CVPR JUNE 3-7, 2026 DENVER COLORADO

KAMNITSAS LAB FOR AI IN MEDICINE

IBME INSTITUTE OF BIOMEDICAL ENGINEERING

DEPARTMENT OF ENGINEERING SCIENCE

UNIVERSITY OF OXFORD

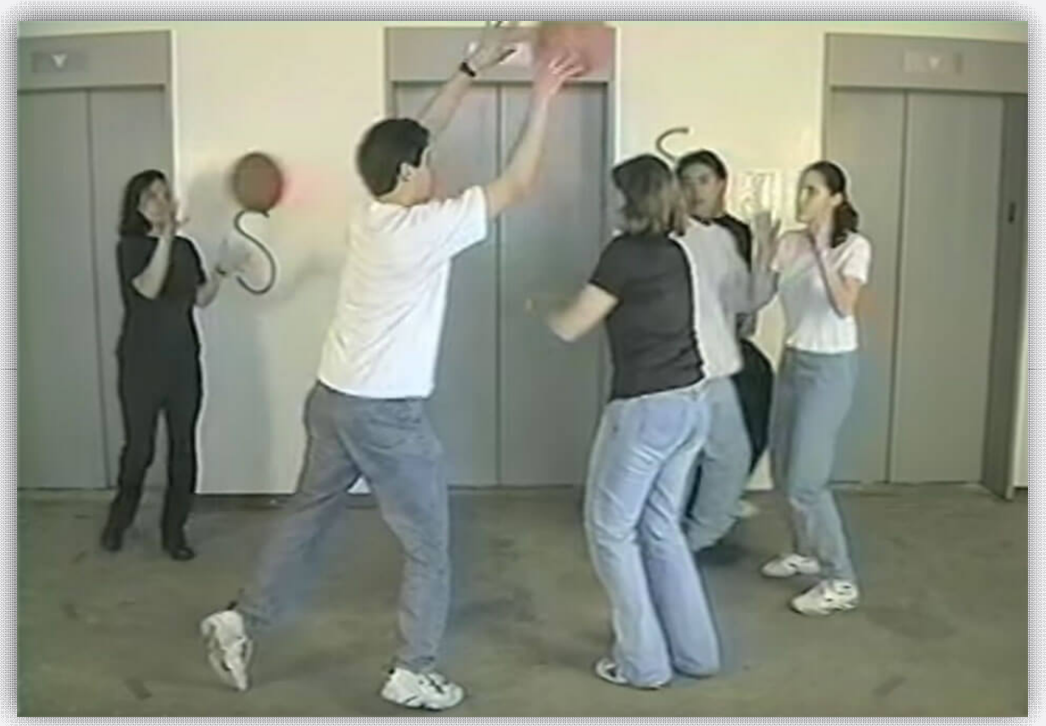
The Invisible Gorilla Effect In Out-of-distribution Detection

Harry Anthony, Ziyun Liang, Hermione Warr, Konstantinos Kamnitsas

CVPR Highlight 2026

Let’s begin with an exercise…

Count the number of basketball passes between
players in white shirts and ignore passes between in black shirts.
https://www.youtube.com/watch?v=vJG698U2Mvo

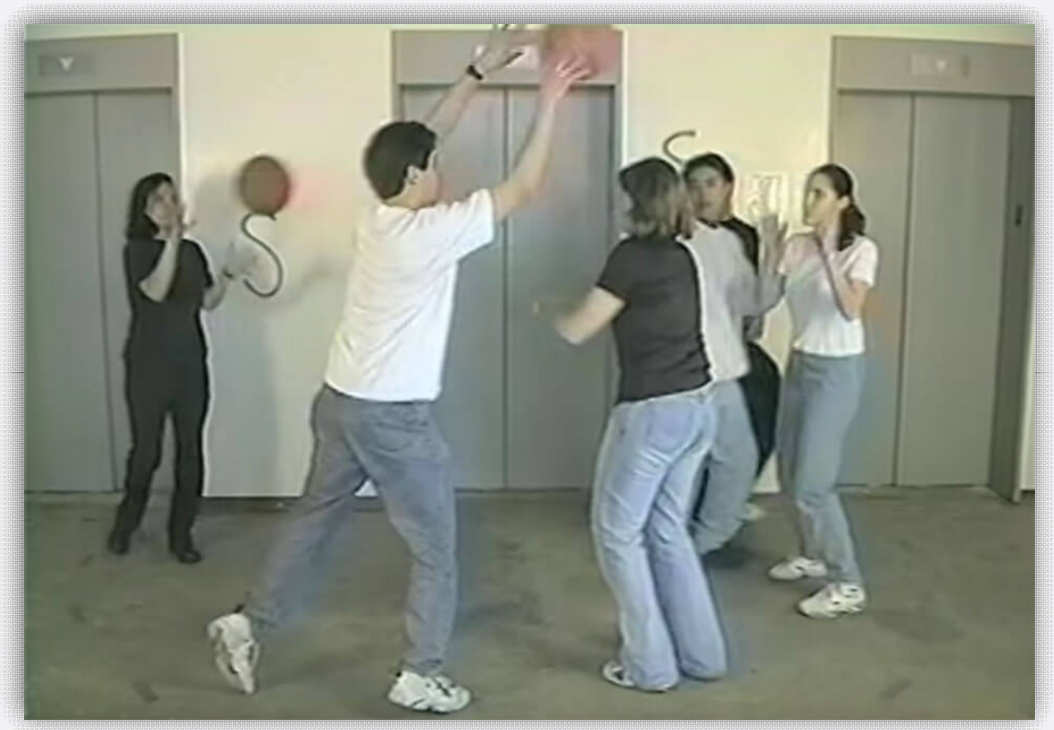


D.Simons (1999)

How many did you count ?

