The Invisible Gorilla Effect in Out-of-distribution Detection

Harry Anthony, Ziyun Liang, Hermione Warr, Konstantinos Kamnitsas
Department of Engineering Science, University of Oxford, Oxford, UK.
Harrya@live.co.uk

CVPR JUNE 3-7, 2026 DENVER COLORADO

1. Introduction

Out-of-distribution (OOD) data differs from model's training data, causing unreliable predictions.

OOD detection performance varies across artefacts, but existing explanations are limited.

We identify a new bias in OOD detection.

Paper Available

Code & Data Available

Original Gorilla Study

Key Findings

OOD detection improves when OOD artefacts are similar to the model's ROI.

Feature-based methods are most impacted by bias.

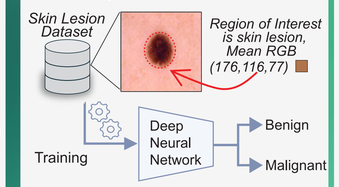
OOD colour sensitive directions correlate with high variance directions in model's latent space.

Projecting out nuisance subspace in model's latent space provides more consistent mitigation than colour jitter augmentation.

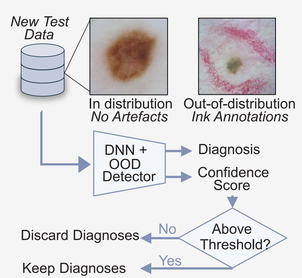
2. What is the Invisible Gorilla Effect

Train Model for Primary Task (e.g. Image Classification)

Training Data with No Artefacts

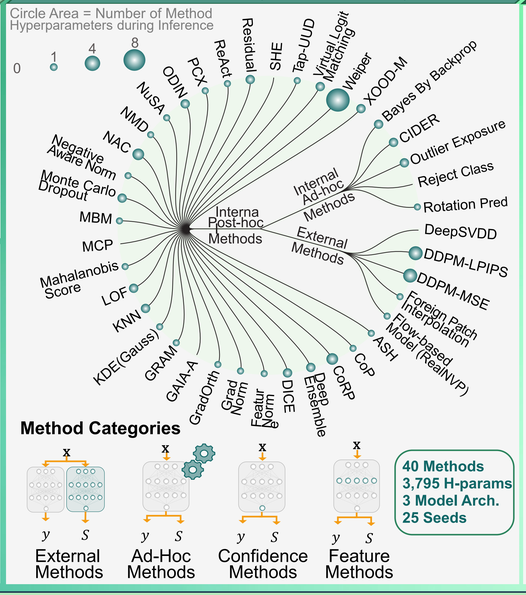


Evaluate Model on Held-out Data

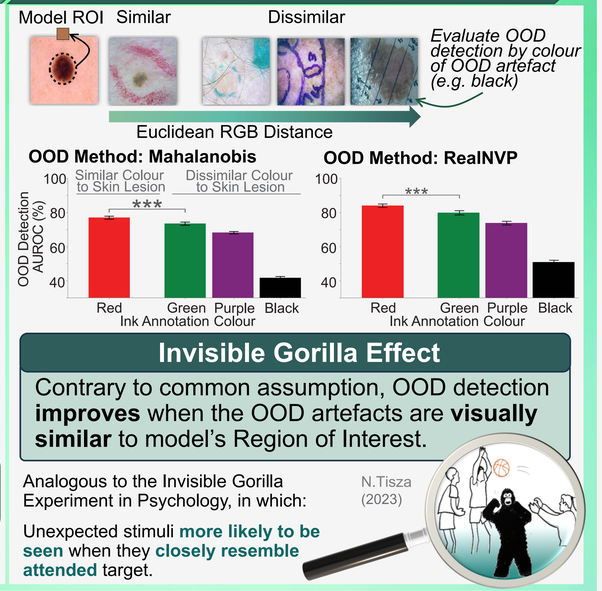


Evaluation Metrics:
AUROC, AURC, FPR@TPR80

Apply Out-of-distribution Detection Methods on Held-out Data (Code Available)



Evaluate Results for the Invisible Gorilla Effect



Invisible Gorilla Effect

Contrary to common assumption, OOD detection **improves** when the OOD artefacts are **visually similar** to model's Region of Interest.

