IASc-INSA-NASI Summer Research Fellowship Program - Final Report

Robustness and Interpretability of Vision Foundation Models (VFMs): A Comparative Survey Under Distributional Shifts

Submitted by:

Suha Roy (SRFP ENGS4751)

Department of Data Science and Analytics

Central University of Rajasthan

Under the Supervision of:

Prof. Sanghamitra Bandyopadhyay

Professor, Machine Intelligence Unit

Director, Indian Statistical Institute, Kolkata

Jun - Aug 2025

Certificate from the Supervisor

This is to certify that **Suha Roy (ENGS4751)** has successfully completed eight weeks of the IASc-INSA-NASI Summer Research Fellowship Program 2025 under my supervision in the Bioinformatics Lab of the **Machine Intelligence Unit** at the **Indian Statistical Institute, Kolkata** from June 12 to August 11.

The research work conducted during this period is titled:

"Robustness and Interpretability of Vision Foundation Models (VFMs): A Comparative Survey Under Distributional Shifts"

Prof. Sanghamitra Bandyopadhyay
Professor, Machine Intelligence Unit
Director, Indian Statistical Institute, Kolkata

Abstract

Reliable deployment of vision models in real-world scenarios requires both robustness to natural perturbations and transparency in model predictions. While convolutional and transformer-based architectures have achieved high accuracy on curated datasets, their stability and explainability often degrade under authentic corruptions and distributional shifts common in practical settings. This study presents a unified, comparative evaluation of leading vision foundation models using real-world datasets containing diverse, naturally occurring perturbations. We benchmark model robustness and systematically assess a range of interpretability methods including GradCAM, LayerCAM, ScoreCAM, EigenCAM, Attention Rollout, and Chefar's Method across both clean and corrupted conditions. Our results highlight substantial differences in how modern models handle real-world challenges and reveal key trade-offs between robustness and interpretability. These findings provide actionable guidance for selecting and deploying vision models that remain reliable and transparent when faced with the complexities of real data.

Code available here: `https://github.com/suharoy/Robustness-Interpretebility-VFM-Survey`

1 Introduction

The rapid evolution of computer vision over the past decade has been driven by increasingly powerful deep learning models and the growing availability of large, diverse, real-world datasets. Convolutional Neural Networks (CNNs) and, more recently, transformer-based architectures such as the Vision Transformer (ViT) and Swin Transformer, have become the foundation of state-of-the-art visual recognition systems. These models are typically benchmarked on curated datasets that are designed to be representative, but real-world deployment presents additional challenges that are not fully captured by such controlled benchmarks.

In practical applications, vision models must operate on images drawn from complex environments where quality, content, and acquisition conditions are highly variable. Real-world datasets often contain images with a wide range of naturally occurring perturbations, including sensor noise, motion blur, lighting variations, occlusions, and geometric distortions. Distributional shifts arising from changes in imaging hardware, environmental context, or even the target population can significantly impact model performance. As a result, robustness to such perturbations has become a critical requirement for any model intended for deployment beyond the laboratory. At the same time, interpretability is essential for understanding and validating model behavior in unpredictable real-world scenarios. The ability to provide transparent, trustworthy explanations for model predictions is particularly important in sensitive domains such as healthcare, autonomous driving, and scientific research, where unexpected data artifacts or corruptions may lead to failures that require rapid diagnosis and intervention.

Recent years have seen the introduction of several robustness benchmarks (such as ImageNet-C and AugMix) that attempt to simulate common corruptions found in real data. Meanwhile, a range of interpretability techniques including GradCAM, LayerCAM, Attention Rollout, and CausalCAM have been developed to visualize and explain the features driving model decisions. However, most prior studies have evaluated these aspects separately, often using artificial perturbations rather than systematically analyzing models on authentic real-world data and naturally occurring shifts.

This work aims to bridge that gap by conducting a comprehensive, unified evaluation of leading vision foundation models spanning both CNN and transformer-based architectures using real-world datasets containing diverse, naturally occurring perturbations. We benchmark model robustness by exposing each architecture to both synthetic and authentic corruptions, and we systematically analyze the reliability and informativeness of modern interpretability methods under these challenging conditions. Through this comparative analysis, we seek to

provide actionable insights for practitioners and researchers concerned with deploying vision models in dynamic, real-world settings. By highlighting the interplay between robustness and interpretability on realistic data, our study offers a foundation for the development and selection of models that are both reliable and transparent under practical, non-ideal conditions.

2 Related Works

The field of computer vision has undergone a marked transformation over the past decade, shifting from manual, feature engineering approaches to the widespread adoption of deep learning architectures. Early breakthroughs, such as AlexNet and ResNet [8], established convolutional neural networks (CNNs) as the foundational architecture for image recognition tasks. These models enabled hierarchical feature extraction directly from raw pixel data, significantly improving performance benchmarks and reducing reliance on handcrafted features.

The subsequent introduction of transformer-based architectures to computer vision, notably with the Vision Transformer (ViT) [9], introduced self-attention mechanisms that allow models to capture long-range dependencies within visual data. This advance facilitated effective training on large-scale datasets with less need for domain-specific architectural modifications. The result was the emergence of Vision Foundation Models (VFMs): large, general-purpose models pre-trained on diverse visual corpora, capable of being fine-tuned for a wide variety of downstream tasks.

The scope of VFMs has further expanded with the development of multi-modal models, such as CLIP [11], which jointly learn from vision and language, and prompt-based segmentation frameworks like SAM [12]. These approaches enable prompt-driven interaction and adaptation to diverse tasks, supporting applications that go beyond single-task classification or detection. Recent survey articles [28, 29, 31] have emphasized the increasing centrality of segmentation and prompt engineering for both research and application in VFMs. VFMs are now being adopted in domain-specific contexts where reliability is critical, such as autonomous driving [33] and medical imaging [38, 3]. These application-driven surveys highlight both the technical advances and the unique deployment challenges associated with large, general-purpose models in regulated or high-stakes environments. Despite these developments, two fundamental challenges remain insufficiently addressed in the literature:

Reliable deployment of VFMs requires consistent performance under a wide range of real-world conditions, including distributional shifts, image corruptions, and various input perturbations [14]. While early CNNs exhibited limited robustness to such variations, newer architectures including transformer-based and self-supervised models like DINO [13] and DINOv2 [59] have reported improvements in transferability and resilience to perturbations. Nevertheless, systematic, empirical comparisons of robustness across different VFM families are still lacking. Existing studies often restrict analysis to a single architecture, dataset, or type of perturbation, limiting broader generalization.

As VFMs increase in size and complexity, model transparency and interpretability become essential, particularly for sensitive or regulated domains. Classical explainable AI (XAI) methods such as GradCAM [17], LayerCAM [18], LIME [23], and SHAP [24] were primarily developed for CNNs and may not be directly applicable to transformer-based or self-supervised models. Recent advances, such as Attention Rollout [57] for transformer attribution, CausalCAM [58] for causal visual explanation, and lightweight saliency frameworks like AD-CAM [32], represent active efforts to adapt XAI methodologies to new VFM architectures. Additionally, models like DINOv2 incorporate dedicated interpretability modules based on attention mechanisms. Despite this progress, comprehensive benchmarking and direct comparison of interpretability techniques across model types remains rare. There is an increasing trend towards both post-hoc (explanation after prediction) and ante-hoc (inherently interpretable) approaches for VFMs [7, 30], but a unified, systematic assessment is still needed.

Although several recent survey papers provide overviews of model architectures, training strategies, and domain-specific adaptations, most lack direct, empirical side-by-side evaluation of robustness and interpretability for VFMs. Studies tend to focus on architectural novelty, application domain, or isolated benchmarking efforts, without offering unified experimental frameworks that span both CNN and transformer-based models or address both robustness and interpretability using the latest XAI techniques.

This gap motivates the present survey. The goal is to perform direct, empirical benchmarking of VFMs under controlled real-world perturbations (such as noise, flips, brightness, blur, and rotation) and to evaluate the applicability and effectiveness of interpretability methods including but not limited to DINOv2, Attention Rollout across both CNN and transformer-based models.

This survey aims to (i) provide a comparative, empirical framework for VFM robustness and interpretability, (ii) systematically synthesize augmentation and XAI techniques as applied to current VFM architectures, and (iii) generate actionable insights for practitioners and researchers targeting deployment in reliability- and transparency-critical scenarios.

Table 1 summarizes recent major surveys, core VFM/model/method papers, and XAI techniques, highlighting the landscape, their robustness/augmentations stance, and open gaps. Newer methods and models are explicitly included.

Table 1: Summary of Recent Survey Papers and Vision Foundation Models (VFMs)

Paper	Year	Domain	Key Models	Benchmarks / Datasets	Robustness / Augmentations	Key Gaps / Limits
Remote Sensing VFMs [1]	2025	Remote Sensing	ResNet, Swin, ViT, SatMAE, DINO	Satellite, Multispectral, SAR	Pretraining, SSL; not systematic robustness	Domain-specific; lacks transform benchmarking
Unification of Gen. & Disc. VFMs [2]	2024	General Vision	GANs, Diffusion, ViT, SAM, CLIP	ImageNet, LAION, custom	Evolution, taxonomy, generalization	Lacks empirical or robustness comparison
Medical Image VFMs [3]	2025	Medical Imaging	ViT, SAM, TransUNet, SwinUNet	MRI, CT, Ultrasound, Pathology	Domain adaptation, adapters, compression	No transform-based robustness benchmarks
Pathology FMs [4]	2025	Pathology	ViT, SSL, CLIP, DINO, Multi-modal	TCGA, GTEx, OpenPath	Data curation, adaptation, eval tasks	No transform-based robustness assessment
FM Era in Vision [5]	2025	General Vision	CLIP, SAM, BLIP, Florence, OpenCLIP	ImageNet, COCO, LAION	High-level: architectures, robustness as challenge	No empirical transform-based robustness
VLM Benchmarks [6]	2025	Vision-Language	CLIP, GPT-4V, BLIP-2, LLaVA, etc.	38+ VLM datasets	Benchmarks, eval metrics, some robustness	Focus on VLMs, not pure image FM robustness
Explainability and VFMs [7]	2025	Explainability	CLIP, BLIP, LLaVA, SAM, DINO	VQA-X, custom	XAI taxonomy, post-hoc or ante-hoc, eval	No robustness under transformations
ResNet [8]	2016	Vision Model	ResNet	ImageNet	Baseline; skip connections; strong but not robust to corruptions	Milestone; sparked robustness and augmentation research
ViT [9]	2020	Vision Model	Vision Transformer	ImageNet, JFT-300M	Transformer backbone; large-scale data; needs augmentation	Early ViTs weak to corruptions, later improved