Correct answer is 15

But did you spot the person in a **Gorilla Costume**?



D.Simons (1999)

Invisible Gorilla Experiment

In Psychology

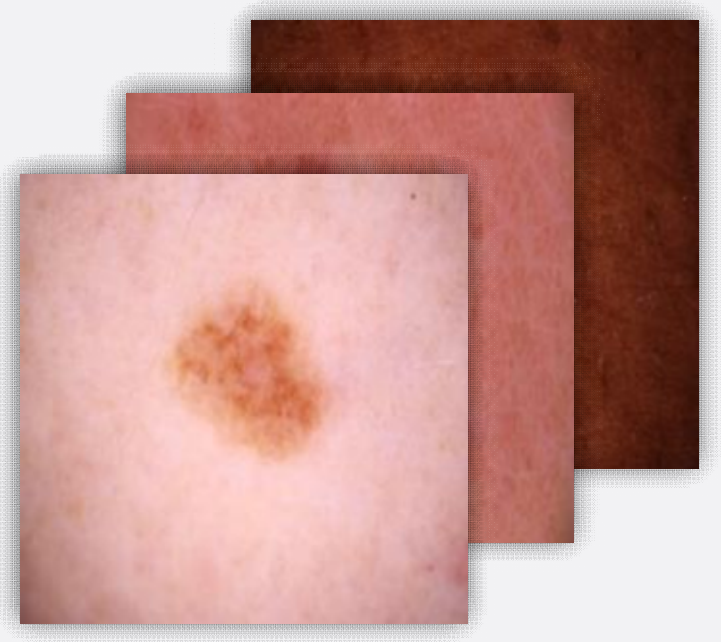
N.Tisza (2023)

Inattentive Blindness

Individuals focusing on a primary task can ignore unexpected stimuli due to a lack of attention.

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

Out-of-distribution detection



$\boldsymbol{x}$

Benign

Malignant

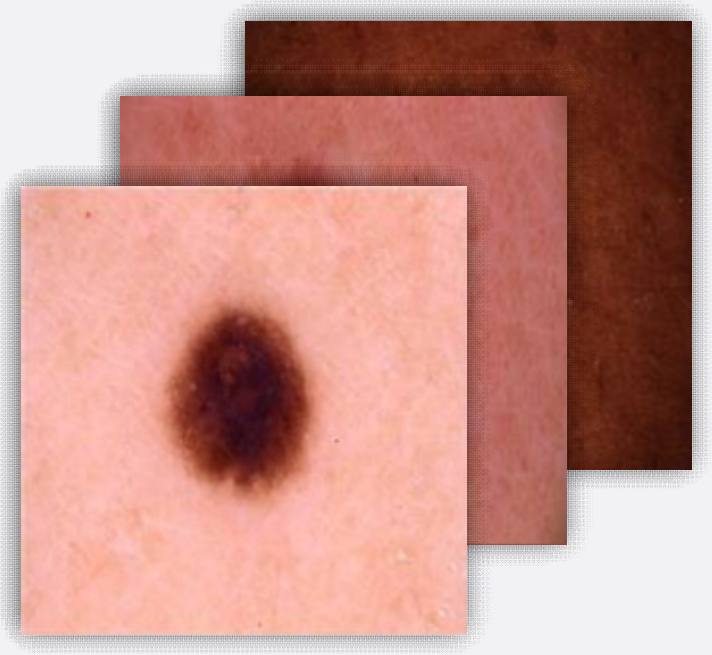
Benign

y

Training Data

$\mathcal{D}_{train} = \{(\boldsymbol{x}_n, y_n)\}_{i=1}^{N}$

Out-of-distribution detection



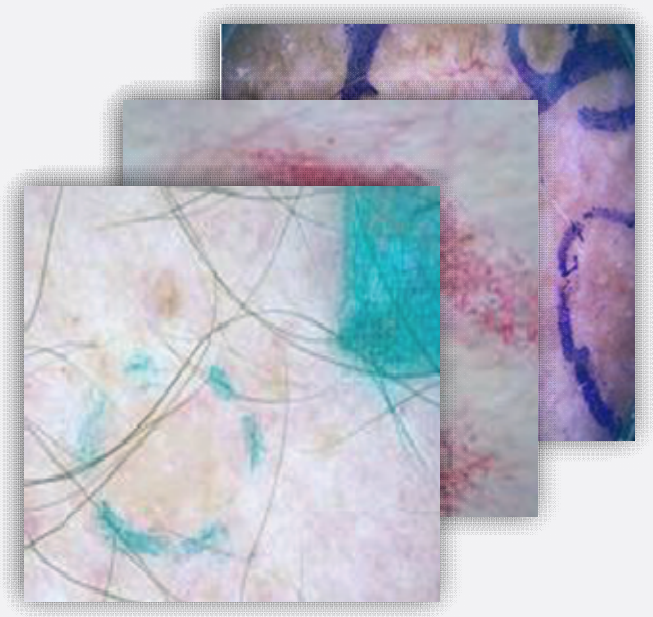
$(\boldsymbol{x}, y) \sim \mathcal{D}_{train}$
In-distribution
(ID)

