A Comprehensive Toolkit for Stress-Testing AI Models with Adversarial Attacks and Out-of-Distribution Data

Pooja Yakkala
py9363@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Sindhu Pasupuleti
sp7289@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Charitha Madhamsetty
cm7513@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

ACM Reference Format:
Pooja Yakkala, Sindhu Pasupuleti, and Charitha Madhamsetty. 2018. A Comprehensive Toolkit for Stress-Testing AI Models with Adversarial Attacks and Out-of-Distribution Data. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Problem Statement

As AI systems become integral to critical domains such as autonomous vehicles, medical diagnosis and content moderation, understanding and securing their behavior in unpredictable environments is more urgent than ever. While current deep learning models achieve high accuracy on standard benchmarks, they remain brittle when exposed to out-of-distribution (OOD) inputs or semantically adversarial examples. These vulnerabilities typically lie concealed behind normal tests but have the ability to result in catastrophic failures upon real deployment. Current robustness protocols largely focus on low level noise or simple perturbations, overlooking the deeper semantic and causal structures central to human perception.

Our research addresses this critical gap by suggesting a three-pronged stress-testing of AI vision models with semantically coherent and causally consistent OOD attacks. In contrast to existing works, which primarily manipulate pixels or overlay Gaussian noise, we attempt to know how models behave when visual inputs are coherent but fall apart semantically or causally. Specifically, we explore how models handle confusing contexts, structural inconsistencies, and causal mismatches challenges that are prevalent in real-world, open-world settings where the assumption of clean, ID data no longer holds.

First, we introduce *Semantic Confusion*, a generative OOD attack that creates misleading yet visually plausible images. Using a BART language model, we generate unusual or contradictory captions (e.g., "a zebra rug in a living room"), then synthesize corresponding images via a diffusion model. These hybrid samples are designed to challenge generalization in models like ResNet-18. Alongside, we propose the *Perceptual Concept Shift* metric, which quantifies semantic misalignment by comparing Grad-CAM attention maps, BLIP-generated region captions, and CLIP-based image-caption consistency. PCS enables quantitative assessment of a model's semantic coherence under confusion.

Second, we propose structural disruption, a structural OOD attack that disrupts part-whole relationships by altering the semantic and spatial integrity of object parts. We use SAM to generate fine-grained segmentation masks and DINOv2 to extract attention maps. We implement two disruption techniques: (1) simply blacking out the high-attention segmentation mask, and (2) swapping two or more segmentation masks. These perturbations retain local realism while disrupting global semantic consistency. To assess the structural impact, we present the Structural Divergence Score that computes model attention shifts by comparing Grad-CAM heatmaps between OOD and ID images.

Third, we introduce Causal Concept Subversion, a novel method for probing a model's reliance on causally critical features. After identifying key causal regions via Grad-CAM, we replace them with perceptually plausible but semantically conflicting patches, selected using both LPIPS and CLIP distances. These modifications are localized and blended to preserve visual integrity while misleading the model's internal reasoning. We further propose the Conceptual Fragility Index, which quantifies the volatility in model predictions across multiple adversarial variants, offering a fine-grained view into its conceptual robustness.

Collectively, our framework extends traditional robustness evaluations by introducing high level semantic, structural and causal stressors, offering a more holistic view of model reliability in real world settings. The paper is organized as follows: Section 2 has the literature review. Section 3 outlines the methodology for each proposed attack and the generative pipeline. Section 4 presents experimental results and analysis. Section 5 discusses the hypotheses, limitations, and insights. Section 6 concludes with directions for future work.

2 Literature Review

This paper summarizes recent advances in out of distribution (OOD) detection and adversarial robustness, outlining approaches and their limitations. One contribution is Adversarial Mixup (AM) training which combines mixup with FGSM perturbations while improving generalization and OOD detection. On CIFAR datasets, AM achieves AUROC of 98.1% outperforming methods like MSP but its reliance on synthetic OOD data and costly adversarial training limits scalability to larger datasets like ImageNet [6]. Adversarial training with DeepFool shifts decision boundaries to increase robustness but lowers attack success on MNIST from 92.4% to 45.7% and is limited by its narrow evaluation scope [12]. Weighted Random Forest and Weight Variance Regularization techniques are away from feature over-reliance, achieving accuracy to 76.4% while

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

maintaining clean accuracy at 92.5% though they are not tested on complicated models [3].

In genomics, MLR-OOD yields higher performance than conventional likelihood ratio methods with an AUROC of 0.822 though computational cost limits application [2]. Replacement of softmax with energy based scoring reduces FPR95 by 18.03% in WideResNet models but is costly and for pretrained models only [10]. JPEG compression with AR-GANs halts adversarial gradients at minimal accuracy loss but reduces performance at very high compression and remains vulnerable to adaptive attacks [5]. CW attacks are slower but yield more transferable and efficient adversarial examples compared to L-BFGS and IFGS [13].

The Detection by Attack model achieves 99.22% recall against adversarial inputs on MNIST but suffers from sensitivity to perturbation size and high computational cost, resulting in possible false negatives in adaptive attacks [14]. The FRODO framework enhances MRI segmentation adversarial robustness but is plagued with computational overhead and feature collapse in U-Net [8]. A federated learning system using deep active learning is 95% accurate for Retinal OCT, but it is confronted with manual labeling, hyperparameter sensitivity, and communication overhead [1].

