i) taking a centered crop retaining an α fraction of the side length, and (ii) resizing back to 224×224 using bilinear interpolation.

Edge-crop for the Substitution Paradox. For a retained ratio $\rho \in (0, 1]$, we perform a center crop retaining ρ of each side length, then resize back to the original resolution. In Table 5 (main paper), `2%/4%/6%/8% edge crop` corresponds to $\rho \in \{0.98, 0.96, 0.94, 0.92\}$.

Naming note (Uninformative). In code and some internal analyses we use `Uninformative` to denote samples where both the standard and robust models are incorrect; this corresponds to the *Uninformative* set $\mathcal{G}$ in the main text.

A.4 Remark: Mechanistic Interpretation via Spectral–Spatial Coupling

Remark A.1 (Mechanistic Interpretation via Spectral–Spatial Coupling). Definition 3.1 is intentionally formulated in operational terms to enable empirical verification. It also admits a mechanistic interpretation in terms of spectral–spatial coupling inside the network.

Let $\phi(x)$ denote an intermediate feature representation and let τ be a (small) semantics-preserving geometric transformation parameterized smoothly. Spectral–spatial locking can be viewed as arising from a non-trivial interaction between (i) frequency-selective sensitivity and (ii) spatial alignment sensitivity, e.g., via a non-zero mixed interaction of the form

$$\frac{\partial f(\phi(x))}{\partial \mathcal{F}(x)} \cdot \frac{\partial f(\phi(x))}{\partial \tau} \neq 0,$$

which expresses that the model's sensitivity to frequency components depends on their spatial alignment. This provides intuition for why jamming effects can be strong on a native host yet collapse when alignment is disrupted.

A.5 Choosing the Grafting Cutoff Radius r (Auxiliary Sensitivity)

The main paper fixes the grafting cutoff radius to $r = 50$. Here we provide a small auxiliary sweep to justify this choice as an operating point that maximizes the contrast between (i) recovery on Jammed samples under LPF and (ii) transferred toxicity under frequency grafting. This auxiliary analysis does not change any main-text results.

Quantities. Let $x_j \in \mathcal{J}$ and $x_c \in \mathcal{C}$ denote Jammed and Common samples (defined on clean inputs under the primary lens).

- **LPF recovery on Jammed:** $\mathrm{Acc}_{\mathcal{J}}(T_f^{(r)})$, the standard-model accuracy on Jammed samples after applying LPF with radius r.
- **Transferred toxicity (infected rate):** the fraction of hybrids that adopt the Jammed donor's wrong label, i.e., $\Pr[f_{\text{std}}(x_{\text{hybrid}}) = f_{\text{std}}(x_j)]$.
- **Gap:** $\mathrm{Gap}(r) = \mathrm{Acc}_{\mathcal{J}}(T_f^{(r)}) - \mathrm{Infected}(r)$.

A larger gap indicates that high-frequency components are harmful on their native host but lose potency when transplanted, consistent with non-transferability.

Table 7: **S^{3}P Flag Rates on the Quad-Set Partition.** We report $\Pr[\text{flag} \mid x \in \mathcal{A}]$ for each set $\mathcal{A}$. The combined probe achieves high coverage on Jammed and Spurious samples, with complementary contributions from GSP and FSP.

SET	GSP	FSP	BOTH	S^{3}P (EITHER)
COMMON ($\mathcal{C}$)	5.9%	15.0%	–	17.8%
JAMMED ($\mathcal{J}$)	56.2%	65.1%	42.3%	**79.0%**
SPURIOUS ($\mathcal{S}$)	30.8%	56.2%	22.0%	**65.0%**

At $r = 50$, LPF yields the largest recovery on Jammed samples while the transferred toxicity is 0.00% over the $n = 2140$ hybrids of this sweep (0.17% averaged over five random pairings; Appendix D), supporting the view that the high-frequency cues driving jamming are strongly context-dependent.