Covariate
Shift

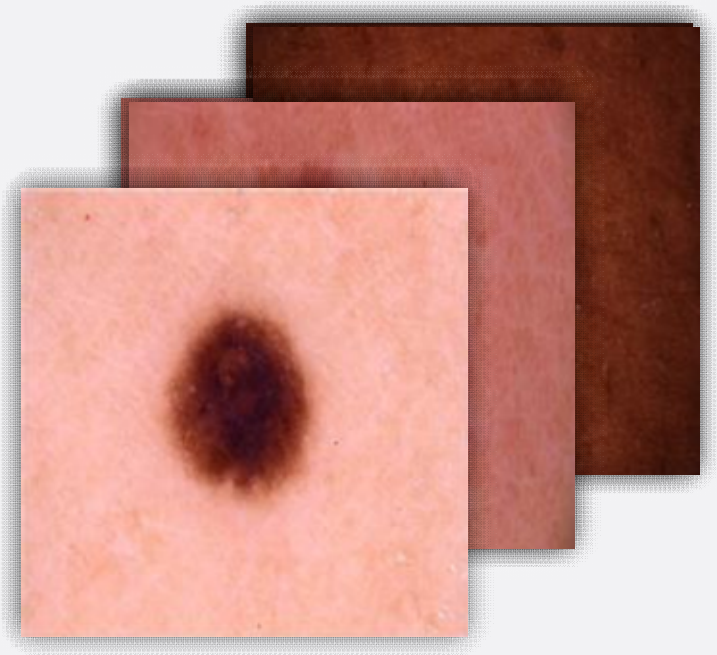
Same Classes

$p_{ID}(y|x) = p_{OOD}(y|x)$

$p_{ID}(x) \neq p_{OOD}(x)$

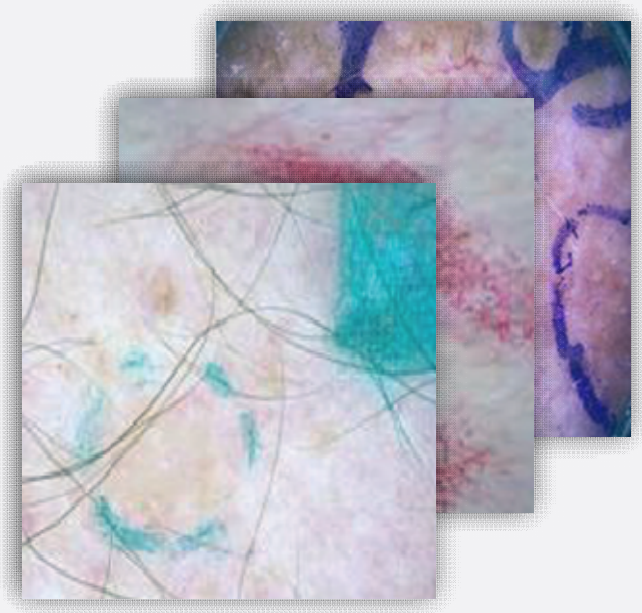


$(\boldsymbol{x}, y) \nsim \mathcal{D}_{train}$
Out-of-distribution
(OOD)



$(x, y) \sim \mathcal{D}_{train}$

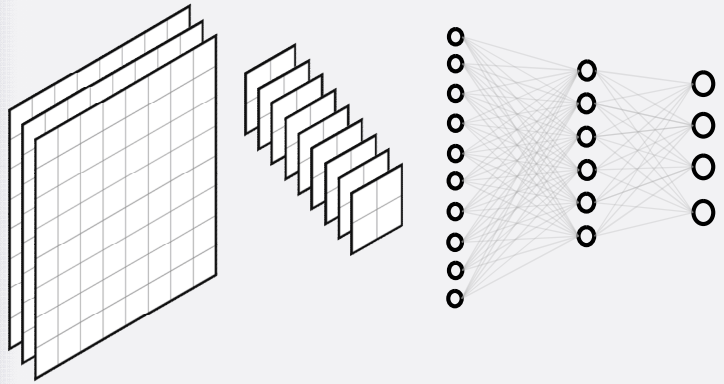
In-distribution



$(x, y) \nsim \mathcal{D}_{train}$

Out-of-distribution

ResNet18



90.8%

Diagnostic Accuracy

74.3%

Diagnostic Accuracy

$(x, y) \sim \mathcal{D}_{train}$

In-distribution

90.8%

Diagnostic Accuracy

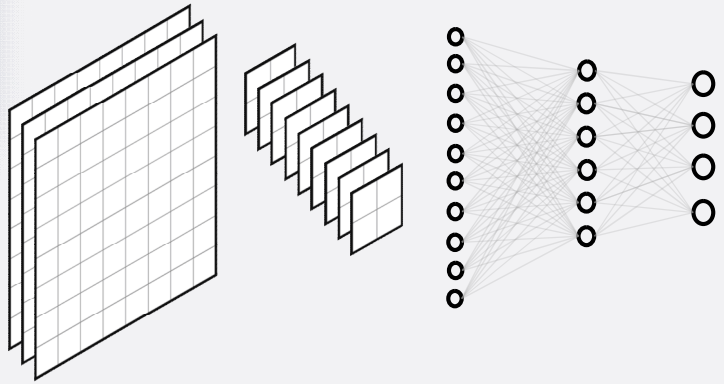
$(x, y) \nsim \mathcal{D}_{train}$

Out-of-distribution

74.3%

Diagnostic Accuracy

ResNet18



Motivates Out-of-distribution detection methods

Aim of flagging diagnoses on OOD data as unreliable

Finding biases in

Out-of-distribution detection

OOD detection methods that perform well in detecting one OOD artefact can perform poorly in another.