Continued on next page

Table 1 – *Continued from previous page*

Paper	Year	Domain	Key Models	Benchmarks / Datasets	Robustness / Augmentations	Key Gaps / Limits
Swin Transformer [10]	2021	Vision Model	Swin Transformer	ImageNet, COCO	Hierarchical, windowed self-attention, better transfer	Gains robustness over vanilla ViT; still tested
CLIP [11]	2021	Vision-Language	CLIP (ViT, ResNet)	LAION, WebImage Text, ImageNet	Large-scale text supervision; zero-shot robustness	Benchmark for VFM; some limitations with noisy data
SAM [12]	2023	Vision Model	Segment Anything (SAM)	SA-1B	Prompts for segmentation; foundation model	Not directly tested for robustness to noise/corruption yet
Segment Anything Survey [28]	2024	Segmentation Survey	SAM, variants, segment anything FMs	SA-1B, other segmentation datasets	Examines promptable segmentation, robustness in benchmarks	Gaps in evaluating under noise, domain shifts
Foundation Model Segmentation Survey [29]	2024	Segmentation Survey	SAM, OneFormer, Mask2Former, others	General segmentation, remote sensing, medical	Covers foundation segmentation model landscape, transfer	Practical robustness or augmentation results still missing
Prompt Engineering VLMs [31]	2024	Prompting Survey	CLIP, BLIP, SAM, GPT-4V, etc.	VLM benchmarks, COCO, LAION	Prompting methods, empirical VLM robustness	Mostly qualitative; little on vision-only robustness
FM Visualization Challenges [30]	2024	Visualization XAI	CLIP, DINO, SAM, others	Visualization datasets, general	Examines XAI and visualization for foundation models	Practical XAI methods on transformers still underdeveloped
AD-CAM [32]	2024	XAI Interpretability	AD-CAM, GradCAM, LayerCAM	Standard vision datasets	Lightweight, interpretable CNN CAM method	Limited to CNNs; adaptation to transformers not shown
Brain Tumor GradCAM [34]	2024	Medical XAI	GradCAM, CNNs	Brain MRI, medical datasets	Examines clinical utility of XAI	Limited generalization; not tested on FMs
Review of Explainable AI in Medical Imaging [36]	2024	XAI Survey	GradCAM, LIME, SHAP, others	Medical imaging datasets	Reviews XAI for medical imaging, practical impact	Lacks benchmarking across new FMs
Why Explainability Matters for LFM [37]	2024	XAI Survey/AI Policy	LFM, VFM, XAI methods	General, policy datasets	Explains XAI's importance in FMs, AI safety	Conceptual; empirical comparisons absent
FM in Autonomous Driving [33]	2024	Auto Driving Survey	BEV, LSS, 3D SegFM, Trans4Trans	Driving, perception benchmarks	Covers FMs for driving, robustness in complex scenes	Limited evaluation of augmentation and robustness methods
FM in Medical Imaging [38]	2024	Medical Imaging Survey	SAM-Med3D, MedCLIP, ViT, CNN, more	20+ medical datasets	Reviews adaptation, robustness, XAI in medical FMs	Side-by-side model comparison, but little on augmentations

Continued on next page

Table 1 – *Continued from previous page*

Paper	Year	Domain	Key Models	Benchmarks / Datasets	Robustness / Augmentations	Key Gaps / Limits
Unifying Understanding VFM [35]	2024	General VFM Survey	CLIP, DINO, SAM, ViT	General vision, multi-task	Seeks unified view, VFM taxonomy	Does not benchmark robustness or augmentation methods
DINO [13]	2021	SSL Vision Model	DINO (ViT, ResNet)	ImageNet, more	Self-supervised; strong robustness properties	Robust, transferable features; needs more evals
DINOv2 [59]	2023	SSL Vision Model	DINOv2 (ViT, ResNet)	ImageNet, custom	Improved SSL, attention heads for robustness, XAI	Transformer-centric; interpretability methods included but empirical comparison lacking
ImageNet-C [14]	2019	Benchmark	Any vision model	ImageNet-C	Standard for common corruptions, perturbation robustness	Limited to 15 types of synthetic corruptions
AugMix [15]	2020	Data Augmentation	Any vision model	ImageNet, CIFAR	Improves robustness via augmentation mixing	Used in many modern robust training regimes
AutoAugment [16]	2019	Data Augmentation	Any vision model	CIFAR, ImageNet	Learns augmentation policies	Requires computation, not universal
GradCAM [17]	2017	Explainability	Any CNN	ImageNet, domain	Visual explanation for decisions	Not robust for ViT/transformer without adaptation
LayerCAM [18]	2021	Explainability	Any CNN	Various datasets	Hierarchical CAM, better localization	Less tested on transformers
Attention Rollout (Chefer et al.) [57]	2021	Explainability	ViT, Transformers	ImageNet, general	Propagates attention for attribution in ViT	Only for transformers; comparison with CNN-XAI is open
CausalCAM [58]	2023	Explainability	CNN, ViT, DINO, SAM	ImageNet, domain	Causal visual explanations, robust to spurious features	Still under evaluation for clinical reliability
LIME [23]	2016	Explainability	Any model	Various datasets	Model-agnostic local explanation	Computationally heavy, not specific for vision
SHAP [24]	2017	Explainability	Any model	Various datasets	Unified feature attribution, theoretical basis	Can be slow, not always visual
AlexNet [25]	2012	Vision Model	AlexNet	ImageNet	First breakthrough deep CNN for ImageNet	Largely outperformed now, historical importance
VGGNet [26]	2015	Vision Model	VGGNet	ImageNet	Deep, simple architecture; baseline	Overfitting, inefficient for modern FMs
GoogLeNet [27]	2015	Vision Model	GoogLeNet or Inception	ImageNet	Inception modules, efficient	Less used today, historical

3 Review of Robustness and Interpretability Challenges in Modern Vision Models

The trajectory of computer vision research over the last decade has been shaped by a succession of architectural innovations and the proliferation of large-scale annotated datasets. Deep convolutional neural networks (CNNs) such as AlexNet, VGGNet, and ResNet [25, 26, 8] fundamentally altered the landscape of image classification by introducing deep feature hierarchies and enabling end-to-end supervised learning. While these models delivered significant gains in accuracy and transferability, initial benchmarks were typically restricted to clean, high-quality datasets, leaving questions about their stability under realistic data corruptions and perturbations. It was not until the systematic introduction of explicit robustness benchmarks and automated data augmentation protocols [14, 15, 16] that these concerns became central to the evaluation of visual models.

The adaptation of transformer architectures to vision, first realized in the Vision Transformer (ViT) and further refined through hybrid and hierarchical designs such as the Swin Transformer [10], has introduced new capabilities and challenges. The Swin Transformer's hierarchical window-based self-attention facilitates both local and global context aggregation, achieving state-of-the-art performance on a variety of recognition and dense prediction tasks. Reviews focused on medical imaging, pathology, and remote sensing [3, 4, 38] underscore the value of these architectures in domains where spatial relationships and multi-scale reasoning are critical. However, the same surveys highlight a persistent gap: most empirical evaluations of transformers emphasize accuracy and transfer learning, with robustness to real-world corruptions typically receiving only qualitative or anecdotal attention [33, 38].

The emergence of large-scale foundation models such as CLIP [11] and SAM [12] has further accelerated progress, enabling models to generalize across tasks and modalities through the use of multi-modal supervision and prompt-driven learning [31, 28, 29]. Despite their demonstrated flexibility, these models exhibit increased architectural complexity and opacity, raising new concerns regarding transparency and reliability in high-stakes applications [7, 37, 30].

Efforts to address model interpretability have produced a range of post-hoc and ante-hoc explanation methods. Classical approaches such as GradCAM [17], LayerCAM [18], LIME [23], and SHAP [24] have been widely applied to CNNs, yielding visualizations of feature importance and model decision logic. However, their direct application to transformer-based architectures, especially those with hierarchical attention mechanisms like Swin, is problematic. Recent surveys [7, 30, 32] point out that standard attention visualizations in transformers do not necessarily capture the class-discriminative or causal structure required for robust model explanations, and new techniques such as Attention Rollout [57], CausalCAM [58], and advances in self-supervised models like DINOv2 [59] are still under active investigation.

In medical and safety-critical domains, the necessity for transparency is particularly acute [34, 36, 38, 4]. While interpretability tools are increasingly leveraged for model validation and regulatory compliance, existing literature demonstrates a strong bias toward CNN-based models, with transformer-based and foundation models often remaining black boxes in clinical workflows. Moreover, empirical studies exploring the interplay between robustness (to corruptions and domain shift) and interpretability are exceedingly rare.

Most existing reviews and taxonomies [1, 35, 6, 29] catalogue VFMs by architecture, training protocol, or application, but do not present integrated, empirical comparisons of robustness and interpretability. Where robustness is considered, it is frequently discussed as a training augmentation strategy rather than an explicit axis of model assessment. Similarly, interpretability is often evaluated only on clean data, without consideration of model behavior under stress. Thus, a unified empirical pipeline for benchmarking both robustness and interpretability, especially as applied to contemporary transformer-based and foundation models remains absent from the literature.

In conclusion, while progress in vision architectures and explainability methods has been substantial, comprehensive studies examining the robustness and transparency of modern VFMs under real-world conditions are conspicuously lacking. Addressing this void is essential for informed model selection, safe deployment, and the development of next-generation VFMs tailored for reliability-critical applications.

Objectives of This Study

This study is designed to fill the identified gap by conducting an empirical, side-by-side comparison of robustness and interpretability across representative families of vision architectures. Specifically, the following objectives are addressed:

1. **Model Coverage:** Evaluate a diverse set of contemporary vision models including classical CNNs, Vision Transformers, Swin Transformers, and self-supervised models using standardized classification benchmarks such as ImageNet-100, ImageNetV2, Tiny ImageNet, CIFAR-10 etc.

2. **Robustness Evaluation:** Systematically assess each model's sensitivity to a spectrum of common image corruptions and perturbations, including geometric and photometric transformations (e.g., flips, rotations, blur, brightness variation, and Gaussian noise). Quantify the effect of these perturbations on classification accuracy and analyze model resilience profiles.