B Spectral-Spatial Stability Probe: A Practical Tool

Sections 3–4 show that many non-robust behaviors on clean images are fragile under mild spectral and geometric perturbations, consistent with spectral–spatial locking and its redistribution under small transformations (the substitution paradox). We leverage this fragility to build a lightweight *diagnostic tool* S^{3}P that flags potentially non-robust samples at test time. It is intended as a practical *probe* rather than a new state-of-the-art classifier: it requires **no retraining**, and uses **only a standard model**.

B.1 The S^{3}P Framework

We define two fixed, semantics-preserving perturbations that are designed to disrupt fragile spectral–spatial locks:

Definition B.1 (Geometric Stability Probe (GSP))**.** Let T_g be a fixed center-zoom transform. Define the geometric instability indicator

$$b_g(x) := \mathbb{I}[f_{\text{std}}(T_g(x)) \neq f_{\text{std}}(x)]\,.$$

Definition B.2 (Frequency Stability Probe (FSP))**.** Let T_f be a fixed low-pass filtering transform. Define the frequency instability indicator

$$b_f(x) := \mathbb{I}[f_{\text{std}}(T_f(x)) \neq f_{\text{std}}(x)]\,.$$

Combined Probe (S^{3}P). We flag a sample if either probe detects instability:

$$\text{S}^3\text{P}(x) := b_g(x) \;\vee\; b_f(x).$$

We fix one default operating point throughout this section: T_g is a $1.27\times$ center zoom and T_f is a low-pass filter with radius $r = 50$. A sensitivity sweep is provided in Appendix B.3 below. S^{3}P uses 3 forward passes of f_{std} (original, T_g, T_f).

B.2 Empirical Validation

We evaluate S^{3}P on the quad-set partition induced by standard-vs-robust disagreement (Section 2). We report the conditional flag rate $\Pr[\text{flag} \mid x \in \mathcal{A}]$ for each set $\mathcal{A}$; for $\mathcal{C}$ this corresponds to a false-positive rate.

As shown in Table 7, GSP and FSP are complementary: many Jammed/Spurious samples are flagged by only one of the two probes. A chi-squared test confirms weak correlation between probe outcomes ($\phi = 0.24$, $p < 10^{-4}$), indicating that they capture different instability modes. We also find that S^{3}P flags are strongly associated with Jammed identity ($\chi^2 = 1963$, $p < 10^{-4}$; odds ratio $= 7.8$).

As a sanity check, trivial uncertainty signals from the same standard model (e.g., low maximum softmax probability or high predictive entropy) are also informative for flagging jammed/spurious samples; we report detailed comparisons and complementarity with S^{3}P in Appendix B.4 below.

Table 8: **S^{3}P vs. TTA (repurposed heuristic).** TTA is designed for accuracy improvement via averaging; when repurposed for instability flagging, it is much weaker than targeted stability probes at similar inference cost.

METHOD	JAMMED	SPURIOUS	INFERENCES
FIVECROP TTA	30.1%	9.7%	5
S^{3}P	**79.0%**	**65.0%**	3

Table 9: **Set sizes for MSP/Entropy analysis (primary lens).**

Set	Common ($\mathcal{C}$)	Jammed ($\mathcal{J}$)	Spurious ($\mathcal{S}$)	Uninformative ($\mathcal{G}$)
Count	33,071	2,140	5,002	9,787

Test-time augmentation (TTA) is designed to improve accuracy by averaging predictions across transformed views Shorten and Khoshgoftaar [2019]. For contrast, we repurpose TTA into an instability-flagging heuristic by flagging a sample when its FiveCrop-averaged prediction differs from its center-crop prediction. As shown in Table 8, this repurposed heuristic flags far fewer Jammed/Spurious samples than targeted stability probes.

Practical Use. S^{3}P is a drop-in diagnostic that can be used without retraining: (1) Pre-deployment screening: flag unstable test samples for manual review, abstention, or targeted curation. (2) Runtime warning: if S^{3}P flags instability, lower confidence, request verification, or trigger a fallback. (3) Cost/coverage trade-off: use GSP alone (2 inferences) for speed; add FSP (3 total) for broader coverage.