Costa et al. (2024) highlight the lack of consistent evaluation metrics for adversarial defenses in Vision Transformers, hindering meaningful comparisons [4]. Lal et al. achieve 99.9% accuracy in the diagnosis of diabetic retinopathy against adversarial attacks with feature fusion and adversarial training [7]. Li, Du, and Zhang suggest preprocessing techniques like rotating images that are not majorly affected by adversarial effects but suffer low clean image accuracy and are not adaptive-resistant unless integrated into other defenses [9]. The Spectrum Simulation Attack manipulates frequency domain characteristics to achieve a 95.4% success rate on robust models like IncRes-v2ens with superior performance compared to spatial domain methods [11].

In contrast to current research that primarily addresses perturbations at the pixel level or changes in distribution, our approach injects semantic ambiguity, causal undermining, and structure destruction as a new dimension of adversarial stress not just aimed at visual appearance but also at semantic coherence of concepts in an image. In contrast to past approaches such as JPEG+AR-GAN that focus on low-level spatial distortion, our approach uses generative modeling to alter concepts and enables controlled experimentation in vision-language models' semantic understanding of high-level concepts. We propose new evaluation metrics, Perceptual Concept Shift, Conceptual Fragility Index and Structural Divergence Score, to identify semantic misalignment, prediction instability and sensitivity to structural reorganizations which is not done by prior metrics like AUROC or FPR95. By leveraging the combination of generative methods, attention analysis, and semantic evaluation, our framework offers a scalable and cognitively consistent approach to robustness testing.

State-of-the-art OOD detection and adversarial defenses often suffer from limited semantic diversity, reliance on handcrafted perturbations and metrics that inadequately capture conceptual or structural weaknesses. Techniques like adversarial training or preprocessing defenses remain focused on low level pixel manipulations and fail to generalize across tasks or domains. Our framework overcomes these constraints by generating diverse, visually coherent, and semantically challenging adversarial examples that probe a model's conceptual understanding. With PCS, SDS and CFI, we introduce a triadic evaluation paradigm that captures failure modes in semantic alignment, structural integrity, and causal reasoning. This work advances the field by exposing overlooked vulnerabilities in vision models and offering interpretable tools to help researchers assess and improve robustness.

3 Methodology

This paper proposes a stress-testing framework for vision-language models through three adversarial attack paradigms: *Semantic Confusion*, *Causal Concept Subversion*, and *Semantic Structure Attack*. The attacks generate OOD images to evaluate the generalization ability of a model outside of its training distribution. Experiments are conducted on two vision backbones, **ResNet18** and **EfficientNetB0** with **ImageNet** and the **Intel Image Classification Dataset**. Attack efficacy is measured using three diagnostic metrics: **Prompt Consistency Score**, **Structural Divergence Score**, and **Conceptual Fragility Index** within an end-to-end framework for estimating model resilience to semantically adversarial perturbations.

3.1 Datasets

This study employs two publicly available data sets: ImageNet-256 and the Intel Image Classification data set, collected from Kaggle.

The ImageNet-256 dataset is the downsampled variant of the original ImageNet dataset and consists of images from 256 object classes. There are between 1,000 and 5,000 RGB images per class, and a total of one million images. The images are standardized to a resolution of 256×256 pixels.

The Intel Image Classification dataset has 25,000 natural scene images categorized into six classes: buildings, forest, glacier, mountain, sea and street. Each class contains 3000 - 5000 images and are resized to 150×150 pixels. This dataset is often used for benchmarking scene classification models and algorithms.

3.2 Semantic Confusion (SC)

Semantic Confusion (SC) is an attack framework for measuring model robustness against semantically anomalous but visually plausible inputs. It evaluates a model's susceptibility to subtle contextual violations by generating such instances and comparing performance on EfficientNet-B0 and ResNet18.

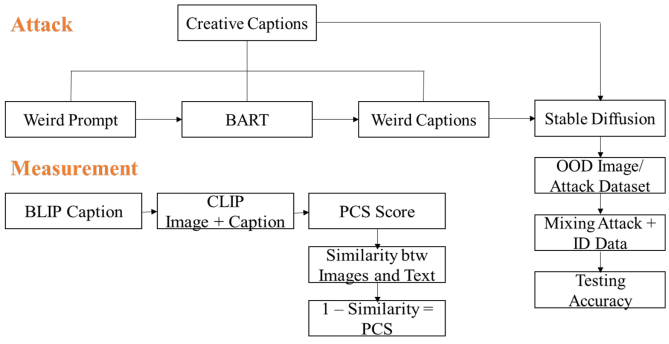

Figure 1: Methodology for SC

3.2.1 Experimental Procedure. The Semantic Confusion attack and measurement procedure involves the following steps:

- Load ResNet-18, EfficientNet-B0, BLIP, and CLIP ViT-B/32 models.
- Normalize dataset images to match dataset statistics.
- Use BART to generate semantically unusual base prompts and stylized variants.
- For each generated prompt:
  - Generate an image using Stable Diffusion v1.4.
  - Classify the image using ResNet-18 and EfficientNet-B0 to get predictions and softmax probabilities.
  - Generate a GradCAM heatmap and extract the salient causal region.
  - Use BLIP to caption the GradCAM region.
  - Compute PCS by measuring cosine similarity between the BLIP caption embedding and the CLIP image embedding.
  - Measure softmax entropy and cosine distance from cached in-distribution (ID) embeddings.
- High PCS (> 0.3), entropy, and cosine distance indicate semantic confusion and possible OOD behavior.