3. **Interpretability Assessment:** Apply both traditional (e.g., GradCAM, LayerCAM) and transformer-adapted (e.g., Attention Rollout, Chefar's Method) explanation techniques to all models, on both unperturbed and perturbed inputs, to evaluate the stability and informativeness of visual explanations under stress.

4. **Unified Evaluation Framework:** Develop and implement an extensible pipeline for direct, quantitative and qualitative comparison of robustness and interpretability outcomes across model families, corruption types, and explanation methods.

5. **Deployment Guidance:** Derive actionable insights regarding the trade-offs between accuracy, robustness, and explainability, with practical recommendations for selecting and deploying vision foundation models in domains with stringent requirements for transparency and reliability.

By integrating robustness and interpretability benchmarking in a single experimental framework, this work aims to advance the empirical understanding of current vision foundation models and to inform their safe and effective adoption in real-world and safety-critical environments.

4 Methodology: Architectural and Methodological Progression

The present study evaluates a diverse spectrum of vision foundation models, drawing from the major architectural milestones that have defined image recognition in the deep learning era. The analysis covers models spanning from classic convolutional neural networks, through hybrid and transformer-based networks, to the latest self-supervised and explainability advances. Robustness and interpretability are benchmarked using a selection of real-world and challenging datasets, providing a holistic view of model performance under practical deployment conditions. Our evaluation begins with standard convolutional networks such as ResNet, DenseNet, Inception, ResNeXt, and the efficiency-focused EfficientNet and MobileNet families. Each

of these architectures incrementally improved the representational power, scalability, and deployment feasibility of CNNs.

The next major leap was the adoption of transformer-based models for vision, initially exemplified by the Vision Transformer (ViT). ViT replaced convolution with patch embeddings and global self-attention, dramatically changing the learning dynamics and enabling long-range dependency modeling. This innovation was refined by data-efficient approaches such as DeiT, which improved training efficiency with limited data, and by hierarchical, scalable models like the Swin Transformer, which introduced local windowed attention for dense prediction and multi-scale processing. Further advances emerged from self-supervised learning, as in DINOv2, which learns robust and transferable representations from unlabeled data using a vision transformer backbone. Such approaches have proven especially resilient to out-of-distribution perturbations, a property critical for real-world robustness. Table 2 provides a concise summary of all evaluated models, including their main innovations and functional differences.

All models are evaluated with ImageNet-pretrained weights unless otherwise stated. Each architecture is analyzed for both robustness and interpretability, using the most appropriate explainability method based on its structure.

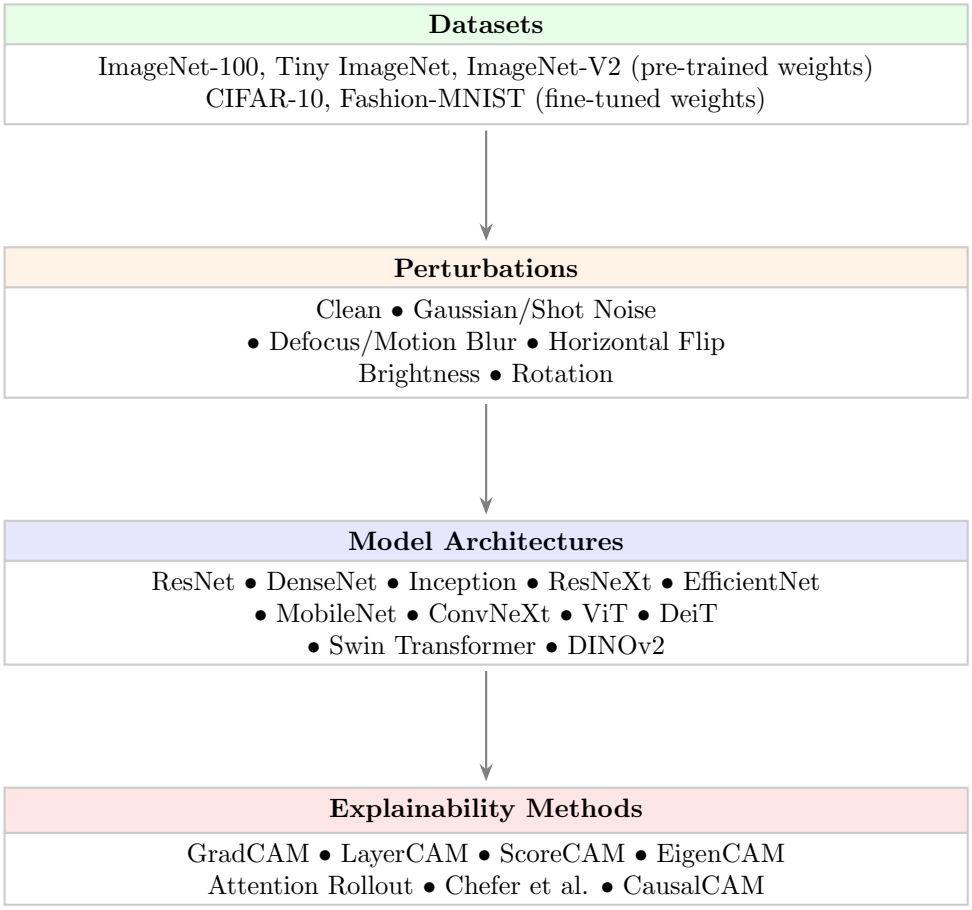


Figure 1: Evaluation pipeline with distinct stages

4.1 Datasets

To provide a thorough assessment of model robustness and interpretability, we evaluate all methods on a diverse set of image classification benchmarks. The datasets used are as follows:

- **CIFAR-10** [21]: Contains 60,000 32×32 color images evenly distributed among 10 object categories: *airplane, automobile, bird, cat, deer, dog, frog, horse, ship,* and *truck*. CIFAR-10 is a standard benchmark for general object recognition and allows testing of both accuracy and robustness on low-resolution, varied images.
- **Fashion-MNIST** [60]: Comprises 70,000 grayscale images (28×28 pixels) of clothing and fashion items, each labeled as one of 10 classes: *T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag,* and *ankle boot*. This dataset emphasizes shape and texture discrimination, and is widely used for evaluating model and attribution reliability on simple, structured visual data.
- **ImageNet-100**: A subset of the original ImageNet dataset, containing 100 balanced classes selected to maintain diversity and realism. Images are high resolution and cover a wide variety of objects and animals, providing a rigorous testbed for large-scale visual recognition and interpretability.
- **Tiny ImageNet**: Contains 100,000 64×64 color images spanning 200 different object categories, each with a high degree of intra-class variation. The smaller image size and increased number of classes, compared to ImageNet-100, introduce additional classification challenges and simulate domain shift.
- **ImageNet-V2** [61]: An out-of-distribution test set for ImageNet, constructed by independently collecting images that match the original ImageNet synsets. This dataset is designed to probe generalization and robustness to natural distribution shift, especially for models pretrained on ImageNet-1K.

The combination of these datasets ensures that models are evaluated under a spectrum of real-world and controlled conditions, across varying image resolutions, semantic complexities, and levels of distributional shift.

4.2 Image Perturbations and Corruptions

To systematically evaluate the robustness of each vision model and explainability method, we subject all test images to a controlled set of perturbations. These perturbations were chosen to represent a range of common real-world image corruptions, each targeting a different aspect of the input signal. The following transformations were implemented:

- **Clean (Identity)**: The original, unmodified image is used as a baseline for comparison against all other perturbations.
- **Horizontal Flip**: Each image is mirrored along the vertical axis (left-right inversion). This transformation simulates scenarios where an object may appear in a different orientation, and tests whether models and attribution methods are invariant to such flips. *(Implemented as: `img.transpose(Image.FLIP_LEFT_RIGHT)`)*
- **Rotation**: Images are rotated by a fixed angle (e.g., 30 degrees). This simulates out-of-plane viewing conditions or camera misalignment, which are common in both natural and clinical imaging environments. Robust models should retain predictive accuracy and stable saliency despite such geometric distortion. *(Implemented as: `img.rotate(degrees)`)*
- **Blur**: Gaussian blur is applied using a specified kernel radius (e.g., radius = 5). This mimics sensor noise, motion blur, or defocus artifacts, and is particularly relevant for real-world and medical images captured under suboptimal conditions. The challenge for models is to preserve salient features when spatial detail is degraded. *(Implemented as: `img.filter(ImageFilter.GaussianBlur(radius))`)*

- **Brightness Adjustment**: Image brightness is scaled by a multiplicative factor (e.g., factor = 1.5). This transformation emulates variations in ambient lighting or camera exposure. Both model robustness and interpretability should ideally remain stable as brightness changes, provided the semantic content of the image is preserved. *(Implemented as: `ImageEnhance.Brightness(img).enhance(factor)`)*

- **Gaussian Noise**: Additive Gaussian noise is introduced to each pixel, simulating sensor noise or digital transmission artifacts. The image array is perturbed by drawing noise values from a normal distribution with a given standard deviation (e.g., sigma = 42), then re-normalized to the valid pixel range. This tests the model's ability to generalize under severe signal degradation. *(Implemented as: array + np.random.normal(...); see `add_gaussian_noise` function)*

These perturbations are applied independently and identically to all datasets. By covering a range of realistic corruptions—spanning geometric, photometric, and noise-based artifacts—our evaluation framework provides a rigorous assessment of both model performance and explanation stability under real-world variation.

4.3 Evaluated Model Architectures

To ensure a fair and representative comparison across the landscape of modern vision models, our study includes a diverse set of architectures spanning both convolutional neural networks (CNNs) and transformer-based models. Table 2 summarizes the key design principles, input requirements, and distinctive features of each architecture evaluated in this work.

- **Convolutional Networks (CNNs):** This group includes ResNet, DenseNet, Inception, ResNeXt, EfficientNet, and MobileNet families, each introducing architectural advances such as residual connections, dense connectivity, multi-scale processing, grouped convolutions, and compound scaling for improved efficiency and accuracy.

- **Hybrid and Transformer Models:** ConvNeXt bridges convolutional and transformer design by incorporating normalization and activation functions typical of transformer networks into a ConvNet structure. Pure transformers (ViT, DeiT) leverage patch-based input and global self-attention, enabling large-scale pretraining and data-efficient transfer learning. The Swin Transformer introduces hierarchical and windowed attention to better capture local and global dependencies.

