On the Incongruencies of Visual Explanations and Their Robustness: From Stability to Adversarial Attacks

Vincenzo Buono[1,*], Peyman Sheikholharam Mashhadi[2], Mahmoud Rahat[1], Prayag Tiwari[1] and Stefan Byttner[1]

[1]*Halmstad University*

Abstract

Visual explanations for neural networks often exhibit incongruencies, failing to capture true model reasoning due to reconstruction errors and insensitivity to latent attention shifts. We explore these failure modes by dissecting how polysemanticity and superposition give rise to circuitry-level redundancies, allowing models to maintain output stability while internal computational pathways—quantifiable by metrics like Relative Attention Shift (RAS) and Feature Dispersion—diverge significantly under perturbation. This work offers a mechanistic understanding of explanation reliability, its connection to adversarial vulnerability, and the challenges inherent in interpreting robust, redundant systems.

Keywords

eXplainable AI, Explanation Robustness, Explanations artefacts, Saliency Upsampling

1. Introduction

Neural networks have become the de–facto foundation of contemporary computer vision, yet our understanding of *why* they succeed remains largely mediated by post–hoc visual explanations—saliency maps, class–activation visualisations, integrated gradients, and countless variants. These techniques promise an interpretable window into opaque mechanisms; in practice, they often provide *mutually contradictory narratives* that shift under minimal perturbations. Saliency can reverse when a single pixel is toggled, attribution heatmaps migrate when the input is resized, and discriminative regions vanish when a different baseline is chosen. Such *incongruencies* undermine scientific and safety arguments for explanation–driven auditing, yet their root causes remain poorly understood.

In this work we consolidate and extend three strands of ongoing research to clarify *when* and *why* visual explanations diverge from the classifier's true decision–making circuitry:

1. **Reconstruction–based incongruency.** Many attribution failures can be anticipated from the *reconstruction error* incurred when an explanation is treated as an information bottleneck. If masking by a saliency map cannot faithfully reconstruct the original signal, the explanation should be mistrusted. We formalize this intuition and derive tight upper bounds linking reconstruction error to attribution mismatch.

SAIS2025: Swedish AI Society Workshop 2025, 16-17 June 2025, Halmstad, Sweden.

*Corresponding author.

†First author.

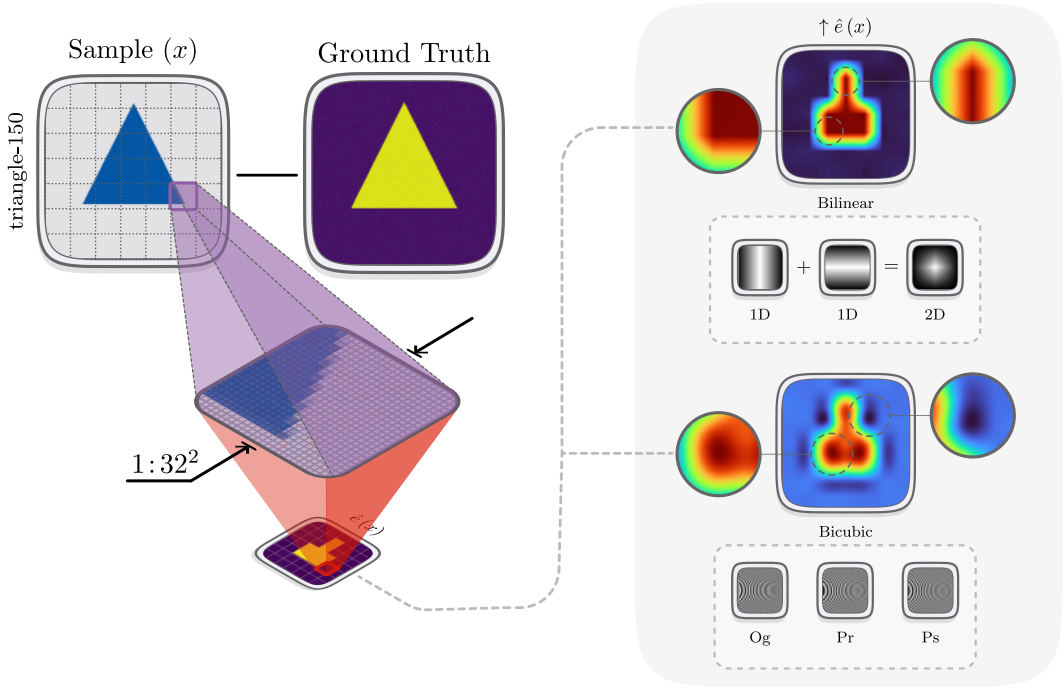

Figure 1: Upsampling corrupts saliency maps through reconstruction errors. Classical saliency upsampling kernels can introduce reconstruction errors, leading to incorrect visual explanations.