3.2.2 Preprocessing. **In-Distribution (ID) Datasets:** The ID datasets used are the ImageNet-256 dataset and the Intel Image Classification dataset. All input images are resized to 256×256 pixels and center-cropped to 224×224 before being normalized using the dataset's statistics:

- Mean: [0.485, 0.456, 0.406]
- Std: [0.229, 0.224, 0.225]

The preprocessing for In-distribution dataset is same for all three methods. For visualization, images are "unnormalized" back to the pixel range $[0, 1]$.

Out-of-Distribution (OOD) Image Generation: A set of atypical or incongruent prompts (e.g., "a zebra rug in a living room") is fed into the `facebook/bart-large-cnn` model to generate variants, with decoding parameters: max_length = 20, temperature = 1.0, top_k = 50, num_beams = 5. These prompts are rendered into images using `Stable Diffusion v1.4` via the HuggingFace Diffusers pipeline, saved at 512×512 resolution in PNG format, and resized/normalized to match the ID input pipeline.

3.2.3 Feature Extraction. **Image Features:** For each sample (ID and generated), we extract a 512-dimensional feature vector $\mathbf{f}_{\text{img}}$ from the penultimate layer of ResNet-18 and EfficientNet-B0:

$$\mathbf{f}_{\text{img}} = \text{Backbone}_{\text{features}}(x), \quad \mathbf{f}_{\text{img}} \in \mathbb{R}^{512}$$

A reference pool of 512 randomly sampled ImageNet features is cached to estimate the distribution of in-domain feature embeddings for cosine distance calculations.

Region Features (GradCAM): GradCAM localizes semantically meaningful regions of the image for the predicted class. For ResNet-18, feature activations are extracted from `model.layer4[-1]` and for EfficientNet-B0 from `model.features[-1]`. These regions are described and compared to full image semantics. We will be using the same GradCAM extraction method for all three attacks.

Textual Semantics (BLIP and CLIP): BLIP is used to caption the Grad-CAM area, generating a text summary of the causal part of the image. The captions, along with the original prompt or image caption, are embedded using `openai/clip-vit-base-patch32`. Semantic similarity is estimated by cosine similarity among both embeddings.

3.2.4 Models and Parameters.

- **ResNet-18 and EfficientNet-B0:** Pretrained on ImageNet and used in `eval()` mode.
- **Stable Diffusion v1.4:** Used to synthesize realistic but semantically anomalous images.
- **BART (facebook/bart-large-cnn):** Rewrites prompts to increase linguistic creativity.
- **BLIP:** Used for region-level captioning.
- **CLIP ViT-B/32:** Used to compute text-image embedding similarity.
- **Grad-CAM:** Extracts causal regions from classifier predictions.

3.2.5 Formulas Used. **Entropy (Classification Uncertainty):**

$$H(p) = -\sum_{i=1}^{C} p_i \log(p_i + \epsilon), \quad \epsilon = 1 \times 10^{-10}$$

where p_i is the softmax probability for class i and C is the total number of classes.

Cosine Distance (Feature Deviation):

$$d_{\cos} = 1 - \frac{\mathbf{f}_g \cdot \mathbf{f}_{\text{id}}}{\|\mathbf{f}_g\| \|\mathbf{f}_{\text{id}}\|}$$

where $\mathbf{f}_g$ is the feature from the generated image and $\mathbf{f}_{\text{id}}$ is an in-distribution reference.

Perceptual Concept Shift (PCS):

$$\text{PCS} = 1 - \cos\left(\mathbf{v}_{\text{text}}, \mathbf{v}_{\text{image}}\right)$$

Here, $\mathbf{v}_{\text{text}}$ is the CLIP embedding of the BLIP-generated caption for the Grad-CAM region, and $\mathbf{v}_{\text{image}}$ is the CLIP embedding of the full image.

3.3 Structural Disruption

Structural Disruption measures the model's robustness when the part-whole relationships in the image are altered by disrupting the semantic coherency. The idea is to assess the drop in accuracy and confidence of the image classifier models when the image parts are swapped, blacked out, and rearranged in a visually consistent manner. This approach identifies the high-attention segmentation masks via DINO-V2 and SAM and applies perturbations like blacking out the high-attention segmentation mask or swapping two masks (swapping one part with another) without disrupting low-level image statistics.

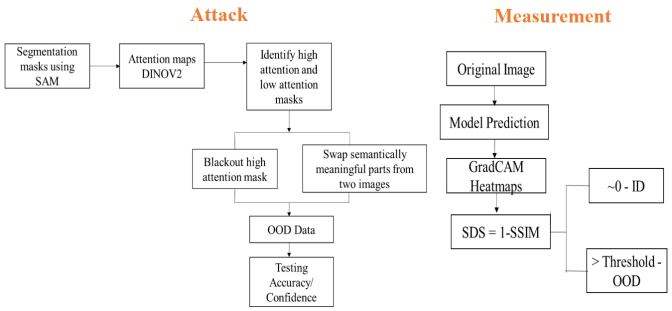

Figure 2: Methodology for Structural Disruption

3.3.1 Experimental Procedure. The procedure for the attack and measurement is as follows:

(1) Load the pre-trained SAM ViT-H model for segmentation, DINOv2-base for attention extraction, and ResNet-18 for classification and Grad-CAM.
(2) Normalize ID dataset images to ImageNet statistics.
(3) For each ID image
   - Apply SAM to obtain segmentation masks for all object parts and extract the attention maps from DINOv2.
   - Overlap the attention maps with the masks of segmentation and choose the high-attention segmentation masks on the IoU threshold (λ).
   - Implement either of the following perturbations on top-k important masks:
     – Blackout: Make the pixels of the selected masks zero without affecting the rest of the image.
     – Part Swapping: Swap one of the high-attention segmentation masks with the other.
   - Pass the perturbed image through the classifier model to obtain the predicted class and softmax score (confidence).
   - Extract the Grad-CAM attention heatmaps for the ID and OOD data.
   - Measure the structural divergence in heatmaps using SDS.
(4) Analyze results: Higher SDS indicate stronger semantic disruption and the OOD data.