- **Self-Supervised and Foundation Models:** DINOv2 exemplifies the latest trend of self-supervised, large-scale pretraining, which improves robustness and transferability by learning from unlabeled data. It also incorporates dedicated heads for interpretability.

Table 2: Architectural comparison and lineage of evaluated models.

Model	Type	Key Innovation	Input Size	Notable Features
ResNet-50/101	CNN	Residual connections	224×224	Skip connections, deep/flexible
DenseNet-121	CNN	Dense connectivity	224×224	Feature reuse, all-to-all connections
Inception V3	CNN	Multi-scale inception-modules	299×299	Parallel convolutions, auxiliary loss
ResNeXt50	CNN	Split-transform merge-blocks	224×224	Grouped convolutions, cardinality

(continued on next page)

(continued from previous page)

Model	Type	Key Innovation	Input Size	Notable Features
EfficientNet-B3	CNN	Compound scaling	300×300	Balanced scaling (depth, width, res)
MobileNetV3-L	CNN	Lightweight, mobile-optimized	224×224	Squeeze-and-Excite, hard-swish
ConvNeXt-Base	CNN/Hybrid	ConvNet-redesigned, transformer cues	224×224	Large kernels, LayerNorm, GELU
ViT-B/16	Transformer	Patch-embedding, global-self-attn	224×224	Patchwise input, pure attention
DeiT-B/16	Transformer	Data-efficient Vision Transformer-training	224×224	Distillation token, improved sample efficiency
Swin-B	Transformer	Hierarchical, windowed-attn	224×224	Local+shifted windows, scalable
DINOv2	Self-supervised Transformer	Self-supervised, foundation model	224×224	Unlabeled pretraining, robust, interpretability heads

4.4 Explainability Methods: Overview and Application

In our experiments, each method was applied to the recommended target layers for every architecture, as summarized above. For CNNs, CAM-based methods were found most stable and interpretable, especially under corruption. For transformer-based models, attention-based methods provided more meaningful attributions, although interpretability decreased under strong perturbations.

Table 3: Summary of explainability methods used in this study.

Method	Key Features	Target Layer(s)	Model Type	Dataset Suitability	Application
GradCAM [17]	Gradient-based localization; heatmap over salient regions	Last convolutional layer (CNN); final attention block (ViT)	CNNs, ViTs (adapted)	Most effective for structured, object-centric datasets (e.g., CIFAR-10, ImageNet-100)	Applied to all CNNs and transformers; most reliable for CNNs under all perturbations
LayerCAM [18]	Aggregates gradients from all spatial locations; improved localization	All convolutional layers (CNN)	CNNs	Best for fine-grained localization on natural images	Used for all CNNs; compared under all perturbations
ScoreCAM [48]	Uses class scores instead of gradients; less sensitive to vanishing gradients	Last convolutional layer (CNN)	CNNs	Suitable for datasets with high inter-class similarity (e.g., Tiny ImageNet)	Applied to all CNNs; notable performance under domain shift
EigenCAM [49]	Based on principal components of feature activations; gradient-free	Last convolutional layer (CNN)	CNNs	Robust to perturbations in structured datasets	Used on all CNNs; good stability on corrupted images
Attention Rollout [57]	Aggregates attention weights across transformer layers for attribution	All transformer attention blocks	Transformers (ViT, Swin)	Best for high-resolution, complex scenes (ImageNet, Tiny ImageNet)	Applied to ViT, DeiT, Swin, DINOv2; more robust for transformers under noise/rotation
Chefer et al. [57]	Combines attention and gradient-based relevance propagation	Attention and MLP blocks in transformers	Transformers (ViT, Swin)	Well-suited for both natural and domain-shifted data	Used for ViT, DeiT, DINOv2; compared for interpretability under perturbations

5 Results

This section presents a systematic evaluation of both robustness and explainability for leading vision models, across diverse datasets and perturbation types. Results are grouped into (1) small-scale, fine-tuned benchmarks and (2) large-scale, transfer evaluations using pretrained models. Both quantitative metrics and qualitative visualizations are used to comprehensively compare methods and architectures.

5.1 Fine-Tuned Evaluation: CIFAR-10 and FashionMNIST

The first analysis examines model performance when fine-tuned on standard benchmarks. Table 4 reports top-1 accuracy for ResNet-18, ViT Small, and Swin Tiny on CIFAR-10 and FashionMNIST under both clean and perturbed inputs. Perturbations include horizontal flip, rotation, blur, brightness adjustment, and Gaussian noise.

Table 4: Robustness (Top-1 Accuracy, %) of Fine-Tuned Models on CIFAR-10 and FashionMNIST.

Model	Clean	Hor. Flip	Rotation	Blur	Brightness	G. Noise
CIFAR-10						
ResNet18	85.09	85.32	66.83	65.55	83.36	31.40
ViT Small	84.76	84.79	75.20	84.83	82.35	60.57
Swin Tiny	86.21	86.65	71.09	86.04	84.34	16.14
FashionMNIST						
ResNet18	79.03	79.35	53.73	47.78	77.06	18.24
ViT Small	83.85	83.47	63.50	83.73	81.17	57.46
Swin Tiny	91.49	91.71	70.31	91.55	90.10	54.05

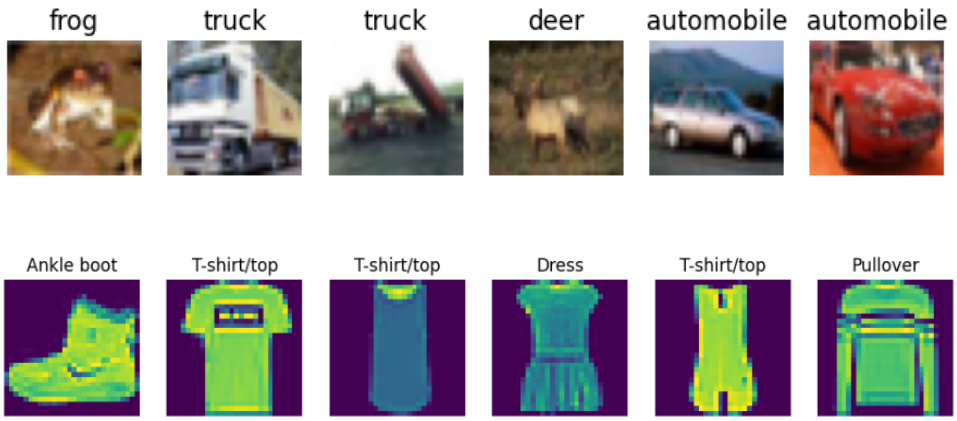


Figure 2: Original images for CIFAR-10 and FashionMNIST

Key Observations: All architectures perform similarly on clean and simply perturbed images (flips, brightness). Significant accuracy drops occur under noise and rotation, especially for ResNet-18. ViT and Swin exhibit relatively higher robustness to rotation, but may degrade quickly under strong noise.

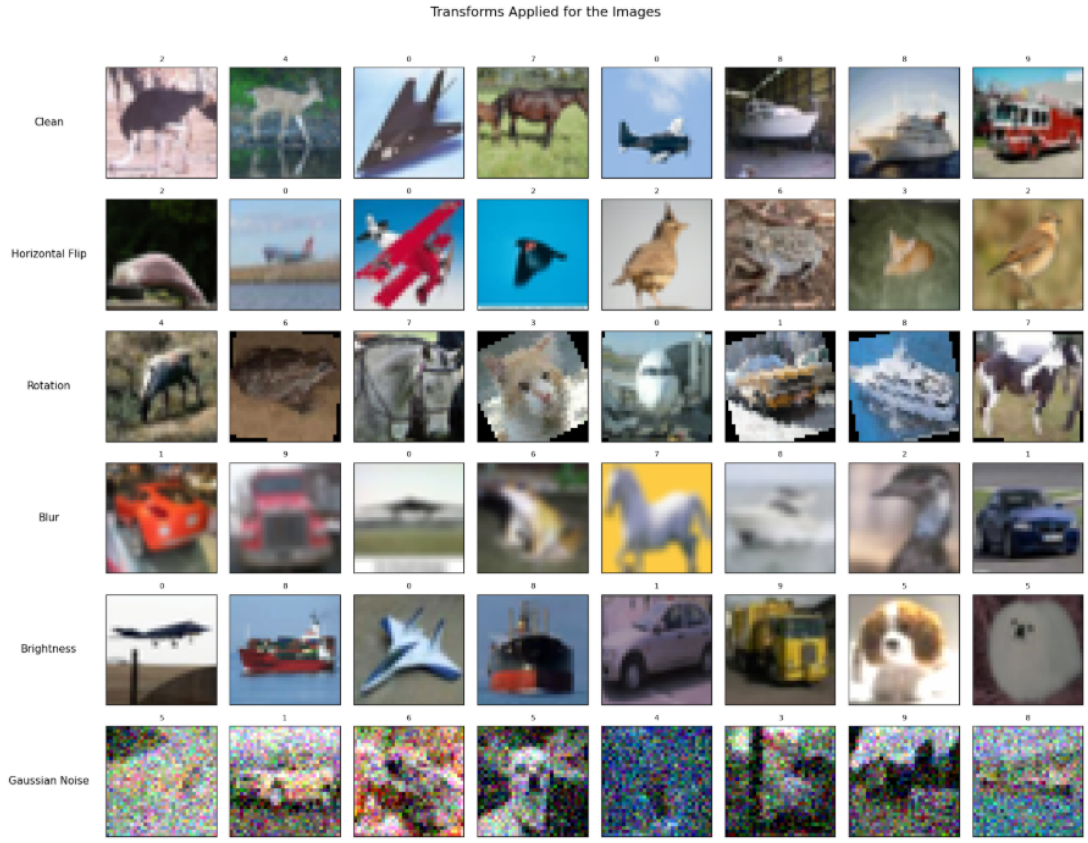


Figure 3: CIFAR-10 images after perturbations

Figure 4: FashionMNIST images after perturbations

5.2 Transfer Evaluation: Large-Scale and Realistic Datasets

For a more challenging assessment, all models are evaluated using pretrained weights on ImageNet-100, Tiny ImageNet, and ImageNet V2, with the same perturbation suite. Results are summarized in Tables 5, 6, and 7.

Table 5: Top-1 match and similarity scores for models on ImageNet-100.

