A Whac-A-Mole Dilemma : Shortcuts Come in Multiples Where Mitigating One Amplifies Others

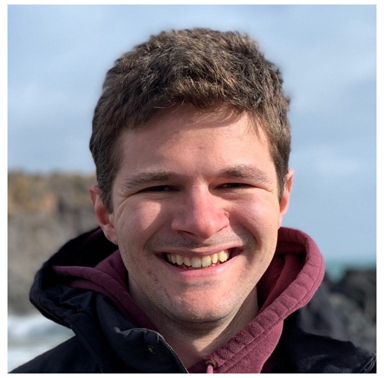
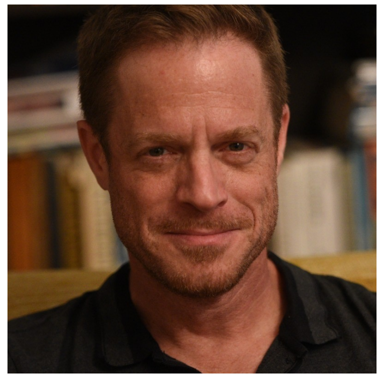
Zhiheng Li[2], *Ivan Evtimov[1], Albert Gordo[1], Caner Hazirbas[1],

Tal Hassner[1], Cristian Canton Ferrer[1], Chenliang Xu[2], *Mark Ibrahim[1] (*Equal Contribution)

[1]Meta AI   [2]University of Rochester

CVPR 2023

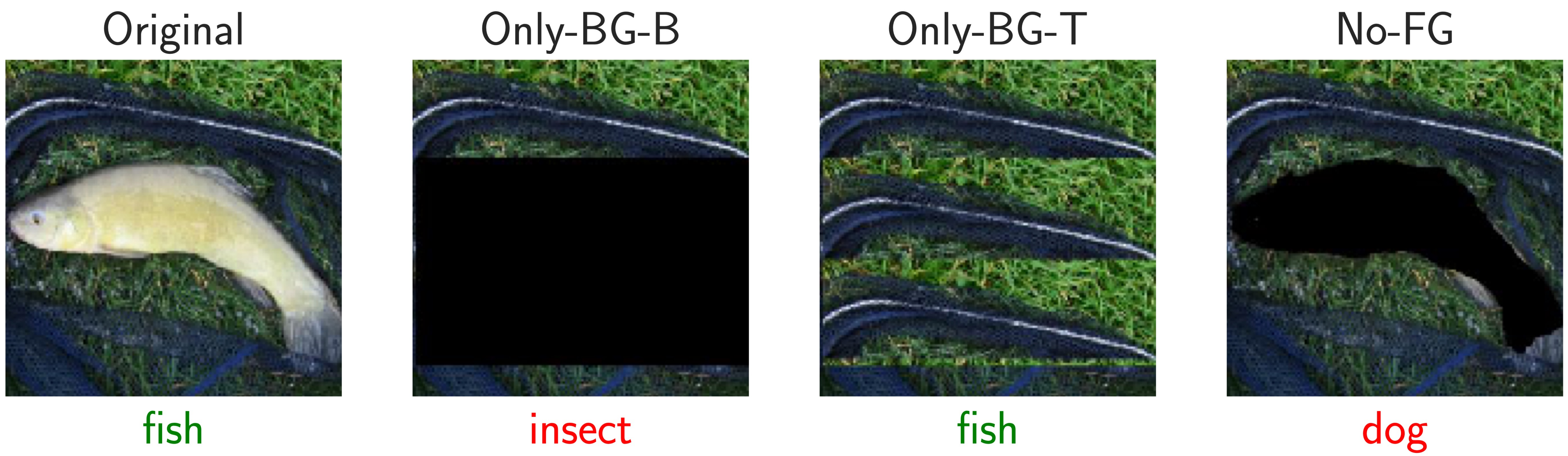
Poster ID: THU-AM-341

Background

Shortcut: unintended decision rules in machine learning models.

background shortcut in object recognition



[1] Xiao et al., “Noise or Signal: The Role of Image Backgrounds in Object Recognition,” in ICLR, 2021.