2. **Perturbation–induced unintuitiveness.** We revisit occlusion, masking, and patch–deletion probes—the workhorse tools of feature importance—through a local Lipschitz analysis. Counter–examples reveal that small, spatially coherent perturbations can trigger *latent attention shifts* that leave model confidence unaltered while drastically re–wiring internal activation paths. Classical occlusion scores therefore conflate sensitivity with representational redundancy and routinely mis–rank critical regions.
3. **Robustness as circuitry redundancy.** Building on recent mechanistic–interpretability insights, we propose a unified framework in which stability emerges from *redundant conceptual detectors* distributed across superposed latent subspaces. We prove a redundancy–implies–robustness theorem, quantify redundancy via a dispersion metric, and connect these theoretical results to adversarial attacks that exploit hidden fail–over circuits.

This paper serves as a *condensed exposition of salient findings* from **three ongoing research trajectories**; it is therefore intentionally circumscribed in scope, eschewing comprehensive coverage or an in-depth exploration of the granular technical details inherent to each individual project. Collectively, these contributions advance a principled account of explanation reliability: we move from a surface–level taxonomy of saliency pathologies to a quantitative, circuit–level diagnosis that predicts when explanations fail, elucidating why naïve perturbations mislead, and situates robustness within the geometry of superposed latent detectors.

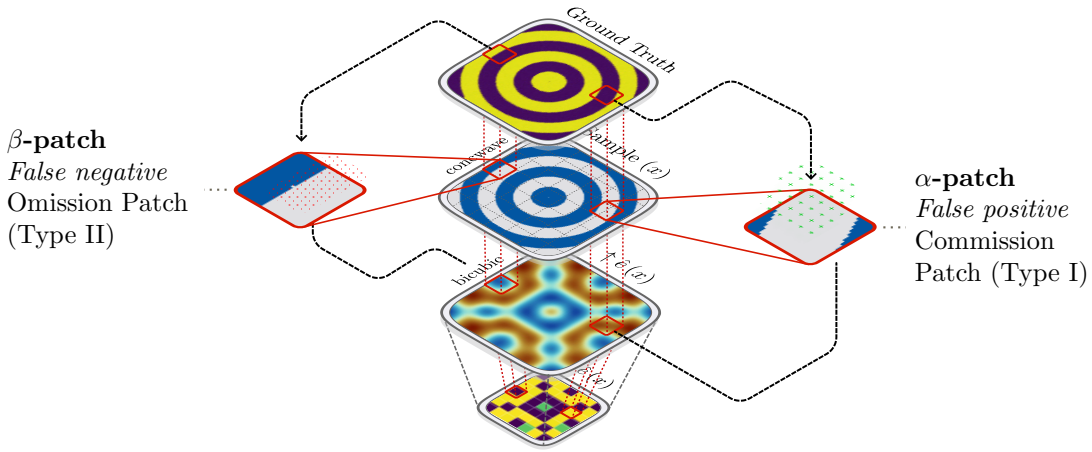

Figure 2: Compression error types. From top to bottom: ground-truth attribution, input image, upsampled saliency map, and raw low-resolution heatmap. Left column shows β-errors where true attributions disappear due to compression (signal loss). Right column shows α-errors where spurious attributions emerge despite absence in ground truth (false signal generation). Both error types corrupt explanation fidelity.

Contributions. Our work offers three primary advances:

- **Theory.** We derive reconstruction–error bounds that formalize incongruency, provide a Lipschitz characterisation of occlusion–induced attention shifts, and prove a redundancy–robustness theorem.
- **Metrics.** We introduce *Relative Attention Shift* (RAS) and *Feature Dispersion* to quantify latent re–configurations and the spread of conceptual load.
- **Empirical validation.** Across ImageNet–scale classifiers and modern vision transformers we demonstrate that high reconstruction error, high RAS, and high dispersion jointly predict explanation failure modes and adversarial vulnerability with state–of–the–art accuracy.

2. Reconstruction–Based Incongruency

Visual explanations, particularly saliency maps, are often conceptualized as filters that isolate the critical subset of input information purportedly used by a neural network. The fidelity of such an explanation can be gauged by treating it as an information bottleneck: if the model's decision-making process, or its eventual outcome, cannot be adequately reconstructed from the information that passes through this bottleneck, the explanation itself is likely incongruent with the model's true internal mechanisms. This *reconstruction error* thus serves as a direct, quantifiable marker for the unreliability of an explanation, signalling a misalignment between what the explanation highlights and what the model computationally prioritizes.