B.3 S^{3}P Hyperparameter Sweep

The S^{3}P probe has two hyperparameters: the zoom factor for GSP and the LPF radius for FSP. We perform a grid search over zoom factors and LPF radii and choose a single operating point used throughout the paper.

Figure 4 reports: Jammed detection, Spurious detection, false positive rate on the Common set, and a combined score. We select the operating point corresponding to a $1.27\times$ zoom and LPF radius $r = 50$.

B.4 MSP/Entropy Sanity Check and Complementarity with S^{3}P

This section compares S^{3}P to single-pass uncertainty signals from the same standard model f_{std}: (i) MSP $p_{\max}(x) = \max_k p(k \mid x)$ and (ii) predictive entropy $H(x) = -\sum_k p(k \mid x) \log p(k \mid x)$.

We report conditional flag rates on the quad-set partition induced by the primary lens. We also quantify complementarity via a *rescue rate*: among samples not flagged by MSP at threshold τ, what fraction are flagged by S^{3}P?

Complementarity via rescue rate. For an MSP threshold τ, define

$$\text{Rescue}(\mathcal{A}; \tau) = \frac{|\{x \in \mathcal{A} : \text{MSP}_\tau(x) = 0 \wedge \text{S}^3\text{P}(x) = 1\}|}{|\{x \in \mathcal{A} : \text{MSP}_\tau(x) = 0\}|}.$$

C Statistical Validation

We report statistical tests under the primary lens (ResNet-50 standard vs. robust ResNet-50 $\ell_2, \epsilon = 1$), $N = 50{,}000$ (ImageNet-Val).

- **Cleansing Paradox is significant:** LPF improves Jammed substantially more than it harms Common ($\chi^2 = 417.3$, $p = 9.33 \times 10^{-93}$; odds ratio $= 2.64$).