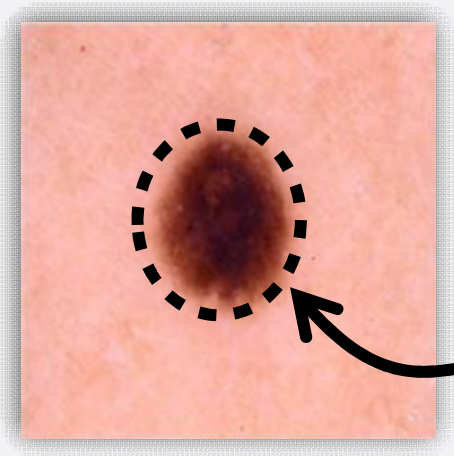
Very limited understanding of what causes these differences.

We introduce an **unreported bias in OOD detection**.

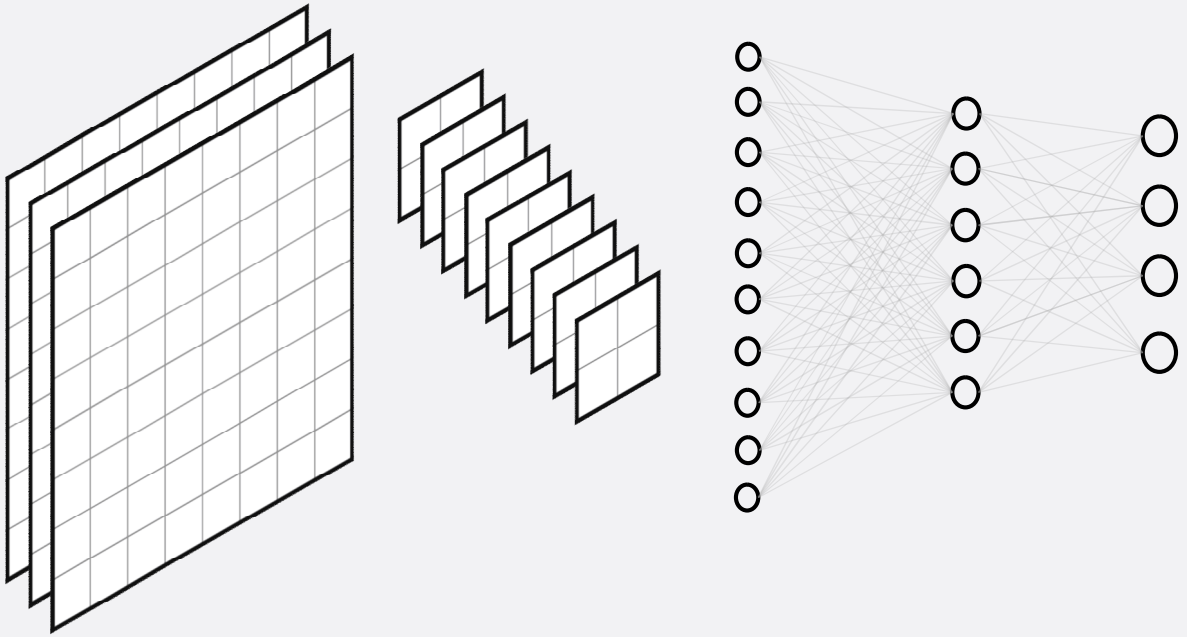
Finding biases in

Out-of-distribution detection

Model's Region of Interest (ROI)

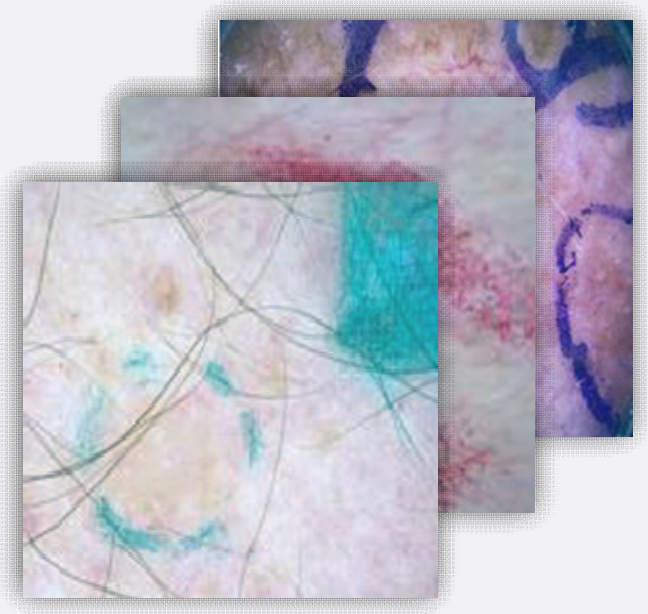


Mean RGB
(176,116,77)