3.3.2 Feature Extraction. **Segmentation with SAM:** SAM (`segment-anything/sam_vit_h_4b8939.pth`), is a vision transformer (ViT) model that is capable of producing high-quality object segmentation masks with no need for labeled data. It uses the pre-trained ViT-H backbone for image encoding and segmentation.

Attention with DINOv2: DINOv2 (Self-Distilled Vision Transformer) is a self-supervised learning model that is used to extract semantic representations and attentions in the image data. `facebook/dinov2-base` checkpoint to extract attention maps and map them with the segmentation masks obtained via SAM using the IoU score. We then identify the segmentation masks corresponding to the high-attention regions.

3.3.3 Models and Parameters.

- **ResNet-18 / EfficientNet-B0** (pretrained on ImageNet): models are used for prediction and Grad-CAM extractions. Both models are frozen (`model.eval()` mode).
- **DINOv2**: self-supervised ViT model that is used for extracting attention regions. No fine-tuning is performed
- **SAM (Segment Anything Model)**: pre-trained ViT-H model that is used to generate fine-grained segmentation masks. No fine-tuning is performed.
- **Grad-CAM Setup**: target layers are layer4[-1] for ResNet18 and model.features[-1] for EfficientNet-B0.

Hyperparameters:

- IoU Threshold for selecting high-attention regions: $\lambda = 0.5$
- Number of alterations per image: $N_{\text{perturbations}} = 3$
- Number of segmentation masks to be altered: $k = 5$

3.3.4 Formulas Used. **Structural Divergence Score (SDS):**

$$\text{SDS} = 1 - \text{SSIM}(G_{\text{OOD}}, G_{\text{ID}})$$

where G_{OOD} and G_{ID} are GradCAM maps of OOd and ID images, respectively.

Structural Similarity Index (SSIM):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Here, μ_x and μ_y are the local means of x and y, σ_x^2 and σ_y^2 are their variances, σ_{xy} is the covariance, and C_1, C_2 are stabilization constants. A higher SSIM score indicates stronger spatial alignment, while SDS quantifies deviation from this alignment.

Intersection over Union (IoU):

$$\text{IoU} = \frac{|M_{\text{SAM}} \cap M_{\text{Attention}}|}{|M_{\text{SAM}} \cup M_{\text{Attention}}|}$$

where M_{SAM} is the SAM segmentation mask and $M_{\text{Attention}}$ is the DINOv2-derived attention map.

3.4 Causal Concept Subversion (CCS)

A framework called **Causal Concept Subversion (CCS)** is proposed to evaluate the fragility of a classifier by selectively perturbing causal regions identified via Grad-CAM. The goal is to introduce carefully selected out-of-distribution (OOD) patches into the causally important areas of the input, maximizing both low-level texture difference and high-level semantic drift, and to measure how unstable the model's predictions become under these controlled perturbations.

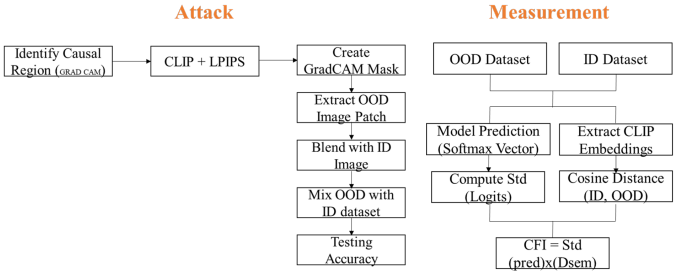

Figure 3: Methodology for CCS

3.4.1 Experimental Procedure. The procedure for the attack and measurement is as follows:

(1) Load ResNet-18 / EfficientNet-B0, CLIP ViT-B/32, and LPIPS AlexNet models.

(2) Normalize ID dataset images to ImageNet statistics and preprocess OOD images appropriately for LPIPS and CLIP feature extraction.
(3) For each ID image:
   - Generate GradCAM heatmaps and threshold to create causal masks.
   - Search for OOD patches maximizing the sum of LPIPS distance and CLIP distance.
   - Blend the selected OOD patch into the causal regions with blending factor α, and classifies the perturbed images using model.
   - Compute semantic shift between original and perturbed images.
   - Aggregate predictions and semantic shifts to compute CFI.
(4) Analyze the results. A higher CFI indicates greater model fragility under causal concept subversion.

3.4.2 Preprocessing. **Out-of-Distribution (OOD) Patches:** OOD images serve as perturbation patch sources. Each OOD image is resized to 224×224 pixels to match the model input size. For the computation of the LPIPS texture distance, these images are further normalized to the interval $[-1, 1]$. Images are fed into CLIP's preprocessing pipeline (resize, center crop as well as normalization) for CLIP semantic feature extraction.

3.4.3 Feature Extraction. **Texture Feature (LPIPS Distance):** A pre-trained LPIPS model with an AlexNet backbone is used to compute perceptual similarity between the original image and OOD candidate patches. A larger LPIPS distance reflects a more significant low level texture difference useful in distinguishing visually different perturbation candidates.

Semantic Feature (CLIP Embeddings): CLIP ViT-B/32 is used to achieve semantic embeddings of both the original images and OOD patches. Semantic dissimilarity between them is calculated as the cosine distance between their corresponding embeddings.