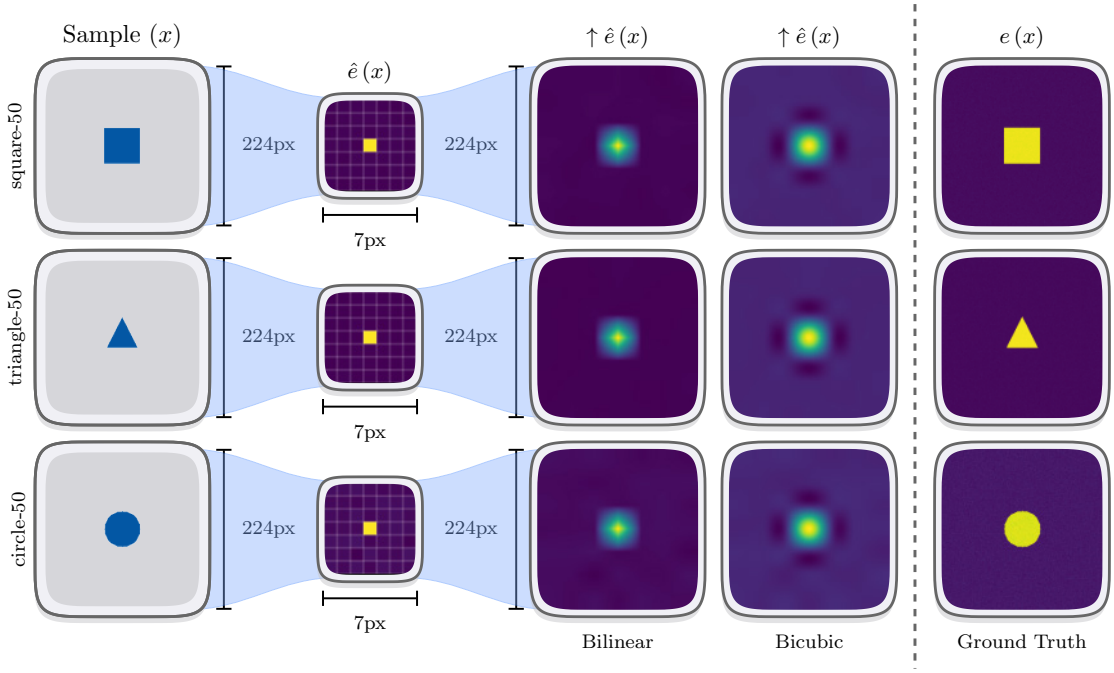

Figure 3: Identical saliency for distinct inputs. Each row shows (left) the input, (middle) the raw heatmap, (right) the upsampled saliency map produced by a canonical attribution method, and (rightmost) the ground–truth attribution priors. Despite distinct ground truths, the visual explanation converges to a single pattern, illustrating reconstruction–based incongruency.

Such incongruencies are not mere abstract concerns but represent a pervasive challenge within the field of interpretability. Commonly, attribution methods exhibit problematic behaviors, such as generating (i) nearly identical saliency maps for visually distinct inputs with differing ground-truth features or class labels (as illustrated in Figure 3), or conversely, (ii) producing disparate explanations for a single input when different explanation techniques are utilized, each potentially optimizing for a distinct facet of eXplainable AI (xAI) fidelity metric. The origins of these inconsistencies are manifold. Qualitative assessments often suffer from inherent subjectivity, while quantitative evaluations frequently yield indeterminate or ambiguous results that extend beyond simple discrepancies in rank-ordering. Although faithfulness remains a universally sought-after characteristic of explanations, existing metrics designed to assess this property often capture only a limited aspect of this multifaceted concept, sometimes even emphasizing undesirable characteristics. While acknowledging this broader context, the present work focuses on incongruencies stemming from reconstruction errors, particularly those introduced by common post-processing steps such as the upsampling of low-resolution heat maps (Figure 1). In these cases, the explanation converges to a generic, input-agnostic pattern rather than faithfully reflecting input-specific discriminative features, thereby failing to reconstruct the unique *reasoning trace* for each input and offering merely a coarse, often uninformative, approximation of the model's attention (Figure 2).