Model	Method	Clean	Hor. Flip	Rotation	Blur	Brightness	G. Noise
ResNet-50	GradCAM	1	0.93	0.74	0.98	1	0.98
ResNet-50	ScoreCAM	1	0.75	0.61	0.95	0.96	0.90
ResNet-50	LayerCAM	1	0.80	0.61	0.90	0.96	0.91
ResNet-50	EigenCAM	1	0.78	0.60	0.94	0.99	0.96
ResNet-101	GradCAM	1	0.93	0.90	0.99	1	0.98
ResNet-101	ScoreCAM	1	0.72	0.71	0.93	0.98	0.88
ResNet-101	LayerCAM	1	0.66	0.64	0.94	0.98	0.94
ResNet-101	EigenCAM	1	0.80	0.75	0.93	0.99	0.96
DenseNet-121	GradCAM	1	0.98	0.94	0.99	1	1
DenseNet-121	ScoreCAM	1	0.87	0.60	0.90	0.82	0.82
DenseNet-121	LayerCAM	1	0.80	0.63	0.87	0.89	0.91
DenseNet-121	EigenCAM	1	0.87	0.69	0.95	0.98	0.93
EfficientNet-B3	GradCAM	1	0.96	0.85	0.98	1	0.99
EfficientNet-B3	ScoreCAM	1	0.35	0.15	0.42	0.83	0.62
EfficientNet-B3	LayerCAM	1	0.88	0.61	0.91	0.94	0.91
EfficientNet-B3	EigenCAM	1	0.89	0.91	0.98	0.98	0.66
MobileNetV3-L	GradCAM	1	0.59	0.71	0.78	0.97	0.84
MobileNetV3-L	ScoreCAM	1	0.56	0.48	0.78	0.92	0.91
MobileNetV3-L	LayerCAM	1	0.68	0.44	0.86	0.98	0.91
MobileNetV3-L	EigenCAM	1	0.76	0.56	0.93	0.98	0.92
ConvNeXt-Base	GradCAM	1	0.60	0.60	0.88	0.98	0.96
ConvNeXt-Base	ScoreCAM	1	1	1	1	1	1
ConvNeXt-Base	LayerCAM	1	0.64	0.54	0.64	0.93	0.82
ConvNeXt-Base	EigenCAM	1	0.83	0.73	0.96	0.96	0.95
ResNeXt50	GradCAM	1	0.85	0.90	0.99	1	0.99
ResNeXt50	ScoreCAM	1	0.77	0.70	0.89	0.97	0.86
ResNeXt50	LayerCAM	1	0.79	0.67	0.94	0.98	0.95
ResNeXt50	EigenCAM	1	0.75	0.77	0.96	0.99	0.94
Inception V3	GradCAM	1	0.95	0.97	0.99	1	1
Inception V3	ScoreCAM	1	0.71	0.60	0.90	0.90	0.87
Inception V3	LayerCAM	1	0.90	0.83	0.95	0.99	0.99
Inception V3	EigenCAM	1	0.77	0.70	0.76	0.94	0.82
ViT-B/16	GradCAM	0.98	0.30	0.22	0.69	0.71	0.63
ViT-B/16	Attn-R	1	0.55	0.41	0.84	0.93	0.83
ViT-B/16	Chefar's	1	0.49	0.46	0.64	0.72	0.52
ViT-B/16 DINOv2	GradCAM	0.99	0.28	0.15	0.81	0.83	0.32
ViT-B/16 DINOv2	Attn-R	1	0.66	0.47	0.93	0.97	0.91
DeiT-B/16	GradCAM	0.92	0.25	0.12	0.75	0.69	0.71
DeiT-B/16	Attn-R	1	0.60	0.56	0.95	0.95	0.96
DeiT-B/16	Chefar's	1	0.68	0.49	0.78	0.88	0.82
Swin-B	GradCAM	0.90	0.53	0.44	0.46	0.68	0.63
Swin-B	Attn-R	1	0.88	0.90	0.73	0.96	0.90

Table 6: Top-1 match and similarity scores for models on Tiny ImageNet

Model	Method	Clean	Hor. Flip	Rotation	Blur	Brightness	G. Noise
ResNet-50	GradCAM	1	0.71	0.71	0.78	0.93	-
ResNet-50	ScoreCAM	1	0.60	0.59	0.51	0.47	0.43
ResNet-50	LayerCAM	1	0.60	0.57	0.73	0.92	0.74
ResNet-50	EigenCAM	1	0.55	0.62	0.63	0.91	0.78
ResNet-101	GradCAM	1	0.53	0.54	0.86	0.98	0.71
ResNet-101	ScoreCAM	1	0.50	0.61	0.65	0.91	0.64
ResNet-101	LayerCAM	1	0.57	0.53	0.63	0.94	0.71
ResNet-101	EigenCAM	1	0.59	0.65	0.72	0.95	0.67
DenseNet-121	GradCAM	1	0.76	-	-	0.97	-
DenseNet-121	ScoreCAM	1	0.58	0.44	0.67	0.92	0.84
DenseNet-121	LayerCAM	1	0.52	0.54	0.84	0.92	0.83
DenseNet-121	EigenCAM	1	0.44	0.44	0.79	0.81	0.67
EfficientNet-B3	GradCAM	1	0.65	0.85	0.88	1	0.89
EfficientNet-B3	ScoreCAM	1	0.57	0.23	0.43	0.82	0.50
EfficientNet-B3	LayerCAM	1	0.53	0.69	0.67	0.97	0.58
EfficientNet-B3	EigenCAM	1	0.99	0.93	0.33	0.93	0.47
MobileNetV3-L	GradCAM	1	0.60	-	-	0.65	-
MobileNetV3-L	ScoreCAM	1	0.44	0.71	0.21	0.97	0.78
MobileNetV3-L	LayerCAM	1	0.45	0.60	0.64	0.95	0.68
MobileNetV3-L	EigenCAM	1	-	0.66	0.86	0.83	0.75
ConvNeXt-Base	GradCAM	1	0.49	0.62	0.83	0.97	0.62
ConvNeXt-Base	ScoreCAM	1	0.75	0.41	0.31	0.66	0.48
ConvNeXt-Base	LayerCAM	1	1	0.56	0.86	1	0.79
ConvNeXt-Base	EigenCAM	1	0.49	0.65	0.81	0.96	0.86
ResNeXt50	GradCAM	1	0.65	0.63	0.89	0.96	0.69
ResNeXt50	ScoreCAM	1	0.54	0.45	0.73	0.89	0.65
ResNeXt50	LayerCAM	1	0.59	0.49	0.51	0.95	0.71
ResNeXt50	EigenCAM	1	0.67	0.76	0.69	0.95	0.78
Inception V3	GradCAM	1	0.95	-	-	1	-
Inception V3	ScoreCAM	1	0.78	0.69	0.79	0.87	0.60
Inception V3	LayerCAM	1	0.83	0.93	0.85	0.99	0.86
Inception V3	EigenCAM	1	0.67	0.89	0.93	0.99	0.76
ViT-B/16	GradCAM	0.94	0.30	0.30	0.51	0.85	0.33
ViT-B/16	Attn-R	1	0.50	0.57	0.79	0.93	0.75
ViT-B/16	Chefar's	1	0.35	0.20	0.56	0.29	0.28
ViT-B/16 DINOv2	GradCAM	0.89	0.13	-	-	-	-
ViT-B/16 DINOv2	Attn-R	1	0.58	0.63	0.95	0.96	0.94
DeiT-B/16	GradCAM	0.91	0.20	0.18	0.46	0.75	0.39
DeiT-B/16	Attn-R	1	0.72	0.73	0.80	0.89	0.79
DeiT-B/16	Chefar's	1	0.35	0.43	0.74	0.81	0.65
Swin-B	GradCAM	0.85	0.73	-	0.76	0.73	0.82
Swin-B	Attn-R	1	0.96	0.86	0.96	0.99	0.74

Table 7: Top-1 match and similarity scores for models on ImageNet V2

Model	Method	Clean	Hor. Flip	Rotation	Blur	Brightness	G. Noise
ResNet-50	GradCAM	1	0.82	-	-	0.99	-
ResNet-50	ScoreCAM	1	0.63	0.51	0.62	0.89	0.63
ResNet-50	LayerCAM	1	0.65	0.58	0.71	0.96	0.81
ResNet-50	EigenCAM	1	0.51	0.44	0.46	0.92	0.84
ResNet-101	GradCAM	1	0.20	-	-	0.99	0.61
ResNet-101	ScoreCAM	1	-	0.59	0.57	0.85	0.66
ResNet-101	LayerCAM	1	0.43	0.59	0.74	0.96	0.70
ResNet-101	EigenCAM	1	0.46	0.77	0.71	0.92	0.66
DenseNet-121	GradCAM	1	0.91	-	0.98	1	-
DenseNet-121	ScoreCAM	1	0.65	0.67	0.74	0.96	0.70
DenseNet-121	LayerCAM	1	0.65	0.62	0.89	0.98	0.84
DenseNet-121	EigenCAM	1	0.32	0.38	0.35	0.95	0.43
EfficientNet-B3	GradCAM	1	0.52	0.90	-	1	-
EfficientNet-B3	ScoreCAM	1	0.36	0.74	0.44	0.94	0.37
EfficientNet-B3	LayerCAM	1	-	0.67	0.64	0.98	0.42
EfficientNet-B3	EigenCAM	1	0.74	0.97	0.97	0.95	0.51
MobileNetV3-L	GradCAM	1	0.53	0.69	0.86	0.99	-
MobileNetV3-L	ScoreCAM	1	0.61	0.45	-	0.89	0.53
MobileNetV3-L	LayerCAM	1	0.64	0.51	0.68	0.97	0.58
MobileNetV3-L	EigenCAM	1	0.61	0.44	0.44	0.99	0.42
ConvNeXt-Base	GradCAM	1	0.18	0.68	-	1	0.98
ConvNeXt-Base	ScoreCAM	1	0.55	0.49	-	0.94	0.69
ConvNeXt-Base	LayerCAM	1	0.56	0.56	0.44	0.99	0.88
ConvNeXt-Base	EigenCAM	1	0.45	0.56	0.49	0.99	0.97
ResNeXt50	GradCAM	1	0.36	0.93	-	0.99	-
ResNeXt50	ScoreCAM	1	0.42	0.70	0.55	0.80	0.59
ResNeXt50	LayerCAM	1	0.55	0.70	0.71	0.97	0.75
ResNeXt50	EigenCAM	1	0.63	0.74	0.77	0.96	0.74
Inception V3	GradCAM	1	0.83	-	-	1	-
Inception V3	ScoreCAM	1	0.61	0.69	0.47	0.93	0.48
Inception V3	LayerCAM	1	0.69	0.77	0.74	1	0.77
Inception V3	EigenCAM	1	-	0.45	-	0.87	0.52
ViT-B/16	GradCAM	1	0.39	0.44	-	0.93	0.74
ViT-B/16	Attn-R	1	0.54	0.44	0.78	0.91	0.65
ViT-B/16	Chefar's	1	0.47	0.43	0.37	0.79	0.46
ViT-B/16 DINOv2	GradCAM	1	0.40	0.17	0.53	0.89	0.29
ViT-B/16 DINOv2	Attn-R	1	0.75	0.72	0.92	0.92	0.86
DeiT-B/16	GradCAM	1	0.27	-	0.46	0.78	-
DeiT-B/16	Attn-R	1	0.69	0.70	0.79	0.95	0.70
DeiT-B/16	Chefar's	1	0.35	0.50	0.31	0.50	0.74
Swin-B	GradCAM	1	-	-	-	-	-
Swin-B	Attn-R	1	0.96	0.84	0.94	0.99	0.70

