A REPRODUCIBLE FRAMEWORK FOR EXPLAINABILITY AND DISTRIBUTIONAL ROBUSTNESS EVALUATION IN VISION MODELS

Nidhi Mithiya

Department of Computer Science and Engineering, IIT (ISM) Dhanbad

ABSTRACT

Understanding whether deep vision models rely on causal visual cues or contextual shortcuts is critical for trustworthy AI. This study introduces a reproducible framework that combines explainability analysis with counterfactual testing to evaluate both spatial focus and distributional sensitivity in convolutional and transformer-based architectures. Using the Oxford-IIIT Pet dataset, we generate counterfactual samples by compositing segmented object foregrounds onto 15 natural backgrounds and analyze attribution patterns using Grad-CAM, Saliency Maps, Integrated Gradients, and Attention Rollout. Quantitative metrics—Foreground Attention Ratio (FAR), Background Dependence Index (BDI), Grad-CAM–Mask IoU, and ΔAccuracy—are jointly used to compare visual interpretability and behavioral robustness. Experiments on three state-of-the-art models pretrained on ImageNet (ResNet-50, Swin-Tiny, and DeiT-Small) show that while FAR and BDI remain stable across counterfactuals, models exhibit notable accuracy drops under background shifts, revealing distributional sensitivity rather than explicit spurious correlation. The proposed framework thus establishes a baseline for reproducible explainability evaluation and highlights the gap between visual attribution stability and true causal robustness.

***Index Terms*—** Explainability, Counterfactual Evaluation, Distributional Sensitivity, Vision Transformers, CNNs, Robustness

1. INTRODUCTION

Deep vision models achieve impressive accuracy yet often depend on unintended contextual cues rather than causal object features [9, 7, 8]. These spurious dependencies undermine reliability, particularly when models encounter unseen environments [16]. Explainable AI (XAI) methods such as Grad-CAM and Integrated Gradients provide insights into where models attend, but visual attribution alone does not guarantee causal understanding or robustness.

This study introduces a unified and reproducible framework that integrates explainability-based visualization with counterfactual evaluation to probe whether model attention and prediction stability align under contextual perturbations.

Fig. 1. Overview of the proposed evaluation framework integrating counterfactual generation, explainability visualization, and quantitative robustness metrics.

By combining attribution metrics (FAR, BDI, IoU) with behavioral measures (ΔAccuracy), we investigate not only how models *see*, but how consistently they *behave* when visual context changes. Our analysis spans both convolutional (ResNet-50) and transformer-based (Swin-Tiny, DeiT-Small) models, offering architectural insights into the distinction between spatial interpretability and distributional sensitivity.

2. RELATED WORK

Explainability and Attribution Methods. Gradient-based visual explanations such as Grad-CAM [1], Saliency Maps [2], and Integrated Gradients [3] are foundational for model interpretability. These methods highlight image regions contributing most to predictions but differ in granularity and stability. Prior studies have shown that CNNs produce spatially localized explanations, while Vision Transformers exhibit globally distributed attention [4, 5].

Spurious Correlations and Counterfactuals. Models trained on biased datasets may rely on contextual cues—a phenomenon termed "Clever Hans behavior" [7]. Counterfactual evaluation provides a way to diagnose such bias by modifying non-causal attributes while preserving the causal signal [6]. In vision tasks, background replacement has been used to test sensitivity to non-object features [8]. However, few works have systematically linked explainability metrics to counterfactual robustness, which is the focus of this framework.

Architectural Sensitivity. CNNs tend to rely on local texture cues [9], whereas transformers leverage global self-attention mechanisms, improving generalization but not fully eliminating contextual bias [4]. Comparative analysis of ex-

plainability methods across these architectures remains limited—this study aims to close that reproducibility gap.

3. METHODOLOGY

3.1. Overview

The framework integrates explainability-based analysis with counterfactual testing to study whether model attributions correspond to causal object regions and whether predictions remain stable under contextual changes. Figure 1 outlines the experimental pipeline—dataset preprocessing, counterfactual generation, model training, attribution map extraction, and quantitative metric computation.

3.2. Dataset and Counterfactual Generation

We use the Oxford-IIIT Pet dataset [10], containing 37 pet categories with pixel-level trimaps distinguishing object, background, and uncertain regions. Counterfactuals are generated by compositing object foregrounds over 15 natural backgrounds sampled from publicly available landscape and texture datasets. This approach follows prior counterfactual construction strategies [6] to isolate background influence while maintaining semantic integrity. However, simple compositing can lead to unrealistic shadows or color mismatches, limiting its capacity to simulate true distributional shifts.