3.4.4 Models and Parameters.

- **ResNet-18 / EfficientNet-B0** (pretrained on ImageNet): used for predictions and Grad-CAM extraction; both models are frozen (`model.eval()` mode).
- **CLIP ViT-B/32**: used for semantic feature extraction; used without fine-tuning.
- **LPIPS Model** (AlexNet backbone): used for perceptual texture distance measurement.
- **Grad-CAM Setup:** target layers are `layer4[-1]` for ResNet-18 and `model.features[-1]` for EfficientNet-B0.

Hyperparameters: Number of perturbations per image: $N_{\text{variants}} = 5$ and Blending factor for patch and causal region: $\alpha = 0.75$.

3.4.5 Formulas Used. **Texture Dissimilarity (LPIPS Distance):**

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \|F_l(x) - F_l(y)\|_2^2$$

where $F_l(\cdot)$ are feature activations at layer l of a network.

Semantic Dissimilarity (CLIP Distance):

$$\text{CLIP\_Dist}(x, y) = 1 - \cos(\theta) = 1 - \frac{x \cdot y}{\|x\|\|y\|}$$

where x and y are CLIP feature embeddings.

Total Patch Selection Score:

$$\text{Total Score} = \text{LPIPS Distance} + \text{CLIP Distance}$$

The patch with the highest total score is selected.

Conceptual Fragility Index (CFI):

$$\text{CFI} = \sigma(\text{Predictions}) \times \mu(\text{Semantic Shifts})$$

where:

- σ(Predictions) is the mean standard deviation of prediction probabilities across all perturbed variants,
- μ(Semantic Shifts) is the average semantic shift (cosine distance) across variants.

3.5 Validation

3.5.1 Success Criteria. Success is defined in terms of how well our semantic, structural, and causal out-of-distribution (OOD) attacks degrade model performance and reveal vulnerabilities that are not already captured by traditional robustness benchmarks. Specifically, a successful attack yields lower classification accuracy, decreased prediction confidence, and model attention changes. Our proposed metrics Perceptual Concept Shift (PCS), Structural Divergence Score (SDS) and Conceptual Fragility Index (CFI) should effectively quantify semantic, structural and causal misalignments, respectively.

3.5.2 Validation Strategy. In evaluating pretrained vision models, we adopt a non-traditional approach that skips the standard train/validation/test splits. Our evaluation protocol involves several key steps: First we perform **Ground Truth Evaluation** by comparing model predictions on standard datasets with true labels, while ground truth labels or captions for generated OOD samples are manually curated. Next we analyze **Confidence and Attention** by examining prediction confidence and Grad-CAM attention maps to understand the model's internal response to OOD interventions. **Metric-Based Evaluation** follows where OOD attacks are assessed using the Perceptual Concept Shift (PCS) for semantic discrepancy, Structural Divergence Score (SDS) for structural shifts and Conceptual Fragility Index (CFI) for causal vulnerabilities. Finally we forgo a traditional validation set, focusing solely on testing the model's behavior under OOD stress to gauge its robustness.

3.5.3 Hypotheses.

- **H1:** *High-level semantic, structural and causal manipulations can systematically degrade the performance and interpretability of state of the art vision models exposing failure modes beyond conventional benchmarks.*
- **H2:** *Models will show a significant drop in prediction confidence when tested on our generated OOD dataset compared to in-distribution samples.*
- **H3:** *Each OOD attack will produce a measurable increase in its corresponding metric PCS, SDS or CFI revealing distinct facets of model brittleness.*

4 Results

4.1 Results of Semantic Confusion Attack

This section presents a comparative analysis of the performance of ResNet18 and EfficientNet-B0 on semantically perturbed and out-of-distribution (OOD) images. We compare the predictions of the models in three representative scenarios to test their susceptibility to texture bias, their handling of semantically perturbed but in-distribution images and their response to OOD inputs. Results are visualized in Figures 1–3 and are interpreted alongside key metrics including confidence scores and prediction entropy.

Figure 4: ResNet18 and EfficientNet-B0 on ImageNet

In Figure 4, we show a zebra-print sofa that is designed to test the model's shape dependence over texture. ResNet18 classifies it as a zebra with very high confidence (>98%) and low entropy indicating high confidence in its incorrect prediction as is characteristic of its widely documented texture bias. EfficientNet-B0 classifies it with low confidence (<30%) and high entropy and indicates greater uncertainty. While it still misclassifies the image, its cautious behavior highlights its reduced susceptibility to confident errors, making it more suitable for detecting potential OOD inputs and improving safety in critical applications.

Figure 5: ResNet18 and EfficientNet-B0 on ImageNet

Figures 5 show a mustache-print coffee mug. ResNet18 correctly classifies mug with >99% confidence exhibiting robustness to minor semantic changes when shape is preserved. EfficientNet-B0 fails to recognize the object and provides a low confidence, high entropy prediction, being conservative but less accurate in in-distribution tasks.

Figure 6: ResNet18 and EfficientNet-B0 on Intel Classification Dataset

In Figure 6, a forest painted van from the Intel Scene Classification dataset is misclassified as a glacier by both the models. ResNet18 again has low entropy and high confidence while EfficientNet-B0 once again predicts with high entropy and low confidence. This demonstrates the overconfidence of ResNet in OOD settings, while EfficientNet expresses uncertainty.