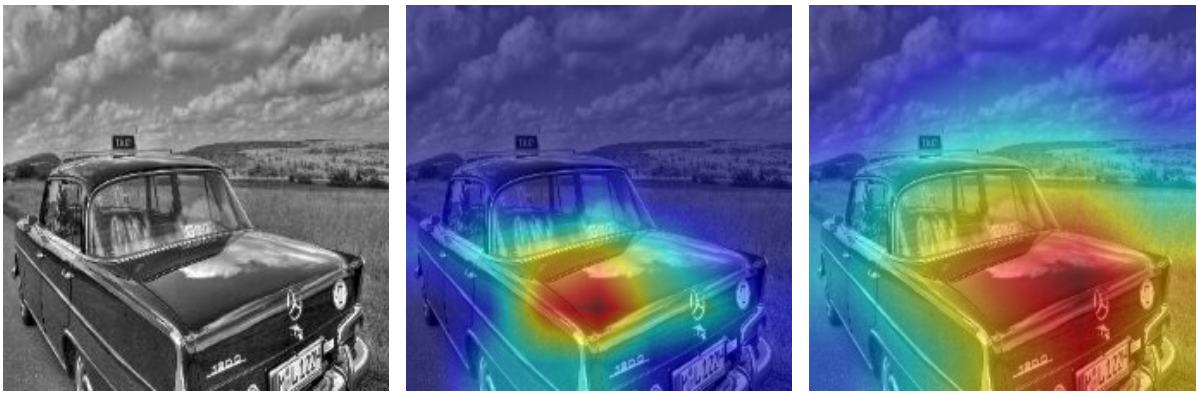
Key Findings:

- **Increased Sensitivity:** Pretrained models show greater drops in performance under corruption, especially for distribution-shifted datasets (Tiny ImageNet, ImageNet V2).
- **Method-Specific Trends:** Attention-based explainability (Attn-R, Chefer's) provides greater robustness for transformers under severe corruptions, while GradCAM and EigenCAM remain stable for CNNs.

- **Variability Across Models:** Some explainability methods (e.g., ScoreCAM) are less stable under domain shift, and no single method outperforms others consistently.
- **Dataset and Perturbation Dependence:** No architecture or explainability tool shows universal superiority—robustness and interpretability are context-dependent.

5.3 Qualitative Analysis: Visual Attribution Under Corruption

To complement the quantitative evaluation, we present qualitative visualizations of attribution maps across representative models and perturbations. These overlays provide intuitive insight into the regions each architecture attends to during prediction, highlighting differences in interpretability and robustness.



Original Image (ImageNetV2) GradCAM: ResNet-50 LayerCAM: ResNet-50

Figure 5: Comparison of original image and XAI overlays for ResNet-50.

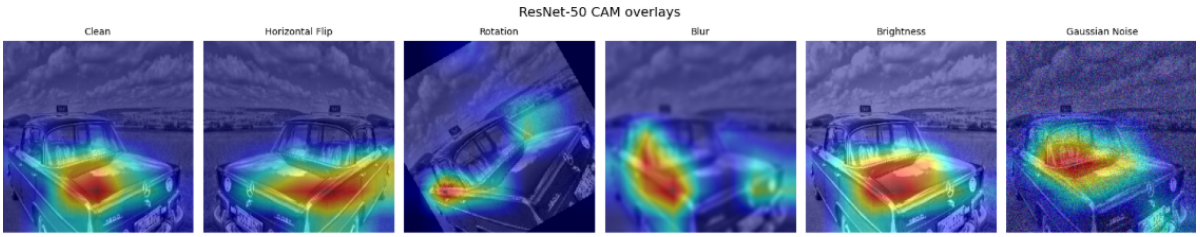


Figure 6: Grad-CAM overlays under all perturbations (L-R : Clean, Horizontal Flip, Rotation, Blur, Brightness, Gaussian)

Interpreting CNN-Based CAM Overlays. The above figures present the qualitative visualizations for CNN architectures, specifically GradCAM and LayerCAM overlays on ResNet-50. In these overlays, the highlighted regions (typically in red and yellow) indicate the spatial locations that most strongly influence the model's prediction. CNN-based attribution methods are inherently local and spatially precise, often resulting in sharply focused attention maps that closely align with salient objects or regions in the image. This spatial precision is evident both for clean images and under various perturbations, where GradCAM and LayerCAM consistently produce interpretable, object-centric heatmaps. Thus, for CNNs, visual explanations can be directly interpreted as the primary evidence driving the model's classification, and changes in attribution across perturbations reflect the model's robustness and sensitivity to image transformations.

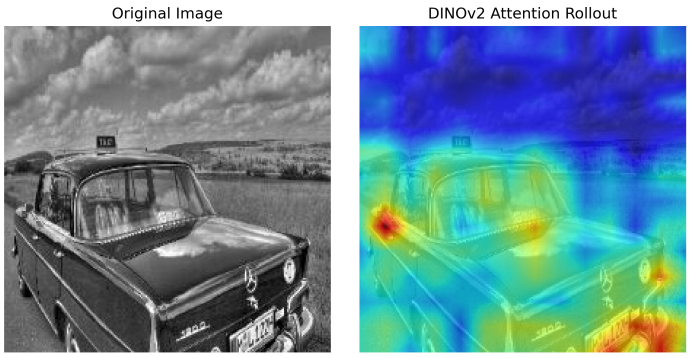


Figure 7: Comparison of original image and DINOv2 Overlay

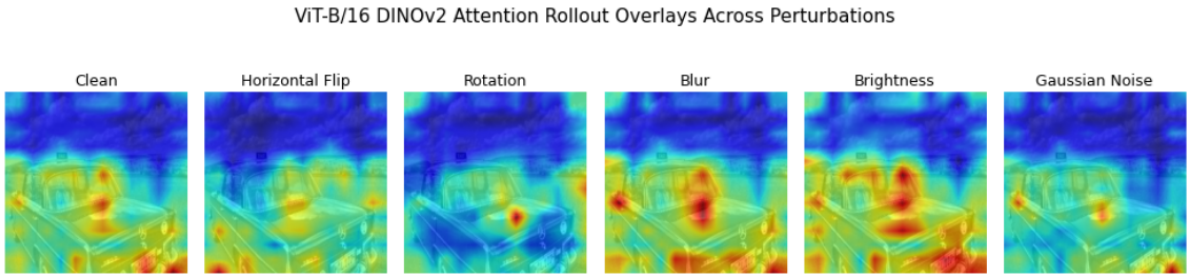


Figure 8: ViT-B16 DINOv2 Attn-R overlays under all perturbations (L-R : Clean, Horizontal Flip, Rotation, Blur, Brightness, Gaussian Noise)

Interpreting Transformer Attention Overlays. The figures illustrate qualitative results for the transformer-based DINOv2 model using attention rollout overlays. Unlike CNNs, vision transformers distribute their attention globally across the image, resulting in more diffuse and spatially distributed attribution maps. The highlighted regions in these overlays represent the aggregate importance assigned by the transformer's self-attention layers, often capturing broader semantic context rather than sharply localized features. As a result, transformer attention maps may appear less focused but offer insight into how the model integrates information from multiple regions. Interpretation of these overlays should thus emphasize the global and context-dependent nature of transformer reasoning. Notably, differences in attention patterns across perturbations reveal both the flexibility and sensitivity of transformers to distributional shifts, as well as the relative stability of global attribution compared to the local focus observed in CNNs.

For addistional viisualizations and overlays, see Appendix 9 10

6 Discussion

This study systematically benchmarks the robustness and interpretability of a range of vision foundation models (VFMs) across both small-scale and large-scale datasets, with special focus on their behavior under realistic distributional shifts. Our findings provide fresh empirical evidence for the nuanced interplay between model architecture, input perturbations, and the effectiveness of explainability techniques—a subject of growing importance in both academic and industrial AI deployment.

6.1 Comparison with Prior Work

Recent literature has emphasized the remarkable performance of transformer-based vision models on curated benchmarks [9, 10, 59], yet concerns about their robustness to real-world corruptions and their interpretability remain only partially addressed [14, 30, 7]. Previous

studies such as [3, 4] have shown that CNNs and hybrid architectures often maintain superior robustness under common corruptions, particularly in medical imaging and remote sensing domains, while ViTs and Swin Transformers excel at capturing global dependencies but are sometimes more sensitive to specific geometric or noise-based perturbations.

Our results reinforce and extend these findings. On clean datasets, transformers such as ViT and Swin achieve state-of-the-art accuracy, consistent with the literature. However, under severe perturbations especially noise and rotation CNNs like ResNet and DenseNet maintain relatively higher stability, a trend reported in robustness-focused benchmarks [15, 38]. Self-supervised models such as DINOv2 demonstrate promising robustness, likely due to their learning of generalizable features from diverse, unlabeled data [59].

In terms of interpretability, the literature has established that GradCAM and related CAM techniques provide focused, class-relevant saliency maps for CNNs [17, 18, 48, 49], but their reliability declines on transformers, where attention-based approaches like Attention Rollout and Chefer et al.'s method are preferred [57, 58]. Our qualitative overlays and similarity scores echo these trends: for CNNs, CAM methods produce sharp, interpretable heatmaps even under moderate corruption, while for ViT/Swin, attention-based methods are more stable but yield diffuse and sometimes less intuitive attributions, especially under distributional shift.