Previous works: single-shortcut

Waterbirds [1]
Shortcut: *background*

Real-world: multiple shortcuts

- E.g., multiple shortcuts for volleyball on ImageNet

Volleyball class in ImageNet

Shortcut #1: *Background*

Shortcut #2: *Co-occurring Object*

[1] Sagawa, et al., "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization," ICLR, 2020
[2] Singla and Feizi, "Salient ImageNet: How to discover spurious features in Deep Learning?," in ICLR, 2022.

Two New Multi-shortcut Benchmarks

1. **UrbanCars** dataset:

background shortcut

co-occurring object shortcut

2. New Multi-shortcut Benchmark for **ImageNet**

texture shortcut

Stylized ImageNet (Geirhos et al., 2019)

background shortcut

ImageNet-9 (Xiao et al., 2021)

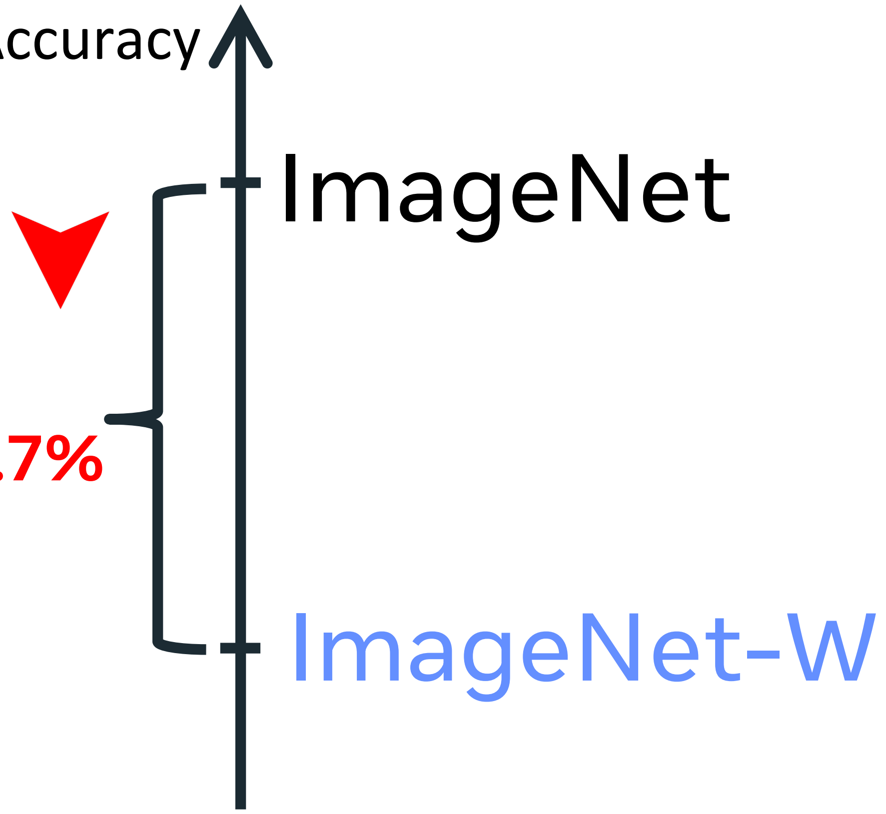
watermark shortcut

ImageNet-W (Ours)