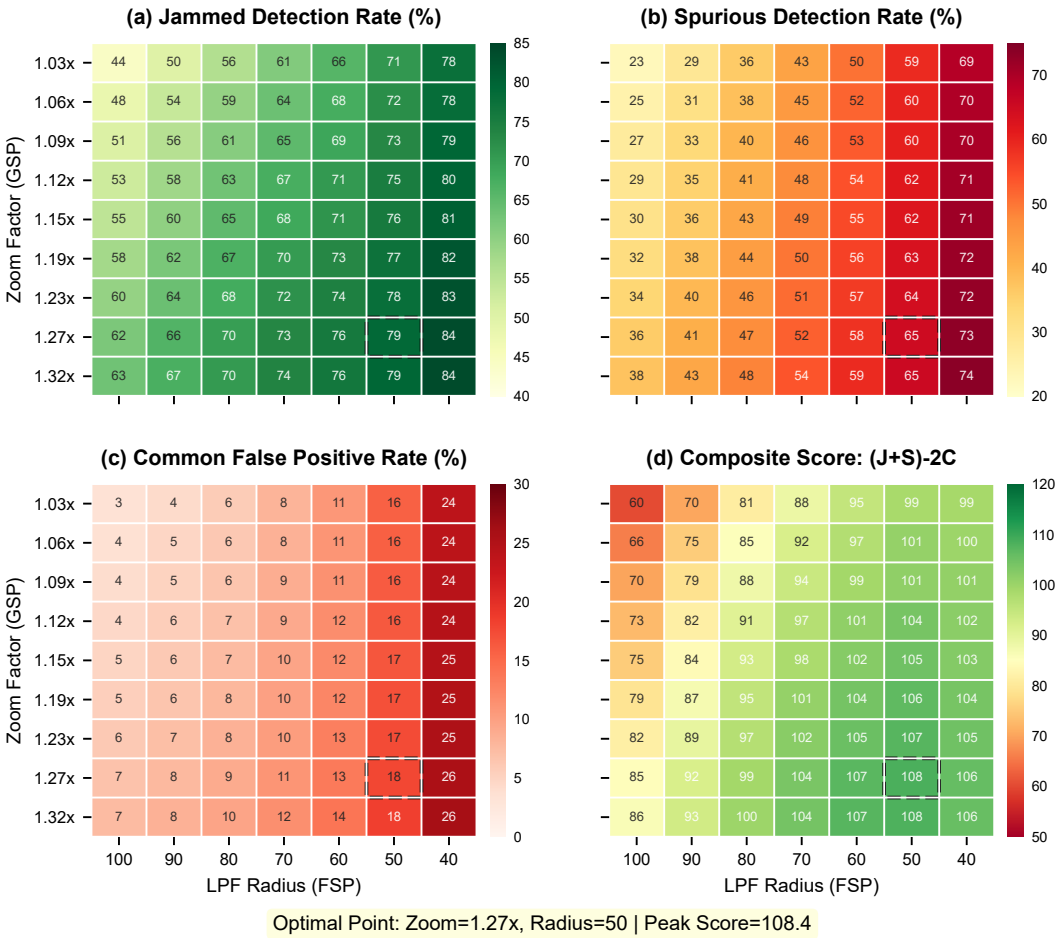

Figure 4: **S^{3}P hyperparameter sweep.** Detection rates for Jammed and Spurious sets, false positive rate on the Common set, and a combined score. The selected operating point ($1.27\times$ zoom, LPF radius $r = 50$) is highlighted. Primary lens.

Table 10: **Fixed-threshold MSP flag rates.** We flag if $p_{\max}(x) < \tau$. Numbers are $\Pr[\text{flag} \mid x \in \mathcal{A}]$.

MSP threshold τ	Common FP ($\mathcal{C}$)	Jammed ($\mathcal{J}$)	Spurious ($\mathcal{S}$)
0.9	25.7%	95.7%	77.1%
0.8	16.6%	89.5%	65.2%
0.7	11.0%	80.3%	53.4%
0.6	7.0%	66.3%	41.4%
0.5	3.7%	46.9%	27.4%

- **S^{3}P is associated with Jammed identity:** $\chi^2 = 1962.7$, $p < 10^{-4}$, odds ratio $= 7.83$.
- **S^{3}P is associated with Spurious identity:** $\chi^2 = 2297.8$, $p < 10^{-4}$, odds ratio $= 4.13$.
- **GSP and FSP are complementary on Jammed:** $\phi = 0.241$ (weak correlation), $\chi^2 = 123.2$, $p = 1.27 \times 10^{-28}$.

The Cleansing Paradox test supports the spectral-sensitivity component of the main paper's spectral–spatial locking claim. The remaining tests support the S^{3}P diagnostic tool (Appendix B).

Table 11: **Fixed-threshold entropy flag rates.** We flag if $H(x) > \tau$. Numbers are $\Pr[\text{flag} \mid x \in \mathcal{A}]$.

Entropy threshold τ	Common FP ($\mathcal{C}$)	Jammed ($\mathcal{J}$)	Spurious ($\mathcal{S}$)
0.5	23.7%	94.3%	76.3%
1.0	10.5%	66.0%	50.2%
1.5	5.1%	43.5%	32.0%
2.0	2.7%	27.9%	20.0%
2.5	1.4%	17.4%	12.0%

Table 12: **MSP–S^{3}P complementarity via rescue rate.** MSP flags if $p_{\max}(x) < \tau$. We report (i) MSP flag rate, (ii) S^{3}P rescue rate among MSP-missed samples, and (iii) union flag rate.