6.2 Nuanced Robustness-Interpretability Trade-offs

A key contribution of this work is its direct comparison of robustness and interpretability under the same experimental pipeline. Our data show that no single architecture or explainability method is universally superior; rather, their effectiveness is highly context-dependent. For instance, while Swin Transformer outperforms on clean, structured data (e.g., FashionMNIST), its attributions become unstable under Gaussian noise. ResNet and DenseNet, in contrast, show a smaller drop in both accuracy and attribution similarity under most corruptions.

Moreover, our findings indicate that robust classification performance does not guarantee stable or interpretable attributions. In some cases, models maintained accuracy but their attribution maps shifted significantly, raising concerns for deployment in safety-critical contexts where explanation stability is paramount [7, 37]. The results align with the growing view that model and explainability selection must be application-specific, as noted by recent surveys [29, 35].

6.3 Practical Implications and Applications

From a deployment perspective, these results suggest that practitioners should avoid relying solely on headline accuracy or a single attribution method when selecting VFMs for real-world use. Instead, both robustness and interpretability should be empirically benchmarked on relevant datasets and expected perturbations. In domains such as healthcare, where explainable and robust predictions are non-negotiable, our unified evaluation pipeline offers a valuable template for model selection and validation.

7 Limitations and Scope for Future Work

While this study covers a broad range of models, datasets, and explainability tools, certain limitations remain. The perturbation suite, though diverse, does not exhaustively cover all forms of real-world shift (e.g., adversarial attacks, rare artifacts, or task-specific corruptions). Likewise, while our evaluation of interpretability is extensive, it remains primarily visual and quantitative; future studies should consider user studies or task-based assessments to better judge explanation utility in practice.

An important direction for future research is the application of this unified framework to specialized domains, particularly medical imaging. Here, both robustness to input variability

and trust in model decisions are critical for clinical adoption. Extending our approach to include medical datasets and expert-in-the-loop evaluation would further advance the field. Additionally, the ongoing development of hybrid and domain-adaptive XAI methods, as well as robustness-aware model training, represents a promising avenue for achieving more reliable and transparent vision AI.

8 Conclusion

This study provides a unified and empirical assessment of robustness and interpretability across a broad spectrum of modern vision foundation models, spanning both convolutional and transformer-based architectures. By systematically evaluating model performance and attribution stability under a range of real-world perturbations and distributional shifts, we demonstrate that no single model or explainability method offers universal superiority. CNNs tend to provide more robust and interpretable attributions under most corruptions, while transformer-based models excel on clean data but are often more sensitive to noise and geometric transformations. Attention-based explainability techniques are essential for transformer models but can yield diffuse and less stable saliency maps in challenging conditions.

Our findings highlight the necessity of application-specific benchmarking for both model selection and explainability pipeline design. Reliable, transparent deployment in real-world or safety-critical domains requires joint consideration of both robustness and interpretability, tested under conditions that reflect actual operational variability. This work lays the foundation for future efforts in developing domain-adapted, robust, and explainable vision systems, with direct applicability to areas such as medical imaging, where these requirements are most acute.

Acknowledgments

I would like to express my heartfelt gratitude to **Prof. Sanghamitra Bandyopadhyay**, **Director and Professor, Machine Intelligence Unit, Indian Statistical Institute, Kolkata**, for her invaluable guidance, encouragement, and constructive feedback throughout the course of this project. Working under her mentorship has been an immensely enriching experience, both academically and personally.

I am deeply thankful to **Dr. Malay Bhattacharyya**, **Associate Professor, Machine Intelligence Unit, ISI Kolkata,** for his guidance, valuable inputs, and support during the project. I sincerely thank the **IASc–INSA–NASI** Summer Research Fellowship Program for giving me this opportunity and for making it possible to be part of such a meaningful learning journey. I am also grateful to the **Bioinformatics Lab of the Machine Intelligence Unit**, ISI Kolkata for providing an inspiring and collaborative research environment, and to the **Department of Data Science and Analytics, Central University of Rajasthan** for their support and encouragement. Last but not the least, my parents, without their support nothing would have been possible.

Declaration

I declare that this report is based on my original work during the Summer Research Fellowship Program (SRFP) 2025 under the IASc-INSA-NASI. Any external code, data, or resources have been cited and acknowledged to the best of my knowledge. All executed codes are available at `https://github.com/suharoy/Robustness-Interpretebility-VFM-Survey`

References

[1] Y. Lu, Y. Guo, S. Chen, J. Li, X. X. Zhu, and P. Ghamisi, "Vision Foundation Models in Remote Sensing: A Survey," *IEEE Geoscience and Remote Sensing Magazine*, Early Access, 2025.

[2] X. Liu, T. Zhou, C. Wang, Y. Wang, Y. Wang, Q. Cao, W. Du, Y. Yang, J. He, Y. Qiao, and Y. Shen, "Toward the unification of generative and discriminative visual foundation model: a survey," *The Visual Computer*, vol. 41, no. 11, pp. 3371–3412, 2024.

[3] P. Liang, B. Pu, H. Huang, Y. Li, H. Wang, W. Ma, and Q. Chang, "Vision Foundation Models in Medical Image Analysis: Advances and Challenges," *arXiv preprint arXiv:2502.14584*, 2025.

[4] D. Li, G. Wan, X. Wu, X. Wu, A. J. Nirmal, C. G. Lian, P. K. Sorger, Y. R. Semenov, and C. Zhao, "A Survey on Computational Pathology Foundation Models: Datasets, Adaptation Strategies, and Evaluation Tasks," *arXiv preprint arXiv:2501.15724*, 2025.

[5] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation Models Defining a New Era in Vision: A Survey and Outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 2245–2268, 2025.

[6] Z. Li, X. Wu, H. Du, H. Nghiem, and G. Shi, "Benchmark Evaluations, Applications, and Challenges of Large Vision Language Models: A Survey," *arXiv preprint arXiv:2501.02189*, 2025.

[7] R. Kazmierczak, E. Berthier, G. Frehse, and G. Franchi, "Explainability and vision foundation models: A survey," *Information Fusion*, vol. 122, p. 103184, 2025.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *CVPR*, 2016.

[9] A. Dosovitskiy, L. Beyer, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2020.

[10] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *ICCV*, 2021.

[11] A. Radford, J. W. Kim, et al., "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021.

[12] A. Kirillov, E. Mintun, et al., "Segment Anything," *arXiv preprint arXiv:2304.02643*, 2023.

[13] M. Caron, H. Touvron, et al., "Emerging Properties in Self-Supervised Vision Transformers," *ICCV*, 2021.

[14] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations," *ICLR*, 2019.

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

[16] E. D. Cubuk, B. Zoph, et al., "AutoAugment: Learning Augmentation Policies from Data," *CVPR*, 2019.

[17] R. R. Selvaraju, M. Cogswell, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *ICCV*, 2017.

[18] P. Jiang, C. Zhang, et al., "LayerCAM: Exploring Hierarchical Class Activation Maps," *IEEE Transactions on Image Processing*, 2021.

[19] J. Deng, W. Dong, et al., "ImageNet: A Large-Scale Hierarchical Image Database," *CVPR*, 2009.

[20] T.-Y. Lin, M. Maire, et al., "Microsoft COCO: Common Objects in Context," *ECCV*, 2014.

[21] A. Krizhevsky, G. Hinton, et al., "Learning Multiple Layers of Features from Tiny Images," Technical report, 2009.

[22] S. Yang, H. Zhang, et al., "MedMNIST v2: A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification," *Scientific Data*, 2022.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," *KDD*, 2016.

[24] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *NeurIPS*, 2017.

[25] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *NeurIPS*, 2012.

[26] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *ICLR*, 2015.

[27] C. Szegedy, W. Liu, et al., "Going Deeper with Convolutions," *CVPR*, 2015.

[28] Z. Shen, Y. He, J. Liu, H. Wang, S. Wang, and Z. Zeng, "A Survey on Segment Anything Model," *arXiv preprint arXiv:2404.02925*, 2024.

[29] Z. Gong, T. Zhang, Y. Li, C. Zhu, X. Wang, X. Lin, Y. Yang, and D. Lin, "Image Segmentation in Foundation Model: A Survey," *arXiv preprint arXiv:2403.15790*, 2024.

[30] Y. Li, X. Han, L. Song, X. Zhang, and H. Qu, "Foundation models meet visualizations: Challenges and opportunities," *arXiv preprint arXiv:2401.12963*, 2024.

[31] J. Wang, S. Wang, J. Liu, J. Liu, Y. Xie, Y. Yang, Y. Wang, R. He, Y. Zhou, and P. S. Yu, "A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models," *arXiv preprint arXiv:2403.12404*, 2024.

[32] C. Zhao, M. Xu, J. Zhang, Y. Li, Y. Zhang, H. Zhou, Y. Lu, and D. Li, "AD-CAM: Enhancing Interpretability of Convolutional Neural Networks With a Lightweight Framework - From Black Box to Glass Box," *IEEE Transactions on Instrumentation and Measurement*, Early Access, 2024. DOI: 10.1109/TIM.2024.3385890

[33] X. Huang, B. Zhang, H. Yu, J. Zhang, and L. Wang, "A Survey for Foundation Models in Autonomous Driving," *arXiv preprint arXiv:2402.03015*, 2024.

[34] G. Asaithambi, S. Dhanasekar, V. Jeyalakshmi, and S. A. Basha, "A Grad-CAM based framework to interpret deep learning models for brain tumor detection and classification using MRI images," *Journal of King Saud University-Computer and Information Sciences*, In Press, 2024. DOI: 10.1016/j.jksuci.2024.04.032

[35] J. Wang, S. Wang, Y. Wang, J. Liu, J. Liu, Y. Xie, Y. Yang, R. He, and P. S. Yu, "Unifying Understanding of Vision Foundation Models: A Survey and Benchmark," *arXiv preprint arXiv:2311.16693*, 2024.

[36] M. Ahmad, A. Tiwari, and S. Prakash, "A review of explainable AI in medical imaging: implications and applications," *Artificial Intelligence Review*, 2024. DOI: 10.1007/s10462-024-10677-w

[37] R. Vinuesa, J. Fuchs, I. Ahmed, et al., "Why Explainability Matters for Large Foundation Models in AI Systems," *Nature Machine Intelligence*, vol. 6, pp. 196–198, 2024. DOI: 10.1038/s42256-024-00818-1

[38] J. Xue, T. Zhou, X. Wang, Z. Jiang, Z. Li, X. Ding, and W. Zhang, "Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision," *arXiv preprint arXiv:2404.02583*, 2024.