ImageNet-W for Watermark Shortcut

- **Pervasive** watermark shortcut in **32 models**
  - E.g., MAE, SWAG, CLIP, …

- **Large accuracy drop**



- Max drop: **-26.7%**
- Average drop: **-10.7%**

watermark
shortcut

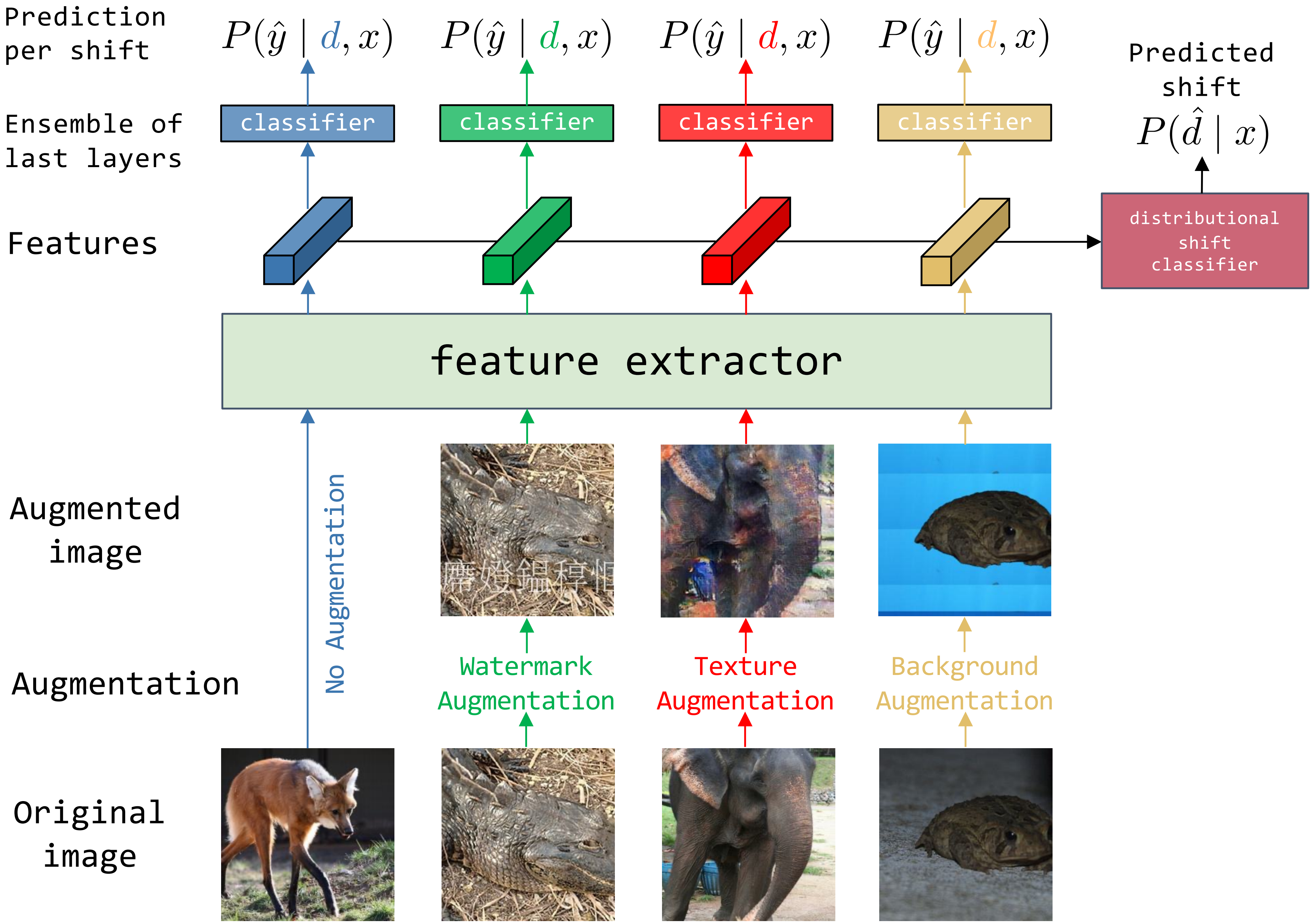
ImageNet-W (Ours)

Through a **comprehensive** benchmark on **26 methods**, we find that:

Multi-shortcut mitigation resembles a Whac-A-Mole game :
mitigating one shortcut amplifies reliance on others .

New Method:
Last Layer Ensemble (LLE)

- mitigate multiple shortcuts jointly



Detailed Introduction

A Whac-A-Mole Dilemma :
Shortcuts Come in Multiples Where Mitigating One Amplifies Others

Zhiheng Li[2], *Ivan Evtimov[1], Albert Gordo[1], Caner Hazirbas[1], Tal Hassner[1],
Cristian Canton Ferrer[1], Chenliang Xu[2], *Mark Ibrahim[1] (*Equal Contribution)