Furthermore, incongruencies can be inadvertently introduced or amplified by the technical

procedures inherent in generating and processing explanation maps. Many contemporary saliency methods, for example, produce low-resolution heatmaps that necessitate upsampling to align with the original input dimensions. This upsampling step, often involving standard interpolation kernels or filtering techniques, can introduce spatial distortions, blur fine-grained details, or erroneously diffuse attributed importance. The result is that the processed explanation map itself becomes a flawed reconstruction, not necessarily of the model's raw internal saliency, but of what a high-fidelity, veridical explanation ought to represent. These artefacts can lead to fundamentally incorrect interpretations regarding the precise localization and significance of features.

Our work aims to move beyond these qualitative observations by establishing a more formal, quantitative understanding of reconstruction-based incongruency. Let $\mathbf{x} \in \mathcal{X}$ be an input processed by a model $f : \mathcal{X} \to \mathcal{Y}$ to produce an output $f(\mathbf{x}) \in \mathcal{Y}$. An explanation for the decision $f(\mathbf{x})$, denoted $\hat{\mathbf{e}}$, is typically generated by a local feature attribution method Φ (i.e., $\hat{\mathbf{e}} = \Phi(\mathbf{x}, f, f(\mathbf{x}); \lambda)$). We contend that the trustworthiness of this explanation $\hat{\mathbf{e}}$ is directly proportional to its ability to enable the faithful reconstruction of $f(\mathbf{x})$ when the model is applied to an input $\mathbf{x}_{\hat{\mathbf{e}}}$ that is conditioned or filtered by $\hat{\mathbf{e}}$ (e.g., if $\hat{\mathbf{e}}$ is a feature relevancy map, $\mathbf{x}_{\hat{\mathbf{e}}}$ might be formed by $\mathbf{x} \odot \hat{\mathbf{e}}$, where $\odot$ denotes element-wise multiplication). More profoundly, this trustworthiness also pertains to the capacity to reconstruct the key internal model representations, denoted $\phi(\mathbf{x})$, and the specific computational pathways within f that culminate in $f(\mathbf{x})$. Consequently, a significant divergence between the decision $f(\mathbf{x}_{\hat{\mathbf{e}}})$ derived from the explanation-conditioned input and the original decision $f(\mathbf{x})$—or a substantial mismatch in their underlying causal features, as might be reflected by comparing $\phi(\mathbf{x}_{\hat{\mathbf{e}}})$ with $\phi(\mathbf{x})$—signifies a high reconstruction error attributable to the explanation $\hat{\mathbf{e}}$. Building upon this premise, we develop a theoretical framework to derive quantitative bounds linking this reconstruction error to the likelihood of attributional mismatch, thereby offering a principled methodology for identifying and potentially rectifying explanations that function as deficient or misleading information bottlenecks.

Our prior research introduced Universal Semantic-Aware Upsampling (USU), a semantically-aware black-box algorithm specifically engineered for eXplainable AI applications. The USU framework was designed to mitigate two critical issues: the information bottlenecks often present in explanation generation, and the pervasive distortions and artefacts attributable to conventional upsampling kernels prevalent in the interpretability methods. In the current work, however, our focus is deliberately constrained. We aim to examine the aforementioned pitfalls and systemic fragilities characteristic of widely adopted interpretability instruments, rather than to provide a detailed exploration of the USU methodology itself.

In what follows, we embark upon an *exploration* of what it means for explanations to exhibit robustness, and how such an elusive property might be *quantified and assessed*. We investigate how *classical interpretability assumptions*—which equate robustness with simple sensitivity measures—fail to capture the *intricate internal dynamics* of modern, large-scale models. These models, as we demonstrate, achieve their remarkable stability not through rigid invariance, but through *circuit redundancy*: multiple, functionally equivalent pathways that enable flexible adaptation while preserving outputs. By bridging recent *mechanistic interpretability* insights with our proposed metrics, we reveal how apparent instability in explanations may paradoxically signal the presence of *robust, multiply-realizable* solutions.

3. Appendices

A. Deconstructing the Black Box: Interpretability in the Age of Complex Representations

Neural networks often exhibit *complex internal representations*, blending multiple features into shared latent dimensions. Their internal representations are *rarely straightforward, one-to-one mappings.* Instead, they learn *dense, distributed codes* where multiple concepts can be *compressed* into shared latent dimensions or even *single neuronal pathways.* This appendix explores, *in a summarized fashion*, the implications of such *representational complexity* for our ability to interpret and understand these models. We explore how phenomena like *polysemanticity* and *superposition* challenge traditional explanation techniques and necessitate new tools for *peering into the latent machinery* of deep learning.