Overall ResNet18 is more accurate in-distribution but overconfident on ambiguous inputs. EfficientNet-B0 is more conservative and therefore better suited for applications where uncertainty estimation is important.

```
--------------------------------------------------
Image: /kaggle/input/testset/generated_image_17.png
Caption: a pizza with pepperoni and pepperoni on it
PCS Score: 0.30719316005706787
ResNet Prediction Class: 963 (Confidence: 1.00)
Status: Likely OOD
--------------------------------------------------
Image: /kaggle/input/testset/generated_image_18.png
Caption: a blue dog bed on the floor
PCS Score: 0.3557380437850952
ResNet Prediction Class: 434 (Confidence: 0.30)
Status: Likely In-Distribution
--------------------------------------------------
```

Figure 7: PCS Score that flags Likeliness

The figure 7 shows PCS (Perceptual Concept Shift) scores and ResNet classification results for two images. The first, captioned "a pizza with pepperoni and pepperoni on it," has a low PCS score (0.307) but a high-confidence prediction (Class 963, Confidence: 1.00), indicating a semantic mismatch and likely Out-of-Distribution (OOD) status. The second image with caption "a blue dog bed on the floor" also has a higher PCS score of 0.356 and lower confidence (Class 434, Confidence: 0.30), showing closer match between image and caption and is Likely In-Distribution. This again highlights PCS's capability to evaluate semantic grounding over model confidence.

4.2 Results of Structural Disruption

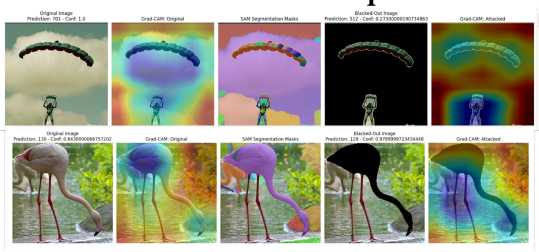

Figure 8: Blackout attack on ResNet18 and EfficientNetB0

In Figure 8, the picture of a flamingo is subjected to the blackout attack. The high attention segmentation mask, i.e., the body of the flamingo, was blacked out. ResNet-18 model predicted the blacked out image as an Ostrich, that too with very high confidence (97%). Figure 8 shows the same blackout attack performed on the EfficientNetB0 model, where the parachute was classified as a cork screw. These results highlight that blacking out high-attention regions can drastically mislead models into confident misclassifications.

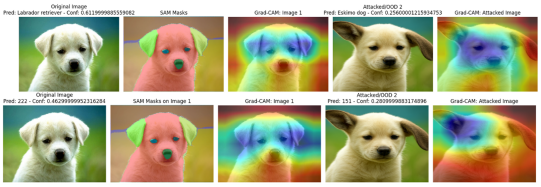

Figure 9: Output of ResNet18 and EffcientNetB0 after Swapping segmentation masks

```
Structural Divergence Score (SDS): 0.1209
Image 1 Top-5: [222 208 257 207 852]
Image 1 Scores: [np.float32(0.832), np.float32(0.091), np.float32(0.029), np.float32(0.007), np.float32(0.007)]
Blackout Top-5: [610 879 649 399 515]
Blackout Scores: [np.float32(0.463), np.float32(0.062), np.float32(0.062), np.float32(0.048), np.float32(0.047)]
```

Figure 10: SDS score after applying shuffling attack

Figure 9 demonstrate the models' performance when the dog's ears (high-attention regions according to DINOv2) were swapped with the kangaroo's ears. Figure 9 is the output of the ResNet18 model. It predicted the attacked image as a different dog breed, and the model's confidence in the prediction has reduced significantly (46% to 28%). The SDS score obtained is 0.2109, indicating high structural divergence. The same was the case with the EfficientNetB0 model. These results indicate that swapping semantically important parts can result in wrong predictions and confidence drops, indicating the model's sensitivity to part-whole relationships in the image.

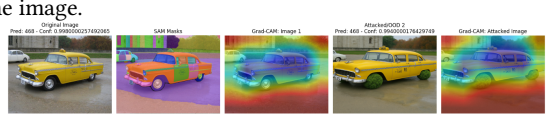

Figure 11: Swapping masks obtained via DINOv2

In Figure 11, we tried to swap the wheels of the car with the patch of bushes. According to DINOv2 attention maps, car wheels are the high-attention regions. But, from the Grad-CAM heatmaps, it is evident that wheels are not of high attention for the ResNet18 model. This change had no effect on the model's prediction and accuracy. SDS score obtained is 0.0008, indicating very low divergence. This shows that just because a self-supervised model like DINOv2 identified wheels as high-attention regions doesn't mean that the classification models depend on those regions for decision making.

Overall, the attack was effective on both ResNet18 and EfficientNetB0 models. There was a drop in the accuracy and confidence of the models. The identified limitation was that DINOv2 attention is not necessarily aligned with discriminative regions utilized by classification models such as ResNet18 and EfficientNetB0. This misalignment will lead to ineffective perturbations in some cases,

4.3 Results of Causal Concept Subversion

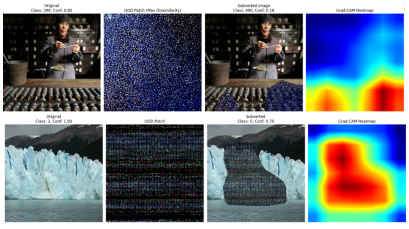

Figure 12: Output of ResNet18 on ImageNet and Intel Classification Dataset

For the majority of cases, combining an out-of-distribution (OOD) patch with an in-distribution image effectively decreases the model's prediction confidence demonstrating the intended impact of our adversarial method. For instance, in the figure 11, we show a person using an abacus. ResNet-18 initially classifies the image with high confidence (0.85), but once an OOD patch is blended into the causal region, confidence drops sharply to 0.18, showing successful disruption of the model's internal reasoning. And same with the image from Intel Classification.