	Jammed ($\mathcal{J}$)			**Spurious** ($\mathcal{S}$)			**Common** ($\mathcal{C}$)		
MSP τ	MSP	Rescue	Union	MSP	Rescue	Union	MSP	Rescue	Union
0.9	95.7%	57.0%	98.1%	77.1%	42.5%	86.8%	25.7%	8.5%	32.0%
0.8	89.5%	61.2%	95.9%	65.2%	46.5%	81.4%	16.6%	10.9%	25.6%
0.7	80.3%	64.1%	92.9%	53.4%	50.6%	77.0%	11.0%	12.7%	22.3%
0.6	66.3%	66.6%	88.7%	41.4%	53.7%	72.9%	7.0%	14.3%	20.3%
0.5	46.9%	70.7%	84.4%	27.4%	57.4%	69.1%	3.7%	15.7%	18.9%

D Additional Controls

D.1 Resampling Kernels and Border Painting

Figure 5 repeats the zoom sweep of Figure 2(b) with six resampling kernels. Recovery on $\mathcal{J}$ appears with every kernel. Painting a 2% border white or black, without resampling, gives 21.0% and 20.8% recovery on $\mathcal{J}$, compared with 22.2% for crop-and-resize.

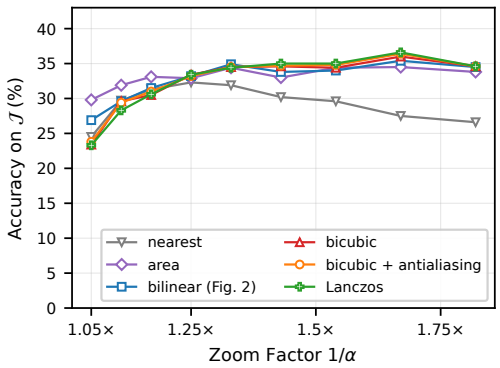

Figure 5: **Zoom recovery with different resampling kernels.** Accuracy on $\mathcal{J}$ under center zoom for six resampling kernels. Primary lens.

D.2 Matched-Damage Noise Control

To separate the effect of a transform from generic damage, we add Gaussian noise of increasing strength and record the accuracy loss it causes on $\mathcal{C}$ and the accuracy it yields on $\mathcal{J}$ (Figure 6). HF phase scramble keeps the Fourier magnitude above $r = 50$ and replaces its phase with a fixed random phase.

D.3 Frequency Grafting: Pairings and Self-Reconstruction

The transferred toxicity in Table 2 is averaged over five random pairings, each pairing every jammer with a randomly drawn Common host; the mean is 0.17% (standard deviation 0.11% across pairings).

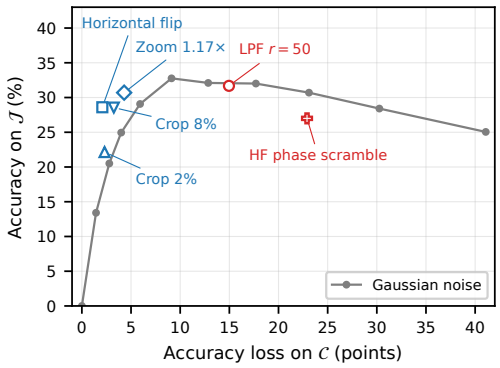

Figure 6: **Matched-damage control.** Accuracy on $\mathcal{J}$ versus accuracy loss on $\mathcal{C}$ for Gaussian noise of increasing strength (curve) and for six transforms (points). A point above the curve recovers more of $\mathcal{J}$ than noise causing the same damage on $\mathcal{C}$. Primary lens.

As a sanity check, we set the host equal to the donor and pass it through the same grafting pipeline: the jammed prediction is preserved for 2,138 of the 2,140 jammers.

D.4 Margin-Matched Boundary Control

We restrict this analysis to images whose true label ranks second under the standard model ($\mathcal{J}$: 1,416 images; $\mathcal{G}$: 3,507 images), where $\mathcal{J}$ is indeed closer to the decision boundary (median PGD-estimated margin 0.0296 vs. 0.0506). We stratify both sets by PGD-estimated margin (deciles), and additionally by margin jointly with the top-2 probability ratio p_2/p_1, and compare, among images whose prediction changes under a perturbation, the probability of moving to the true label (Table 13).