A.1. The Entangled Web: Polysemanticity and Superposition

At the heart of the interpretability challenge lie two key representational strategies commonly adopted by neural networks:

- **Polysemanticity**: This refers to the phenomenon where a single internal unit of the network—be it a neuron or a direction in activation space—encodes multiple, often unrelated, conceptual features. It's as if a single word in the network's internal language carries several distinct meanings, decipherable only in specific contexts.
- **Superposition**: Here, different concepts are not neatly segregated but instead overlap, sharing portions of the model's internal representational space. The network effectively learns to store a richer palette of features than its raw number of dimensions might suggest, forcing these features to become entangled.

This *intertwined encoding* poses a *fundamental problem*: if concepts are not *cleanly separated*, how can we *reliably determine* which specific aspects of the model's architecture or learned parameters are responsible for recognizing or processing a particular feature? This ambiguity is a central hurdle for *interpretability.*

A.2. Conventional Methods and Their Limitations

Many *standard interpretability methods*, such as *occlusion-based sensitivity analyses*, attempt to gauge the "importance" of an input region by observing the impact its removal or perturbation has on the model's *final output.* For instance, if obscuring a patch of an image significantly degrades the model's classification confidence for a particular object, that patch is deemed *salient* for that object.

However, a *critical limitation* arises: these methods typically focus on the conditional *terminal output.* If perturbing an input region *doesn't* alter the final prediction, it might be dismissed as unimportant. But what if the model, faced with a slightly degraded input, *subtly reshuffles its internal computational strategy*, re-allocating its "attentional resources" or leveraging *redundant*

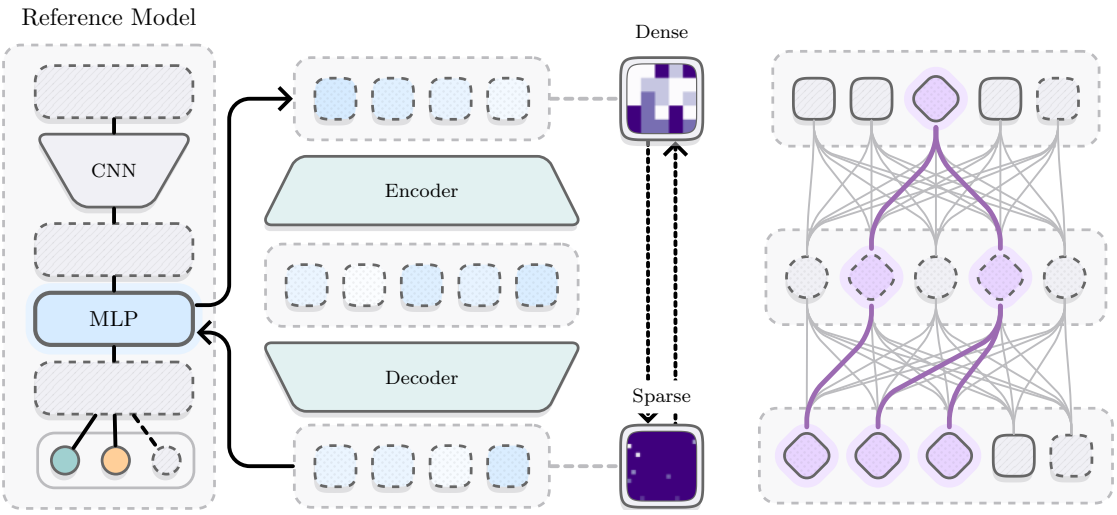

Figure 4: Conceptual depiction of circuit-level robustness and latent attentional shifts. Perturbations might appear to have minimal impact on the final output. However, internally, the model may re-route its processing through alternative, functionally equivalent circuits or feature encodings. Relative Attention Shift (RAS) aims to quantify such internal reconfigurations that are not apparent from output changes alone.

pathways to arrive at the same conclusion? Such *internal gymnastics* often go unnoticed by *modern, contemporary explanation techniques.*

A.3. How Many Ways Does a Network Know? Counting Solutions with RAS

The Relative Attention Shift (RAS) metric is engineered to penetrate beyond the veil of stable outputs, offering a lens into the internal adaptability of a neural network. Its core function is not merely to track feature importance, but to quantify the *multiplicity of computational pathways*—or distinct internal "solutions"—that a model has learned to solve a given problem while holding its final prediction constant. When an input is perturbed, if the model must significantly reconfigure its internal circuitry or attentional focus to preserve the original output, RAS registers this internal effort. A high RAS, therefore, suggests a rich internal landscape where the model can switch between alternative strategies, drawing upon different combinations of features or internal computations to achieve the same end. This capacity for substantial internal rearrangement in the face of input variation, without altering the outcome, is a hallmark of a system that possesses a *diverse array* of solutions.