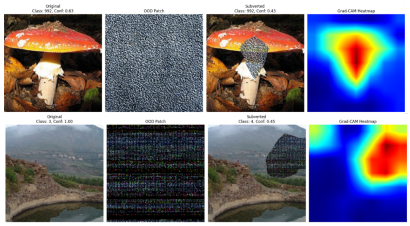

Figure 13: Output of EfficientNet-B0 on ImageNet Dataset

In Figures 12, we observe the same tendencies but EfficientNet-B0 is stronger. Its predictions are less confident overall but it also predicts classes correctly, and the entropy is greater even after patch insertion, which demonstrates a better calibrated uncertainty response. This suggests EfficientNet is better at signaling ambiguity when exposed to OOD content, making it more suitable for safety-critical applications.

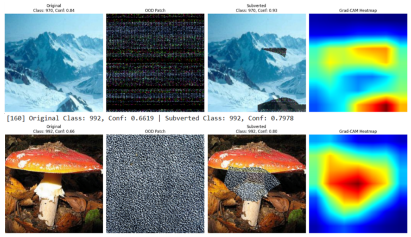

Figure 14: Output of ResNet18 on ImageNet Dataset and EfficientNet-B0 on Intel Classification Dataset

However, there are certain cases that are counterintuitive. In the second image of Figure 13 with mushroom and landscape introducing semantically irrelevant patches unexpectedly increased model confidence. Such unexpected behavior reflects a vulnerability: models like ResNet18 and EfficientNet-B0 can spuriously become confident when patch textures incidentally match learned local features.

This highlights a broader issue that such models heavily rely on local texture information rather than global scene understanding. Even absurd patches can support existing evidence if their surface statistics align with class priors. While EfficientNet is generally more conservative, it also shows overconfidence under such conditions pointing towards the vulnerability of today's CNNs to localized, texture based OOD attacks and the need for models with stronger global reasoning capabilities.

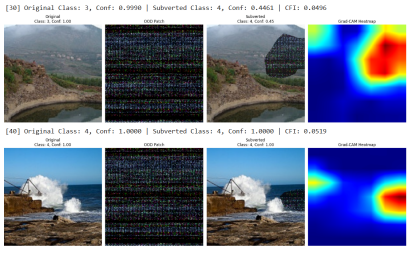

Figure 15: CFI Score for ResNet18 on Intel Classification Dataset

The Causal Fragility Index or CFI captures how quickly the prediction of a model changes upon adding a small, localized patch to an image. Low CFI in the case of the examples (0.0496 and 0.0519) suggests the decision of the model is quite stable and not dependent much on the patch. This makes CFI a useful metric for assessing the robustness and causal reliability of visual models under perturbations. This shows that ResNet18 is quite stable from casual perturbations in this dataset.

5 Discussions

The overall goal of this work was to evaluate the performance of two of the most common convolutional neural network (CNN) architectures, ResNet-18 and EfficientNet-B0, on semantic, structural, and causal distortions in image input. We had three working hypotheses regarding the robustness and calibration of the two models. The results of our experiments gave strong validation for all of them, leading us to an in-depth understanding of the functioning of both architectures.

Semantic, structural, and causal manipulations had significantly degraded the models' performance. In the case of causal subversion, ResNet-18 repeatedly made confident predictions even on distorted images with irrelevant textures or deceptive patches, displaying overconfidence that would be dangerous in real-world applications. EfficientNet-B0, though it had misclassifications, performed much better in signaling uncertainty through higher entropy values and lower prediction confidence scores. In the case of semantic disruption, ResNet-18 is more accurate in-distribution but overconfident on ambiguous inputs, whereas EfficientNet-B0 is more conservative. In the case of structural disruption, both ResNet-18 and EfficientNet-B0 showed degraded performance, i.e., both misclassified the images in most cases. So, this supports the initial two hypotheses that these attacks systematically degrade the performance of the state-of-the-art vision models and reduce the prediction accuracy and confidence.

To measure the effectiveness of each of these attacks, we proposed three different metrics - PCS, SDS, and CFI. In the case of semantic disruption, the OOD samples had higher PCS values compared to the original data, indicating more semantic misalignment. The same is the case with structural disruption; higher SDS values indicated more dissimilarity between ID and OOD images. For causal subversion, higher CFI indicated that the model is more fragile to disturbances in ID image. This supports our final hypothesis that there is an increase in each of these metrics for the generated OOD data.

One of the key limitations of the current work is that only image-based models such as ResNet-18, BLIP, and DINOv2 were used for the experimentation. So, it is not possible to generalize the results over other modalities or model architectures, such as transformers used for processing multi-modal or sequential inputs. Therefore, the effectiveness of the proposed method to generate OOD data in real-world environments with diverse inputs and tasks is not evaluated. The reliance on manual parameter tuning may hinder the automation in the attacking pipeline. In addition, the introduction of synthetic artifacts in creating the perturbations can bound realism. The applicability of such human-crafted distortions, while effective in controlled settings, may fail to capture the complexity or randomness of natural distributional shifts, which can impact the robustness and ecological validity of the evaluation.

Future work should focus on automating the attack pipeline through generative models or reinforcement learning methods for adaptively producing attacks. Also, employing the framework for cross-modal system testing and validating it against real-world OOD scenarios, for instance, domain shifts in medical imaging or autonomous driving, would enhance its applicability and impact. Lastly, having human-in-the-loop feedback for attaining increased realism in perturbations and longitudinal studies in retraining resilience would provide insights for an increased comprehension of model robustness in ongoing deployment.