Finding biases in

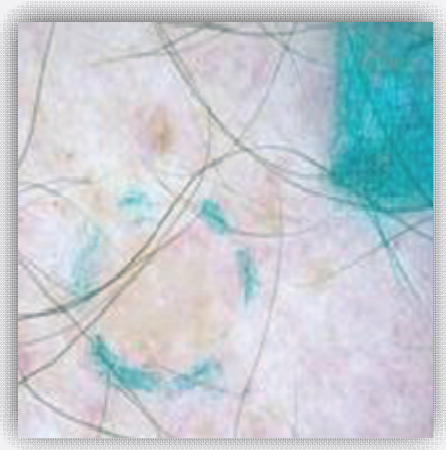
Out-of-distribution detection



$(x, y) \nsim \mathcal{D}_{train}$

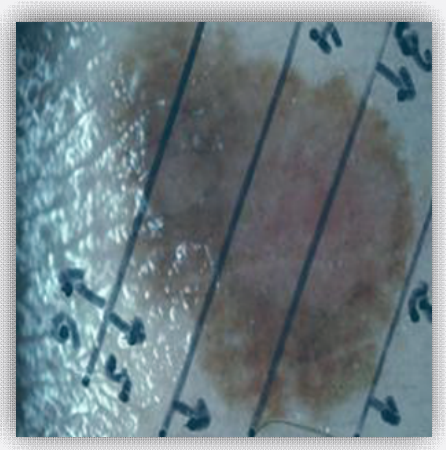
Out-of-distribution

Red



Green

Purple



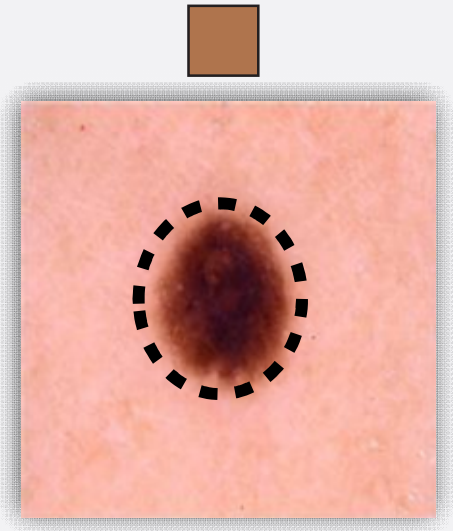
Black

@HarryEJAnthony

Finding biases in

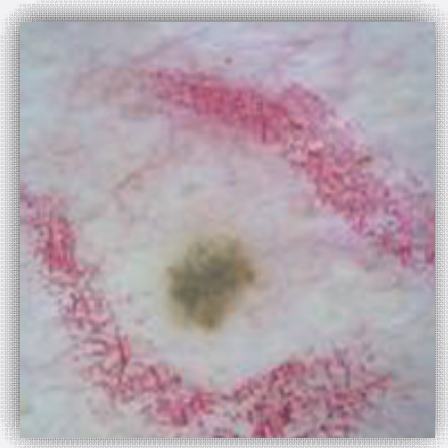
Out-of-distribution detection

Model's ROI
(176,116,77)