The substrate for such a *diverse array* is provided by the very nature of learned representations in deep networks—specifically, *polysemanticity* and *superposition.* As discussed previously, these phenomena mean that features are not isolated but are encoded in dense, overlapping, and distributed ways. This entanglement provides the raw material for multiple, functionally equivalent circuits to emerge, as conceptually illustrated in Figure 4. When one cue or pathway is disrupted by a perturbation, the network can seamlessly—though with significant internal reconfiguration detected by RAS—pivot to an alternative.

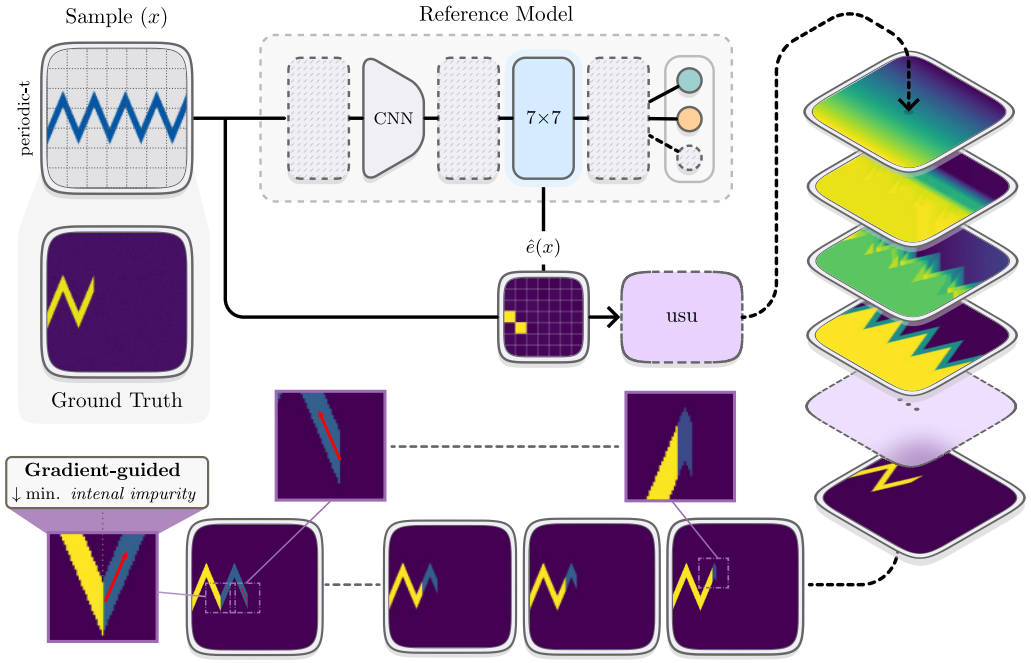

Figure 5: Illustrative example of gradient-guided impurity segmentation, a technique employed by USU (Semantic-Aware Upsampling) in post-processing visual explanations.

A.4. Rethinking Robustness: Unstable Models are Robust Models

The insights gleaned from RAS fundamentally challenge *prevailing notions of model and explanation robustness*, particularly as these notions are *applied and measured* within the XAI field. The *de facto standard* for assessing such robustness—both of the model itself and of the explanations it engenders—often hinges on demonstrating *stability under infinitesimal perturbations*. Consequently, many contemporary stability metrics, paradigmatically are designed to penalize any significant internal representational divergence following minor input perturbations. These metrics presuppose that reliable models and faithful explanations should exhibit minimal internal change; thus, substantial internal shifts are typically interpreted as a sign of instability or unreliability. RAS, conversely, operates from a diametrically opposing premise: it values the capacity for *extensive and diverse internal reconfigurations* (i.e., a maximized number and varied types of internal solutions, as might be further characterized by dispersion) precisely when output stability is maintained. A model exhibiting a high RAS score—indicative of substantial internal adaptation to preserve its prediction—would, by the criteria of these standard stability-focused metrics, be paradoxically deemed *representationally unstable* and its explanations potentially unreliable.

This leads to a provocative reinterpretation: a model that is robust in the sense of possessing a *diverse array* of solutions (high redundancy) might be classified as "unstable" by metrics prioritizing internal representational invariance. This directly confronts current paradigms in

xAI that advocate for enhancing robustness by training or fine-tuning models to specifically *ignore* or down-weight so-called "non-robust" features. These are often characterized as features with high predictive utility but which are easily "flipped" or manipulated, potentially leading to adversarial vulnerabilities or reliance on spurious correlations (shortcuts). The conventional approach implies that a robust model is one that has learned a parsimonious, invariant mapping, ideally relying only on a core set of causally robust features.