[39] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *CVPR*, 2017.

[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *CVPR*, 2016.

[41] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *ICML*, 2019.

[42] I. Radosavovic, X. Wang, R. Dollár, J. Girshick, and P. Dollár, "Designing Network Design Spaces," *CVPR*, 2020.

[43] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," *ICCV*, 2019.

[44] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," *CVPR*, 2017.

[45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training Data-efficient Image Transformers & Distillation through Attention," *ICML*, 2021.

[46] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *CVPR*, 2022.

[47] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," *WACV*, 2018.

[48] H. Wang, Z. Wang, H. Du, Z. Yang, Y. Liu, and X. Wang, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," *CVPR Workshops*, 2020.

[49] H. Wang, Z. Wang, H. Du, Z. Yang, Y. Liu, and X. Wang, "EigenCAM: Visual Explanations for Deep Convolutional Networks via Low-rank Approximation," *arXiv preprint arXiv:2008.00299*, 2020.

[50] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images (CIFAR-10, CIFAR-100)," Technical Report, 2009.

[51] `https://github.com/PatWie/imagenet100`

[52] D. Hendrycks and T. Dietterich, "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations (ImageNet-C)," *ICLR*, 2019.

[53] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *NeurIPS*, 2019.

[54] Ross Wightman, "PyTorch Image Models (timm)," GitHub repository, 2019. `https://github.com/huggingface/pytorch-image-models`

[55] J. Jacob Gildenblat and contributors, "pytorch-grad-cam: Gradient-based class activation maps for PyTorch," GitHub repository, 2021. `https://github.com/jacobgil/pytorch-grad-cam`

[56] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, 12, pp. 2825–2830, 2011.

[57] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," *CVPR*, 2021.

[58] B. Pan, H. Chefer, E. Arnold, and L. Wolf, "CausalCAM: Causal Explanations for Image Classification Models," *arXiv preprint arXiv:2302.01333*, 2023.

[59] M. Oquab, T. Darcet, M. M. Assran, et al., "DINOv2: Learning Robust Visual Features without Supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[60] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[61] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet Classifiers Generalize to ImageNet?" *International Conference on Machine Learning (ICML)*, pp. 5389–5400, 2019. `https://arxiv.org/abs/1902.10811`

Appendix

Figure 9: CAM Overlays for an image from ImageNet-100

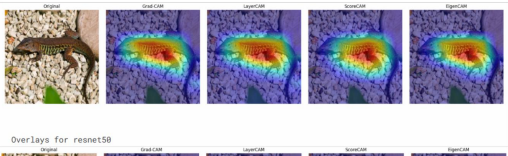
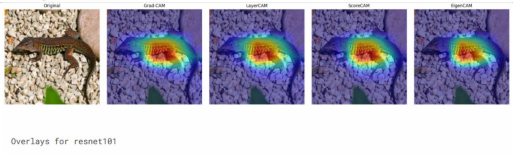
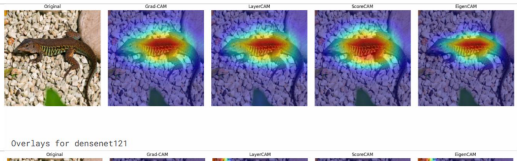
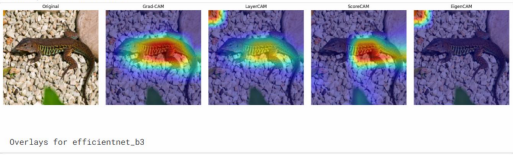
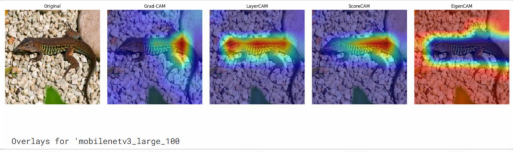
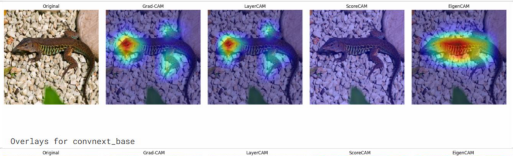
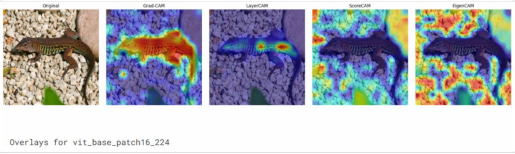
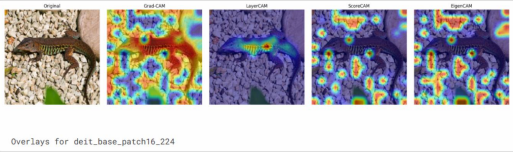
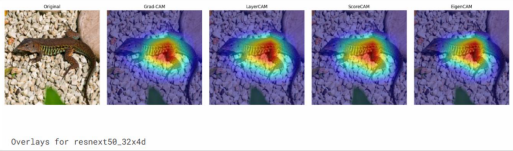
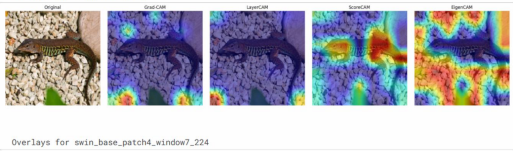
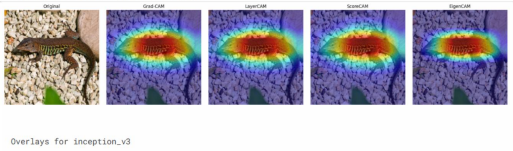
Model	Original | GradCAM | LayerCAM | ScoreCAM | EigenCAM	Target Layer
Res50		model.layer4[-1]
Res101		model.layer4[-1]
Dense121		model.features[-1]
EffB3		model.blocks[-1]
MobV3-L		model.blocks[-1]
ConvNeXt-B		model.stages[-1].blocks[-1]
ViT-B/16		model.blocks[-1].norm1
DeiT-V/16		model.blocks[-1].norm1
ResNeXt50		model.layer4[-1]
Swin-B		model.layers[-1].blocks[-1].norm1
Inception V3		model.Mixed_7c

Fig : Visual Comparison of CAM overlays for all baseline models on ImageNet-100

Figure 10: Perturbation GradCAM Overlays

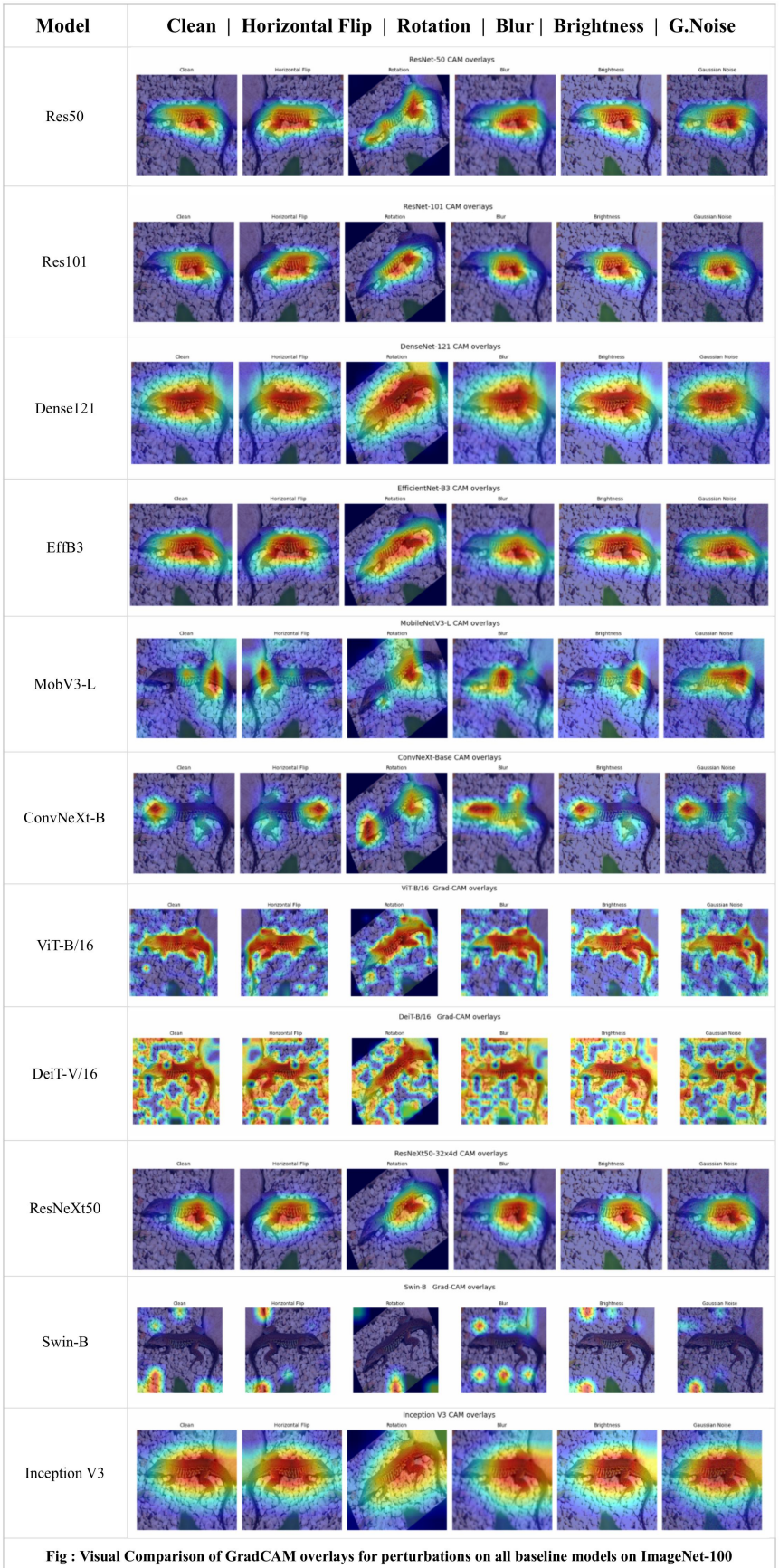


Fig : Visual Comparison of GradCAM overlays for perturbations on all baseline models on ImageNet-100