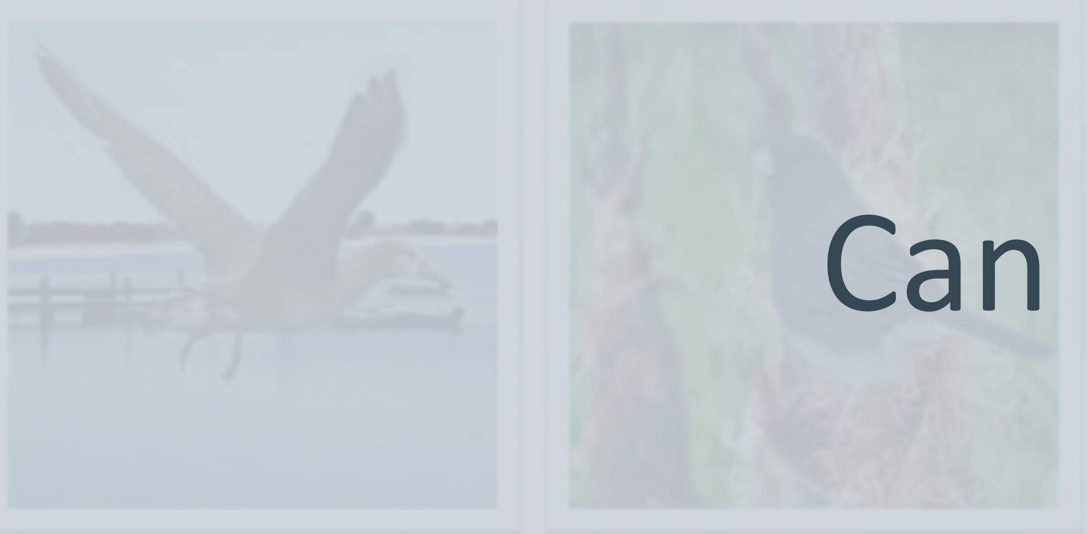
[1]Meta AI [2]University of Rochester

CVPR 2023

Previous works: **single-shortcut**

Real-world: **multiple shortcuts**

- E.g., multiple shortcuts for volleyball on ImageNet



Waterbirds [1]
Shortcut: *background*

Volleyball class in ImageNet

Shortcut #1: *Background*

Shortcut #2: *Co-occurring Object*

Can existing methods overcome **multiple** shortcuts?

[1] Sagawa, et al., "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization," ICLR, 2020
[2] Singla and Feizi, "Salient ImageNet: How to discover spurious features in Deep Learning?," in ICLR, 2022.

UrbanCars Dataset:
A new multi-shortcut dataset

UrbanCars
dataset

- **UrbanCars**:
  - semi-synthetic for controlled setting
- **Advantages**:
  - Commonly seen shortcuts
  - Controllable spurious correlation
  - Shortcut labels

UrbanCars Dataset

Main task:
urban car vs. *country car*

Shortcut #2:
co-occurring object (CoObj)

urban CoObj
(**common**)

country CoObj
(uncommon)