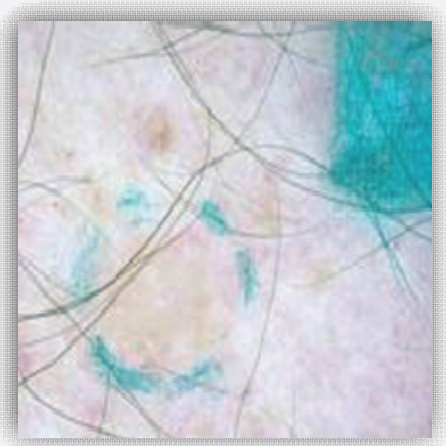


Similar to ROI

Dissimilar to ROI

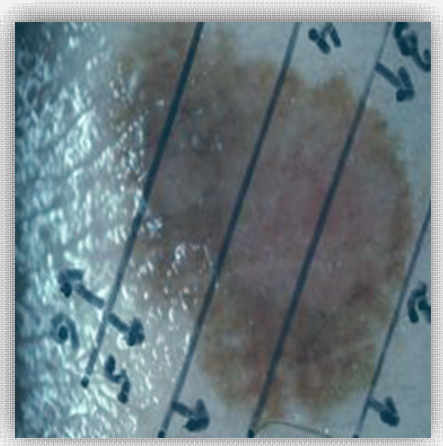


Red



Green

Purple

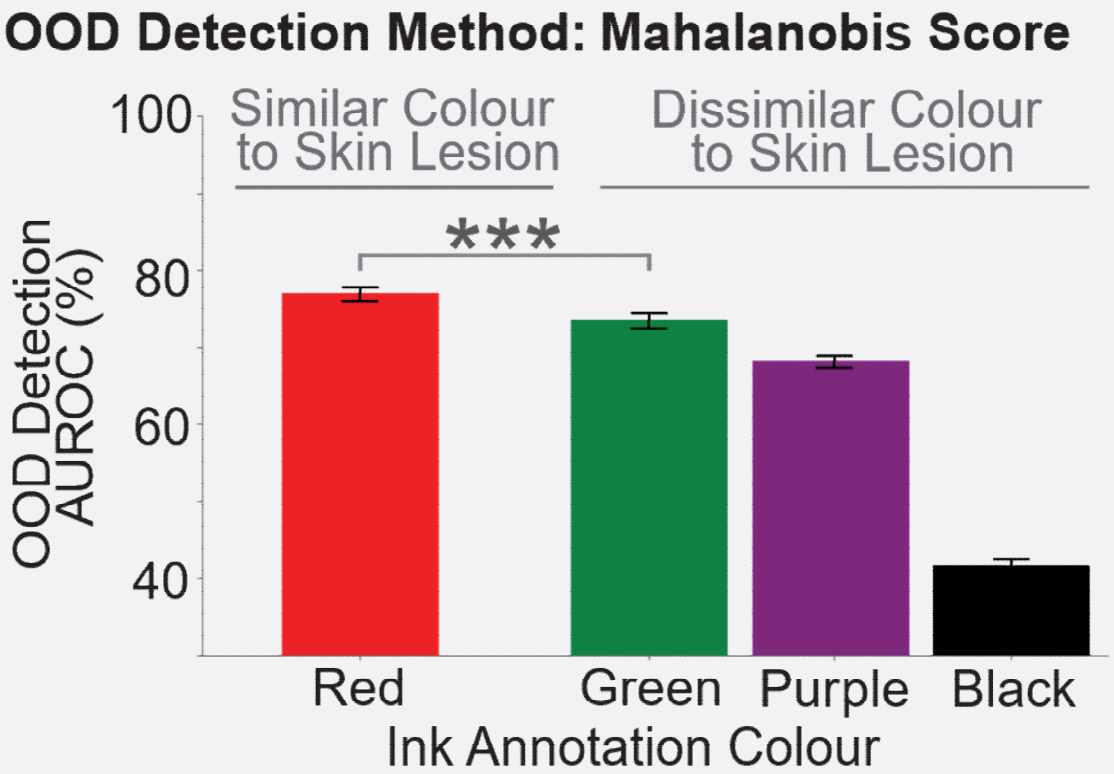


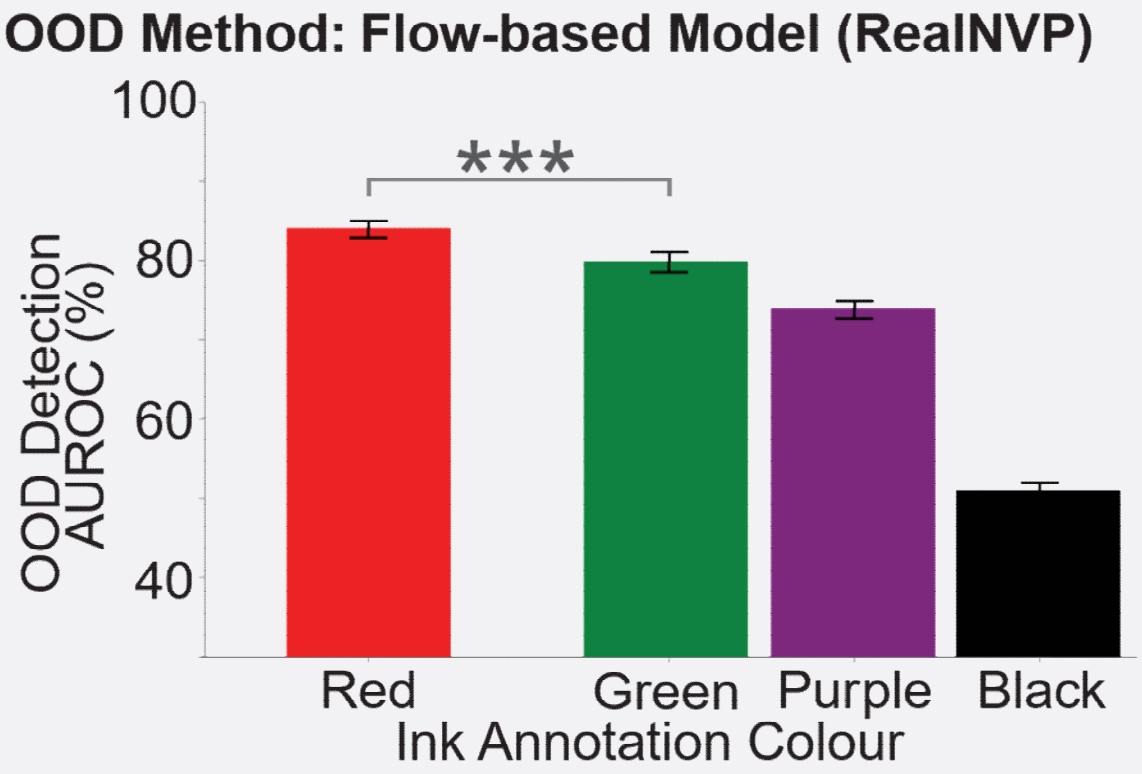
Black

Euclidean RGB Distance

Finding biases in

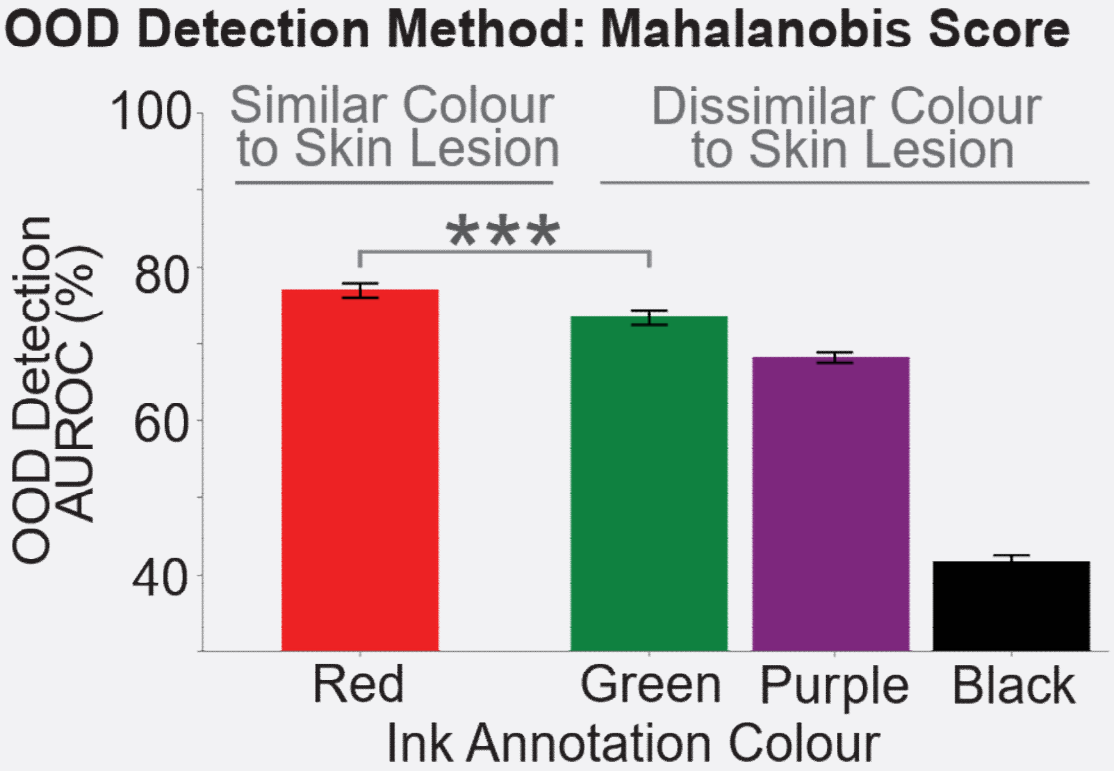
Out-of-distribution detection



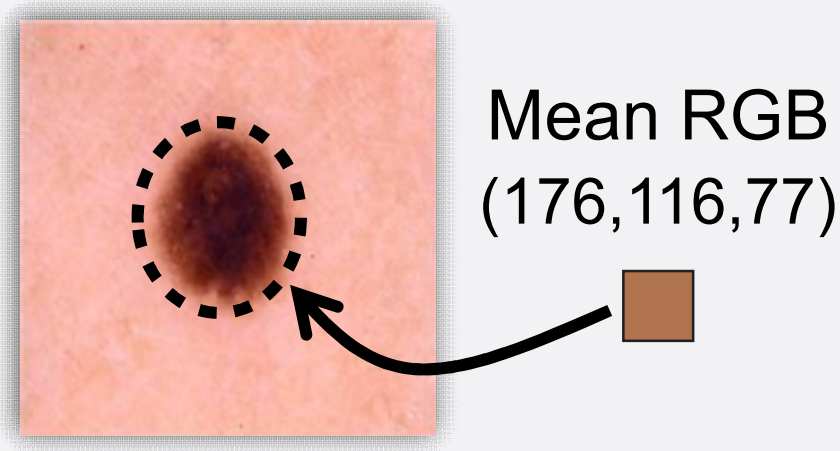


Finding biases in

Out-of-distribution detection



OOD detection performance **improves** when the OOD artefact is **visually similar** to model's Region of Interest.

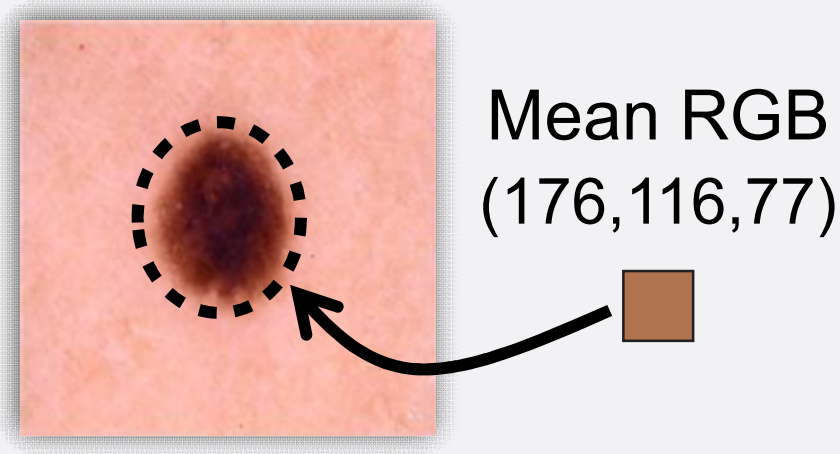


Finding biases in

Out-of-distribution detection

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

OOD detection performance **improves** when the OOD artefact is **visually similar** to model's Region of Interest.



Finding biases in

Out-of-distribution detection

The Invisible Gorilla Effect

OOD detection performance **improves** when the OOD artefact is **visually similar** to model's Region of Interest.

Community Assumptions

The prevailing assumption:

The more similar OOD data is to the training data, the harder it is to detect

Community Assumptions

Our findings challenge that assumption.
We reveal a bias case where the **opposite is true**.