Table 13: **Margin-matched direction test.** Difference between $\mathcal{J}$ and $\mathcal{G}$ in the probability of moving to the true label among images whose prediction changed, in percentage points, before and after matching. Primary lens.

Perturbation	Raw	Margin-matched	Margin $\times$ p_2/p_1 matched
LPF ($r = 50$)	+21.3	+22.3	+22.0
Gaussian noise (same ℓ_2 norm as LPF)	+21.7	+22.0	+21.7
HF phase scramble	+18.6	+19.0	+18.8

D.5 Different Random Seeds

Table 14 compares two robust lenses of the same architecture that differ only in the training seed.

Table 14: **Same-architecture, different-seed baseline.** IoU is computed between the sets induced by the two robust lenses.

Pair	Arch	Robust acc. A / B	Prediction agreement	$\lvert\mathcal{J}\rvert$ A / B	IoU($\mathcal{J}$)	IoU($\mathcal{S}$)
Seed A vs. B (different seeds)	ResNet-18	62.32 / 62.27	76.65%	2,456 / 2,429	.492	.540
Seed A vs. B (different seeds)	ViT-B/32	66.39 / 66.62	80.78%	1,788 / 1,835	.527	.600

E Additional Datasets

For ImageNet-V2 Recht et al. [2019] we use the primary lens directly. For Oxford Pets Parkhi et al. [2012], Flowers102 Nilsback and Zisserman [2008], DTD Cimpoi et al. [2014], Food101 Bossard et al. [2014], and SUN397 Xiao et al. [2010], we freeze the standard and robust ResNet-50 backbones and train a new linear classification head on each; the head on the robust backbone is trained with ℓ_2 PGD ($\epsilon = 1$). Heads are selected on the validation split and evaluated on the test split of the splits of Zhou et al. [2022]. Table 15 reports the results.

Table 15: **Additional datasets.** LPF and Zoom columns give the standard model's accuracy on the original $\mathcal{J}$ (0% before the transform) and $\mathcal{S}$ (100% before) after LPF ($r = 50$) or a $1.17\times$ zoom. The last two columns give the size of $\mathcal{J}$ recomputed after a 2% crop and its IoU with the original $\mathcal{J}$.

				LPF		Zoom		2% crop	
Dataset	N	$\lvert\mathcal{J}\rvert$	$\lvert\mathcal{S}\rvert$	$\mathcal{J}$	$\mathcal{S}$	$\mathcal{J}$	$\mathcal{S}$	$\lvert\mathcal{J}\rvert \to \lvert\mathcal{J}_T\rvert$	IoU($\mathcal{J}$)
ImageNet-val	50,000	2,140 (4.28%)	5,002 (10.00%)	31.7%	43.8%	30.7%	73.9%	2,140 → 2,275	.465
ImageNet-V2	10,000	508 (5.08%)	1,127 (11.27%)	28.1%	40.8%	28.5%	71.9%	508 → 531	.482
Oxford Pets	3,669	158 (4.31%)	212 (5.78%)	36.1%	67.0%	36.7%	72.2%	158 → 153	.474
Flowers102	2,463	101 (4.10%)	136 (5.52%)	27.7%	61.0%	30.7%	75.7%	101 → 121	.563
DTD	1,692	140 (8.27%)	205 (12.12%)	29.3%	62.4%	27.1%	80.5%	140 → 126	.565
Food101	30,300	2,390 (7.89%)	3,877 (12.80%)	23.6%	47.6%	31.7%	70.0%	2,390 → 2,497	.559
SUN397	19,850	1,971 (9.93%)	2,227 (11.22%)	22.8%	44.1%	25.8%	66.2%	1,971 → 1,990	.518

F Scope of the Substitution Paradox

Table 16 compares the 2% crop with mild low-pass filters.