Analogous to the Invisible Gorilla Experiment in Psychology, in which:

Unexpected stimuli **more likely to be seen** when they **closely resemble attended** target.

N.Tisza (2023)

3. New Annotated Benchmarks (Made Publicly available!)

Dataset Summary

ISIC (skin lesion)
ID: Planar images with no ink, rulers or colour charts.
Primary Task: Benign vs. Malignant classification

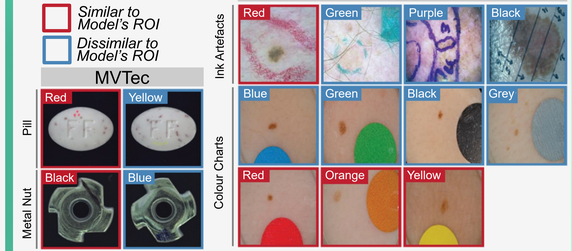
MVTec-AD (Industrial)
ID: Industrial equipment with and without defects.
Primary Task: Default Classification.

Operationalising Similarity
Segment ROI and artefacts. Define similarity as Euclidean RGB.

Meta SAM 2 +

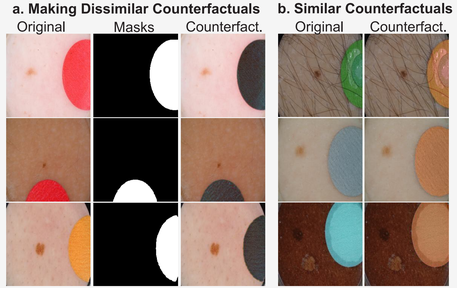
Summary of Manual Annotations on Real Data

11,345 Manually Annotated Images being made public.



Summary of Colour-swapped Counterfactuals

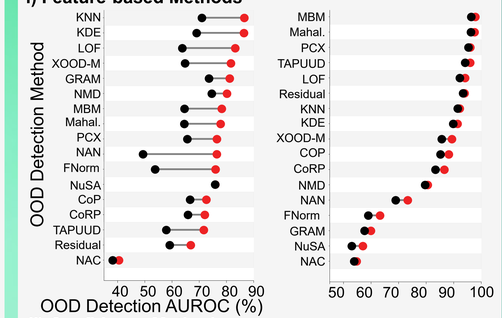
Repeat study on counterfactuals to rule out dataset bias



4. Large-Scale Evaluation

Ink Artefacts
Similar to Model's ROI: Red
Dissimilar to Model's ROI: Green, Black, Purple

Colour Chart Artefacts
Similar to Model's ROI: Red, Orange, Yellow
Dissimilar to Model's ROI: Blue, Green, Black, Grey



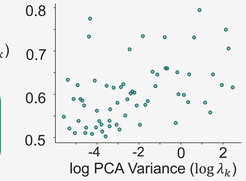
5. Hypothesised Mechanism

Colour variation aligns with **high-variance directions** in the model's latent space.

Feature-based methods typically downweigh high variance directions.

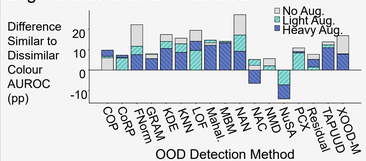
Empirical Evidence

1. PCA on training data to identify eigenvectors (v_k) and corresponding variances (λ_k).
2. Project Similar and dissimilar OOD onto Principal Components.
3. Compute ability to discriminate between similar and dissimilar OOD
$I_k = \max(AUC(Z_{Diss.}v_k, Z_{sim.}v_k), 1 - AUC(\cdot))$
4. Measure correlation between I_k with high-variance directions (λ_k)?



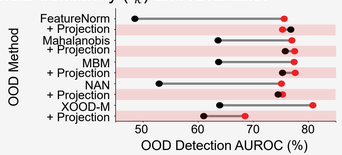
6. Mitigation Strategies

Confidence Jitter Augmentations
Trained models with colour jitter augmentations and revaluated.



→ **Inconsistent in reducing the effect.**

Nuisance subspace projection
Project out 5 directions with the highest colour sensitivity (I_k) and revaluate.



→ **More consistent in reducing the effect.**