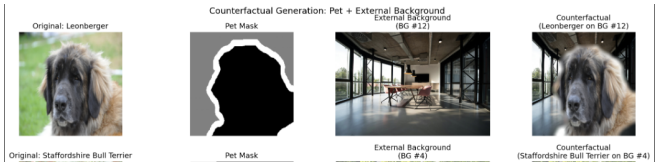

Fig. 2. Counterfactual generation pipeline: foreground objects are extracted from trimaps and composited over 15 natural backgrounds to probe contextual effects.

3.3. Model Selection and Training

Three models were evaluated—ResNet-50 (CNN), Swin-Tiny (hierarchical Transformer), and DeiT-Small (ViT-based Transformer). All were fine-tuned for 30 epochs using pretrained ImageNet weights, AdamW optimization ($lr = 3 \times 10^{-4}$, weight decay $= 0.01$), and cosine learning rate scheduling. AugMix-based augmentations [11] improved robustness to natural variations. Validation included 736 original samples and corresponding counterfactuals.

3.4. Explainability Methods

Explainability techniques were applied to analyze visual focus:

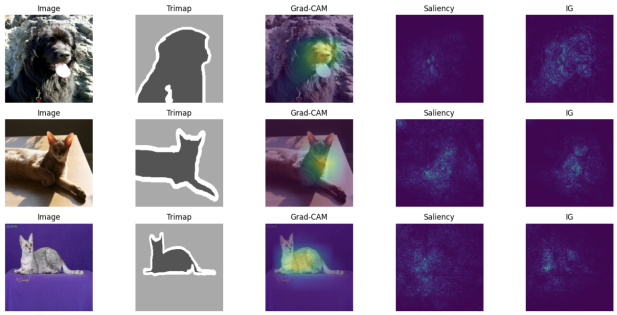

Fig. 3. Attribution visualizations using Grad-CAM, Saliency, and Integrated Gradients across models. CNNs show localized focus, while Transformers exhibit global patterns.

- **Grad-CAM** [1]: Highlights spatial attention in CNNs via feature map gradients.
- **Saliency Maps** [2]: Compute pixel-wise sensitivity to input perturbations.
- **Integrated Gradients (IG)** [3]: Integrates gradients along an input path for smoother attributions.
- **Attention Rollout** [12]: Aggregates attention weights across transformer layers to obtain global spatial relevance.

Grad-CAM was used for CNNs, while DeiT-Small relied on Attention Rollout due to its non-convolutional architecture.

3.5. Quantitative Metrics

Four complementary metrics were used to link interpretability and robustness:

1. **Foreground Attention Ratio (FAR)** – Fraction of attention inside the ground-truth mask [5].
2. **Background Dependence Index (BDI)** – Computed as $1 - \text{FAR}$, indicating contextual reliance [6].
3. **Grad-CAM–Mask IoU** – Overlap between Grad-CAM maps and segmentation masks [1].
4. **ΔAccuracy** – Drop in classification accuracy between original and counterfactual samples [4].

FAR, BDI, and IoU measure visual consistency, whereas ΔAccuracy reflects behavioral sensitivity under context changes.

Table 1. Unified explainability and counterfactual evaluation results. Baseline FAR values are derived from attribution maps; counterfactual FAR/BDI/IoU quantify contextual robustness.

Model	Method	Baseline FAR	Counterfactual FAR	BDI	IoU
ResNet-50	Grad-CAM	0.6716 ± 0.1803	0.8788 ± 0.0709	0.1212 ± 0.0709	0.3000 ± 0.0000
ResNet-50	Saliency	0.5195 ± 0.1622	–	–	–
ResNet-50	Integrated Gradients	0.5553 ± 0.1616	–	–	–
Swin-Tiny	Grad-CAM	0.3103 ± 0.1566	0.8861 ± 0.0408	0.1139 ± 0.0408	0.3000 ± 0.0001
Swin-Tiny	Saliency	0.4653 ± 0.1743	–	–	–
Swin-Tiny	Integrated Gradients	0.4646 ± 0.1611	–	–	–
DeiT-Small	Saliency	0.5515 ± 0.1500	–	–	–
DeiT-Small	Integrated Gradients	0.4910 ± 0.1401	–	–	–
DeiT-Small	Attention Rollout	0.4677 ± 0.1578	–	–	–

4. EXPERIMENTS AND RESULTS

4.1. Quantitative Results

FAR and BDI remained largely consistent across counterfactual samples, indicating that the models' spatial attention did not shift significantly under background perturbation. However, the observed ΔAccuracy across models ranged from 10–20%, revealing substantial sensitivity to contextual changes despite stable attribution maps. This demonstrates that spatial interpretability does not directly imply causal robustness—models may visually "focus" correctly while still encoding distribution-specific dependencies.