Table 16: **Substitution under mild low-pass filtering.** Sets are recomputed after each transform; IoU is with respect to the original sets ($|\mathcal{J}| = 2{,}140$, $|\mathcal{S}| = 5{,}002$). Primary lens.

Transform	$\mathcal{C}$ acc.	$\lvert\mathcal{J}_T\rvert$	$\lvert\mathcal{S}_T\rvert$	IoU($\mathcal{J}$)	IoU($\mathcal{S}$)
Crop 2% (reference)	97.7	2,275	5,171	.465	.596
LPF $r = 110$	99.4	2,183	4,947	.813	.877
LPF $r = 105$	99.0	2,222	4,920	.730	.829
LPF $r = 100$	98.5	2,295	4,849	.651	.769

G Limitations

Our main analysis is conducted on ImageNet-Val using the standard-vs-robust disagreement lens; six further datasets show the same recovery and collapse pattern (Appendix E), but results with other reference model pairs or modalities may vary. Our audit covers both CNN and ViT architectures (Table 3), but a systematic study of transformer-dominant vision models and other modalities remains an open direction.

H Broader Impacts

This paper presents work whose goal is to advance the field of machine learning, specifically in understanding the robustness of deep neural networks against natural non-robust features. By identifying "Natural Jammers," our work highlights potential vulnerabilities in current vision systems that could be relevant for safety-critical applications. There are no other specific societal consequences of our work that we feel must be highlighted here.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the three contributions (Jammed Set, spectral–spatial locking, Substitution Paradox). Their scope is stated in the paper: evidence on ImageNet-Val (Sections 3–4) and six further datasets (Appendix E), and prevalence conservation stated as approximate and limited to the tested mild transformations (Section 4.3).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix G discusses the dependence on ImageNet-Val and on the standard-vs-robust lens, the scope of the architectures studied, and other modalities; Section 4.3 limits the Substitution Paradox to the tested mild transformations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper is empirical. Definition 3.1 is an operational definition, and Proposition 4.3 is explicitly labeled empirical; both are examined experimentally rather than proven.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix A.1 lists all models, checkpoint sources, and training recipes; Appendix A.2 and Appendix A.3 define the evaluation protocols and transform operators; the additional controls and datasets are described in Appendices D–F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code to construct the quad-set partition and run the core evaluations, the membership list of the primary lens, and our robust ViT weights are released through the project page `https://jiamingzhang.netlify.app/naturaljammers`. All datasets are public; robust CNN lenses are the public MadryLab checkpoints.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits, preprocessing, optimizers, schedules, attack budgets, and transform hyperparameters are given in Appendix A.1, Appendix A.3, and Appendix E; the choice of the cutoff radius and the S^3P operating point is justified by the sweeps in Appendix A.5 and Appendix B.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Appendix C reports χ^2 tests for the main effects; the grafting result is averaged over five random pairings, with its standard deviation reported in Appendix D; variability across training seeds is reported in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A.1 (Compute Resources) states the GPU type (NVIDIA A100 80GB), the time per diagnostic evaluation, and the training cost of the robust ViT lenses.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work analyzes image classifiers on public benchmarks, involves no human subjects, and releases no sensitive data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H discusses the relevance of natural jamming for safety-critical vision systems.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work analyzes existing ImageNet classifiers on public benchmarks and poses no high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite ImageNet Russakovsky et al. [2015], ImageNet-V2 Recht et al. [2019], the five transfer datasets and their splits Parkhi et al. [2012], Nilsback and Zisserman [2008], Cimpoi et al. [2014], Bossard et al. [2014], Xiao et al. [2010], Zhou et al. [2022], torchvision pretrained models (BSD-3-Clause), and the MadryLab robust checkpoints Salman et al. [2020] (MIT license). Images are used under the datasets' research terms and are not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released code, membership list, and model weights are documented in the repository README and the model card (evaluation protocol, intended use, and license).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the methods studied. LLM assistance was limited to editing and to facilitating or running experiments, as declared at submission.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.