6 Conclusion

In this work, we assessed the robustness of ResNet18 and EfficientNet-B0 models using a three-pronged attack framework, targeting the structure, meaning and cause-effect relationships of images. We proposed a Semantic Confusion Attack, generating misleading images with contradictory captions and evaluated semantic coherence with the Perceptual Concept Shift (PCS) metric, which measures misalignment between Grad-CAM attention maps and image-caption consistency. The Structural Disruption attack aimed to disturb part whole relationships by swapping high-attention segmentation masks, with the Structural Disruption Score (SDS) quantifying the difference in attention maps. Lastly we identified critical causal regions with GradCAM and replaced them with contradictory patches, evaluating the impact using the Conceptual Fragility Index (CFI) that tracks model prediction instability. These attacks were evaluated on three types of OOD test data, comparing the metrics (PCS, SDS, CFI) to in-distribution (ID) data.

Our findings aligned well with our expectations, yet there were some interesting findings. The Semantic Confusion attack revealed ResNet18's overconfidence due to texture bias, while EfficientNet-B0, though less accurate in-distribution, demonstrated higher uncertainty on ambiguous inputs, indicating better calibration. Structural disruption attacks in the form of blacking out or swapping out high-attention parts lowered both models' prediction accuracy significantly. But, the high-attention regions identified by DINOv2 did not align with those used by the classification model in some cases, limiting the effectiveness of the attack. Causal Concept Subversion attack successfully reduced the confidence of the models significantly by perturbing causally relevant areas. Interestingly, there were surprising spikes of confidence in some cases that revealed models' heavy dependence on local texture features instead of global scene understanding. These results emphasize the development of more robust vision models that are not only accurate with ID data but also resilient to these kinds of attacks. This work contributes to the development of more robust AI models that are reliable in adversarial and ambiguous situations, which are critical in applications like autonomous driving, healthcare, and security.

References

[1] Usman Ahmed, Jerry C. Lin, and Gautam Srivastava. 2022. Mitigating adversarial evasion attacks by deep active learning for medical image classification. *Multimedia tools and applications* 81, 29 (2022), 41899–41910.

[2] Xin Bai, Jie Ren, and Fengzhu Sun. 2022. MLR-OOD: A Markov Chain Based Likelihood Ratio Method for Out-Of-Distribution Detection of Genomic Sequences. *Journal of molecular biology* 434, 15 (2022), 167586–167586.

[3] Ning Cao, Guofu Li, Pengjia Zhu, Qian Sun, Yingying Wang, Jing Li, Maoling Yan, and Yongbin Zhao. 2019. Handling the adversarial attacks: A machine learning's perspective. *Journal of ambient intelligence and humanized computing* 10, 8 (2019), 2929–2943.

[4] Joana C. Costa, Tiago Roxo, Hugo Proenca, and Pedro R. M. Inacio. 2024. How Deep Learning Sees the World: A Survey on Adversarial Attacks Defenses. *IEEE access* 12 (2024), 1–1.

[5] Claudio Ferrari, Federico Becattini, Leonardo Galteri, and Alberto D. Bimbo. 2023. (Compress and Restore)N: A Robust Defense Against Adversarial Attacks on Image Classification. *ACM transactions on multimedia computing communications and applications* 19, 1s (2023), 1–16.

[6] Kyungpil Gwon and Joonhyuk Yoo. 2023. Out-of-Distribution (OOD) Detection and Generalization Improved by Augmenting Adversarial Mixup Samples. *Electronics (Basel)* 12, 6 (2023), 1421.

[7] Sheeba Lal, Saeed U. Rehman, Jamal H. Shah, Talha Meraj, Hafiz T. Rauf, Robertas Damaševičius, Mazin A. Mohammed, and Karrar H. Abdulkareem. 2021. Adversarial Attack and Defence through Adversarial Training and Feature Fusion for Diabetic Retinopathy Recognition. *Sensors (Basel, Switzerland)* 21, 11 (2021), 3922.

[8] Benjamin Lambert, Florence Forbes, Senan Doyle, and Michel Dojat. 2023. *Multi-layer Aggregation as a Key to Feature-Based OOD Detection.* Springer Nature Switzerland, Cham, 104–114.

[9] Feng Li, Xuehui Du, and Liu Zhang. 2022. Adversarial Attacks Defense Method Based on Multiple Filtering and Image Rotation. *Discrete dynamics in nature and society* 2022, 1 (2022).

[10] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2021. Energy-based Out-of-distribution Detection. (2021).

[11] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. 2022. *Frequency Domain Model Augmentation for Adversarial Attack.* Vol. 13664. Springer, Switzerland, 549–566.

[12] Mohamed Ouriha, Youssef El Habouz, and Omar El Mansouri. 2022. *Decision Boundary to Improve the Sensitivity of Deep Neural Networks Models.* Vol. 449. Springer International Publishing AG, Switzerland, 50–60.

[13] Utku Ozbulak, Manvel Gasparyan, Wesley De Neve, and Arnout V. Messem. 2020. *Perturbation Analysis of Gradient-based Adversarial Attacks.* Technical Report. Cornell University Library, arXiv.org.

[14] Qifei Zhou, Rong Zhang, Bo Wu, Weiping Li, and Tong Mo. 2020. *Detection by Attack: Detecting Adversarial Samples by Undercover Attack.* Vol. 12309. Springer International Publishing AG, Switzerland, 146–164.