However, this deliberate feature suppression is counterintuitive from an optimization perspective. Neural networks are designed to exploit any and all predictive signals present in the training data to maximize performance. To compel a model to discard learned, useful shortcuts or highly predictive (albeit potentially brittle) features seems to artificially limit its capabilities. The framework suggested by RAS, and the underlying representational complexity it probes, offers an alternative perspective on this dilemma. The critical vulnerability may not lie in the *utilization* of individually "non-robust" features, but rather in an *exclusive reliance* upon them due to a paucity of alternative computational pathways (i.e., low redundancy, which would manifest as a low RAS). A model with high representational redundancy, evidenced by a high RAS, can afford to leverage a diverse array of features, including those that might be individually fragile, because it is not singularly dependent on any one of them. Its true resilience stems from this adaptive capacity and the richness of its learned solution manifold, rather than from a constrained adherence to a pre-defined set of "robust" features.

A.5. Beyond Magnitude: Dispersion in Representational Space

While RAS primarily provides a measure of the *multiplicity of internal solutions* or computational pathways a network can leverage (effectively, how many ways it has learned to solve a problem while maintaining output stability), a complementary metric, termed Feature Dispersion, helps characterize the *distribution* or *topography* of the internal reconfigurations associated with switching between these solutions within the model's high-dimensional representation space.

- **High Dispersion** suggests that the internal representational adjustments are diffuse, spread broadly across many latent dimensions or directions.
- **Low Dispersion** indicates that the changes are more concentrated, primarily occurring along a few specific representational axes.

Considered together, RAS and dispersion provide a richer understanding of the model's internal response. We can ascertain not only that the model reconfigured its internal feature reliance (high RAS) but also whether this reconfiguration involved a localized adjustment or a more globally distributed restructuring of its latent activations.

A.6. Robustness from Redundancy: A Double-Edged Sword

The phenomena of polysemanticity and superposition, and the distributed internal dynamics they enable (as highlighted by RAS and dispersion), naturally lead to a form of conceptual redundancy. The model, in essence, learns multiple "detectors" or internal circuits for the same concept. If one pathway or encoding is perturbed or becomes unreliable, the model can often fall back on alternative, superposed encodings.

This inherent redundancy is a cornerstone of the robustness often observed in deep neural networks. It makes them resilient to minor input variations and certain types of adversarial attacks. However, this same redundancy presents a formidable challenge for targeted interventions. Trying to "remove" or "unlearn" a specific concept from a trained model becomes akin to trying to remove a particular pigment from a deeply blended paint—targeting one instance of the concept may be insufficient if it is encoded through numerous, entangled pathways. The model's inherent ability to reconstruct or reroute conceptual processing can frustrate such post-hoc modification attempts.

A.7. The Challenge of Post-Hoc Concept Ablation

Consider the task of forcing a pre-trained model to cease using a particular concept—for example, attempting to de-identify a model by removing its ability to recognize a specific individual. Given the prevalence of polysemanticity and superposition, such post-hoc concept ablation is extraordinarily difficult. An intervention targeting one identified neural correlate of the concept might be ineffective, as the model can leverage its distributed and redundant representational structure to reroute processing through alternative latent pathways, effectively preserving or reconstituting the targeted concept. It's analogous to trying to stop water flowing through a porous sponge by blocking a single channel; the water simply finds other routes.

A.8. Concluding Perspectives: Navigating the Entangled Landscape

In essence, neural networks achieve their remarkable capabilities in part by learning to compress a vast number of features and concepts into their internal latent spaces in highly efficient, albeit entangled, ways. Polysemanticity and superposition are hallmarks of this compression. Metrics like Relative Attention Shift (RAS) and Feature Dispersion provide crucial tools for detecting and characterizing the complex internal reconfigurations that models undertake to maintain stable outputs in the face of perturbations, revealing a landscape of multiple, latent computational solutions. This inherent representational redundancy is a primary source of their robustness. Simultaneously, it underscores the profound difficulty in achieving fine-grained interpretability and reliable post-hoc control over unwanted learned behaviors. Acknowledging and further investigating this intertwined relationship between representational flexibility, redundancy, robustness, and the limits of interpretability is central to advancing our capacity to build more transparent, controllable, and ultimately trustworthy AI systems.