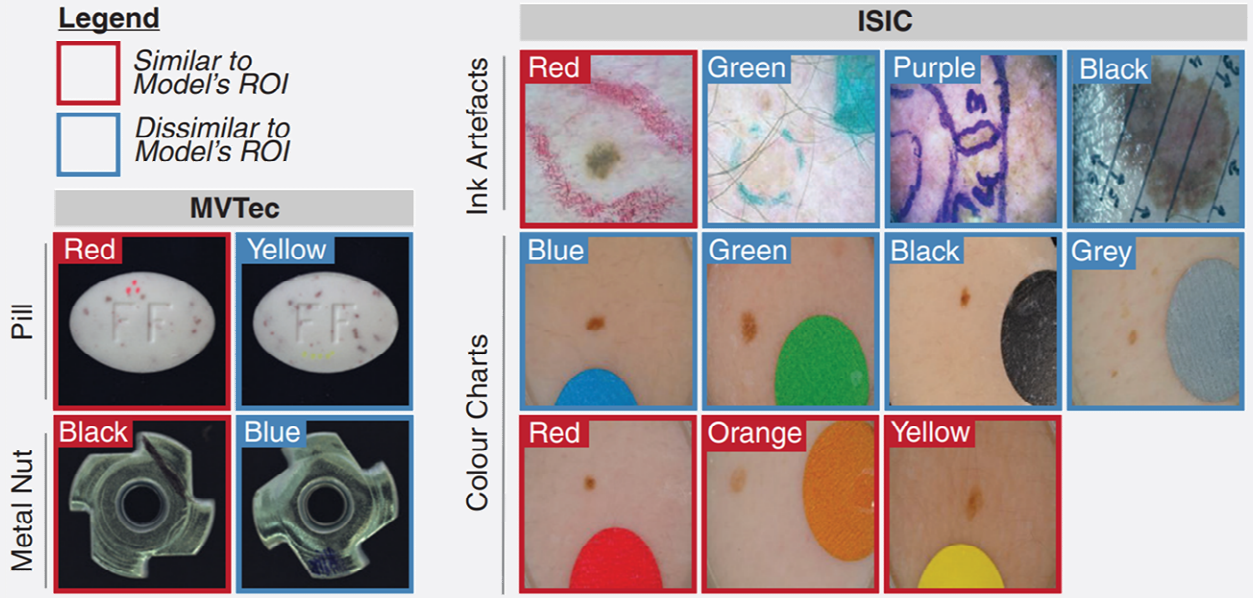
Suggests that OOD task difficulty is not a simple function of global similarity to the training data, but depends on:

- What the model attends to
- How that interacts with the OOD detection method

Contributions of this work

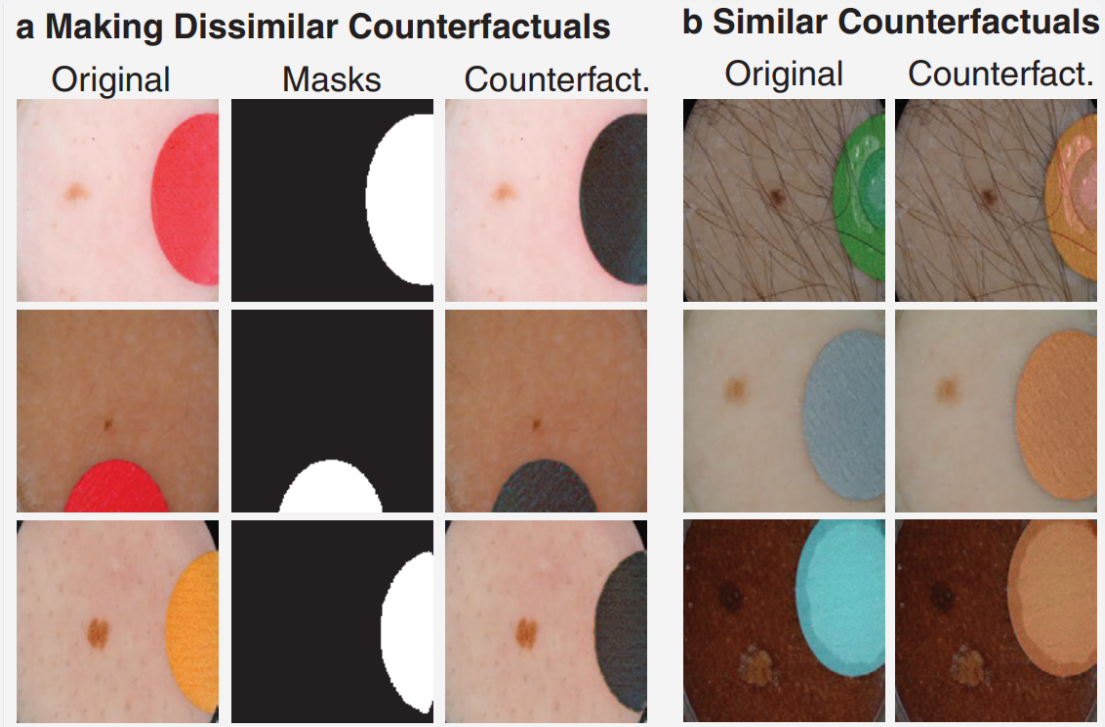
1. Create Benchmarks and Conduct a Large-Scale Evaluation
   - 7 **Benchmarks made public**

11,355 images



Contributions of this work

1. Create Benchmarks and Conduct a Large-Scale Evaluation
   - 7 **Benchmarks made public**
   - Colour-swapped Counterfactuals



Contributions of this work

1. Create Benchmarks and Conduct a Large-Scale Evaluation
   - 7 **Benchmarks made public**
   - Colour-swapped Counterfactuals
   - 3 Model Architectures
   - 40 OOD methods
   - 3,795 Hyperparameters
   - 25 Random Seeds

Contributions of this work

1. Create Benchmarks and Conduct a Large-Scale Evaluation
2. Identify Trends
3. Hypothesise the Causes for the Effect.
4. Use Insights to Inform Mitigation Strategies

Thank you for listening

I welcome any questions

Paper available at:

Data available at:

Full Presentation at:

@HarryEJAnthony