4.2. Interpretability Insights

Grad-CAM provided consistent, localized activations for CNNs, while transformers exhibited globally distributed focus patterns. Saliency and IG maps showed moderate variability due to gradient sensitivity. Attention Rollout effectively captured transformer reasoning but often produced diffuse attention. These findings align with prior work [5, 4], reaffirming that architectural inductive biases shape explainability outcomes.

4.3. Framework Limitations

The main limitation stems from the compositing-based counterfactual generation process, which may introduce unrealistic lighting or texture boundaries. This limits the strength of causal conclusions and may inflate perceived sensitivity. Future extensions using diffusion-based harmonization [13] could yield more reliable counterfactuals and isolate causal factors more effectively.

5. CONCLUSION

This work presented a reproducible framework for evaluating explainability and robustness in vision models under contextual perturbations. While no strong evidence of spatial spurious correlation was found (FAR and BDI remained stable), significant accuracy degradation under counterfactuals indicates distributional sensitivity to unseen backgrounds. The framework reveals a key gap between visual interpretability and predictive stability—highlighting that attribution maps alone cannot capture causal reliability.

Future directions include improving counterfactual realism using diffusion-based generation [13], and incorporating structured debiasing strategies such as **Background Randomization (BR)** [14] and **Contextual Bias Fine-tuning (CBF)** [15] to enhance causal generalization.

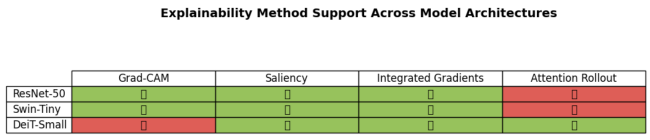

Fig. 4. Applicability of explainability methods across architectures. Green = applicable; Red = incompatible. Grad-CAM suits CNNs; IG and Rollout generalize to Transformers.

6. REFERENCES

[1] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *ICCV*, 2017.

[2] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv:1312.6034*, 2013.

[3] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *ICML*, 2017.

[4] H. Wang, M. Raghu, and K. Zhong, "Do Vision Transformers See Like Convolutional Neural Networks?," in *CVPR*, 2023.

[5] M. Moayeri, H. Qin, and Y. Sun, "A Comprehensive Study of Image Classification Model Sensitivity to Foregrounds and Backgrounds," in *CVPR*, 2022.

[6] Z. Li, "Investigating Spurious Correlations in Vision Models Using Counterfactual Images," in *CVPR Workshops*, 2025.

[7] W. Ye *et al.*, "The Clever Hans Mirage: A Comprehensive Survey on Spurious Correlations in Machine Learning," *arXiv preprint arXiv:2402.12715*, 2024.

[8] S. Beery, G. Van Horn, and P. Perona, "Recognition in the Wild: A Domain Generalization Benchmark for Wildlife Species Classification," in *CVPR*, 2018.

[9] R. Geirhos *et al.*, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in *ICLR*, 2019.

[10] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and Dogs," in *CVPR*, 2012.

[11] D. Hendrycks *et al.*, "AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty," in *ICLR*, 2020.

[12] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," in *ACL*, 2020.

[13] C. Chang *et al.*, "Counterfactual Generation and Evaluation for Visual Causal Analysis," in *ICCV*, 2023.

[14] Q. Xie, M. Tan, and Q. V. Le, "Delving into Deep Data Augmentation for Robust Visual Recognition," in *ICLR*, 2020.

[15] S. Santurkar, A. Iwata, and A. Madry, "Whose Bias Is It Anyway? Mitigating Bias by Improving Model–Dataset Alignment," in *NeurIPS*, 2023.

[16] S. Singla and A. Feizi, "Understanding and Mitigating Spurious Correlations in Deep Learning," in *NeurIPS Workshop*, 2021.