The practical challenges of interpreting these models are also compounded by the very tools we use. For instance, visual explanations often require post-processing, such as upsampling low-resolution heatmaps. Standard interpolation techniques used for this can introduce their own distortions and artifacts, further obscuring the model's true decision-making process, as illustrated by the comparative examples in Figure 6. Similarly, other refinement techniques, like the gradient-guided impurity segmentation shown in Figure 5, while aiming to improve clarity, operate on explanations that are already products of a complex, entangled system.

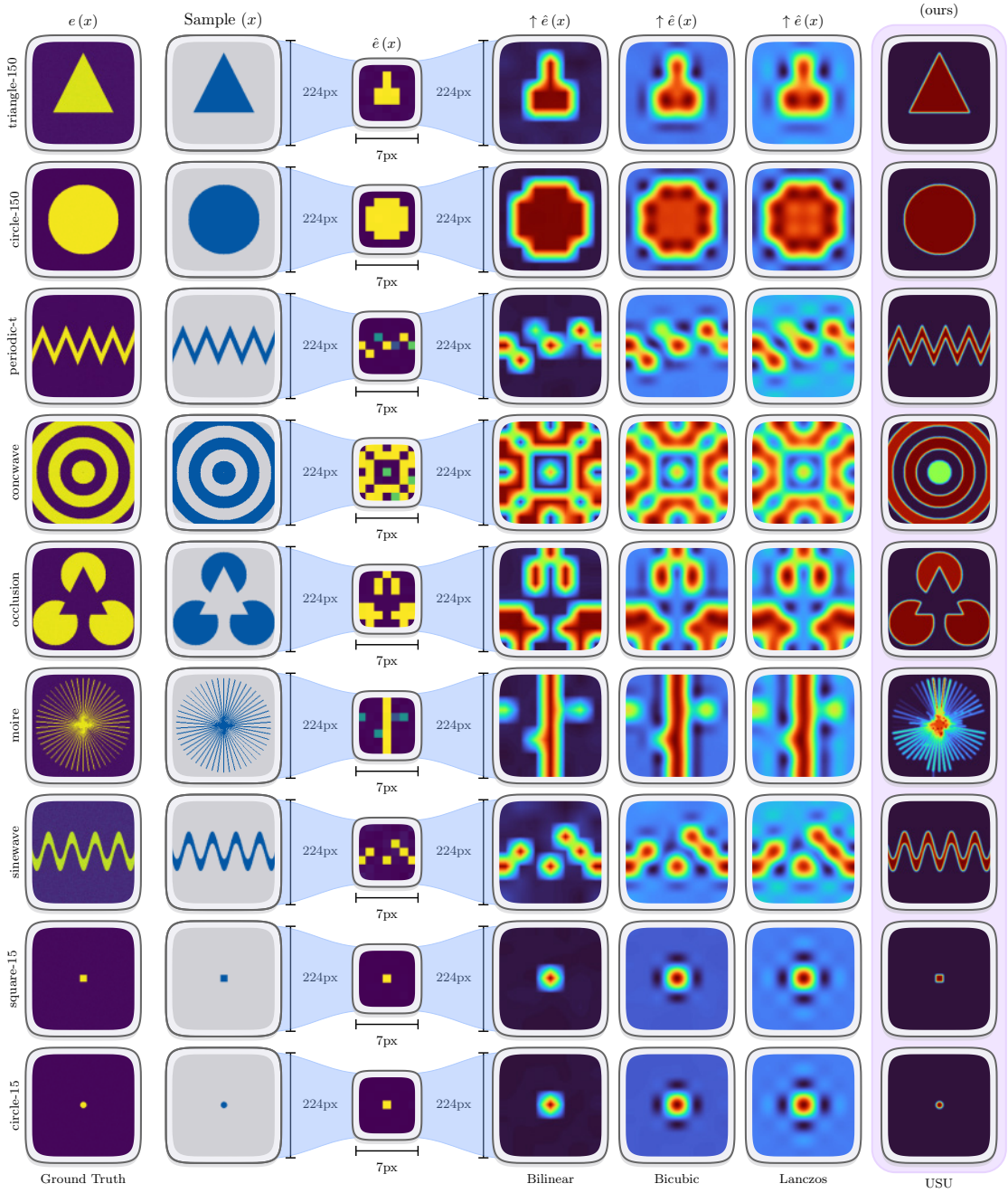

Figure 6: Extended visual comparison highlighting common artifacts (e.g., blurring, spatial misattribution) introduced by standard upsampling algorithms when applied to low-resolution explanation maps in XAI. These contrast with the qualitative improvements achievable via semantically-aware upsampling technique (USU), which aim to preserve critical details from the model's internal representations.