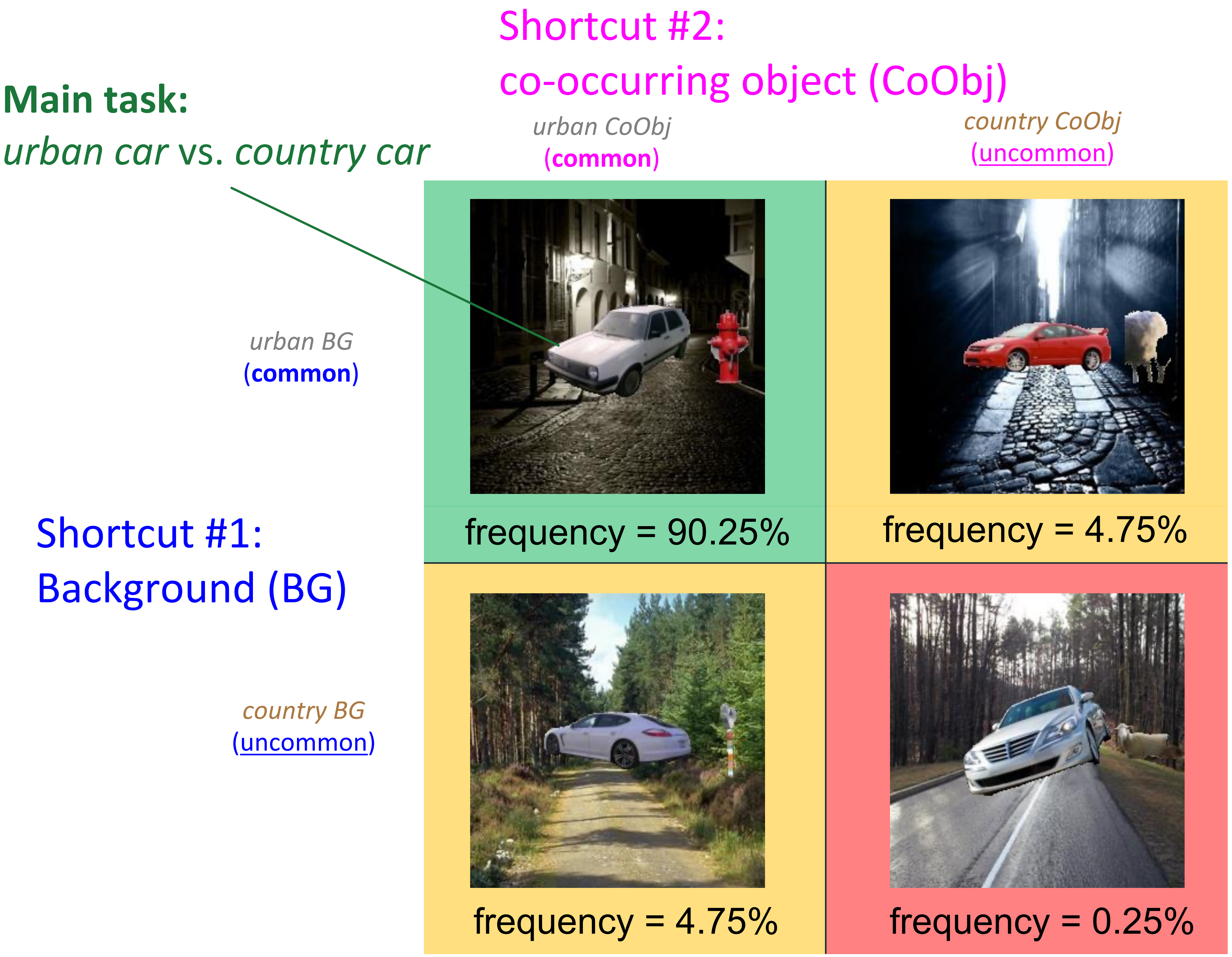
urban BG
(**common**)

Shortcut #1:
Background (BG)

country BG
(uncommon)

Example images of *urban car* in the UrbanCar's training set.

UrbanCars Dataset



Example images of *urban car* in the UrbanCar's training set.

ImageNet-W: Multiple Shortcuts in ImageNet

A new ImageNet variant to evaluate watermark shortcut reliance

Stylized ImageNet (SIN)

Evaluate texture shortcut

Mixed-Same

Mixed-Rand

ImageNet-9

Evaluate background shortcut

ImageNet-W (Ours)

Evaluate watermark shortcut

[1] Geirhos, et al., “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness,” in ICLR, 2019.
[2] Xiao, et al., “Noise or Signal: The Role of Image Backgrounds in Object Recognition,” in ICLR, 2021.

Multiple Shortcuts in ImageNet

Stylized ImageNet (SIN)

Evaluate texture shortcut

Mixed-Same

Mixed-Rand

ImageNet-9

Evaluate background shortcut

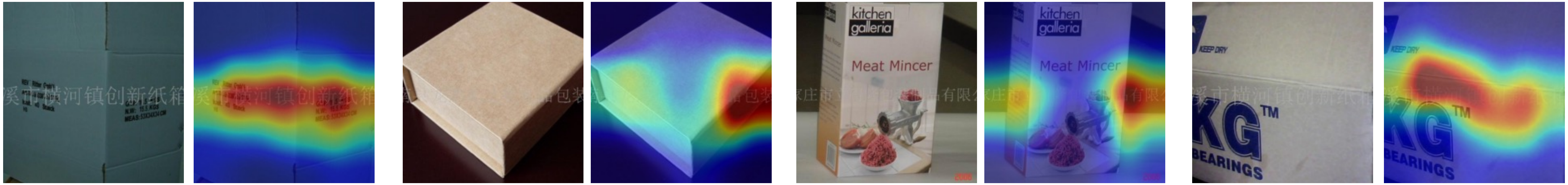
ImageNet-W (Ours)

Evaluate watermark shortcut

[1] Geirhos, et al., "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in ICLR, 2019.
[2] Xiao, et al., "Noise or Signal: The Role of Image Backgrounds in Object Recognition," in ICLR, 2021.

Watermark Shortcut in ImageNet

- ImageNet **training** set: many carton images have watermark

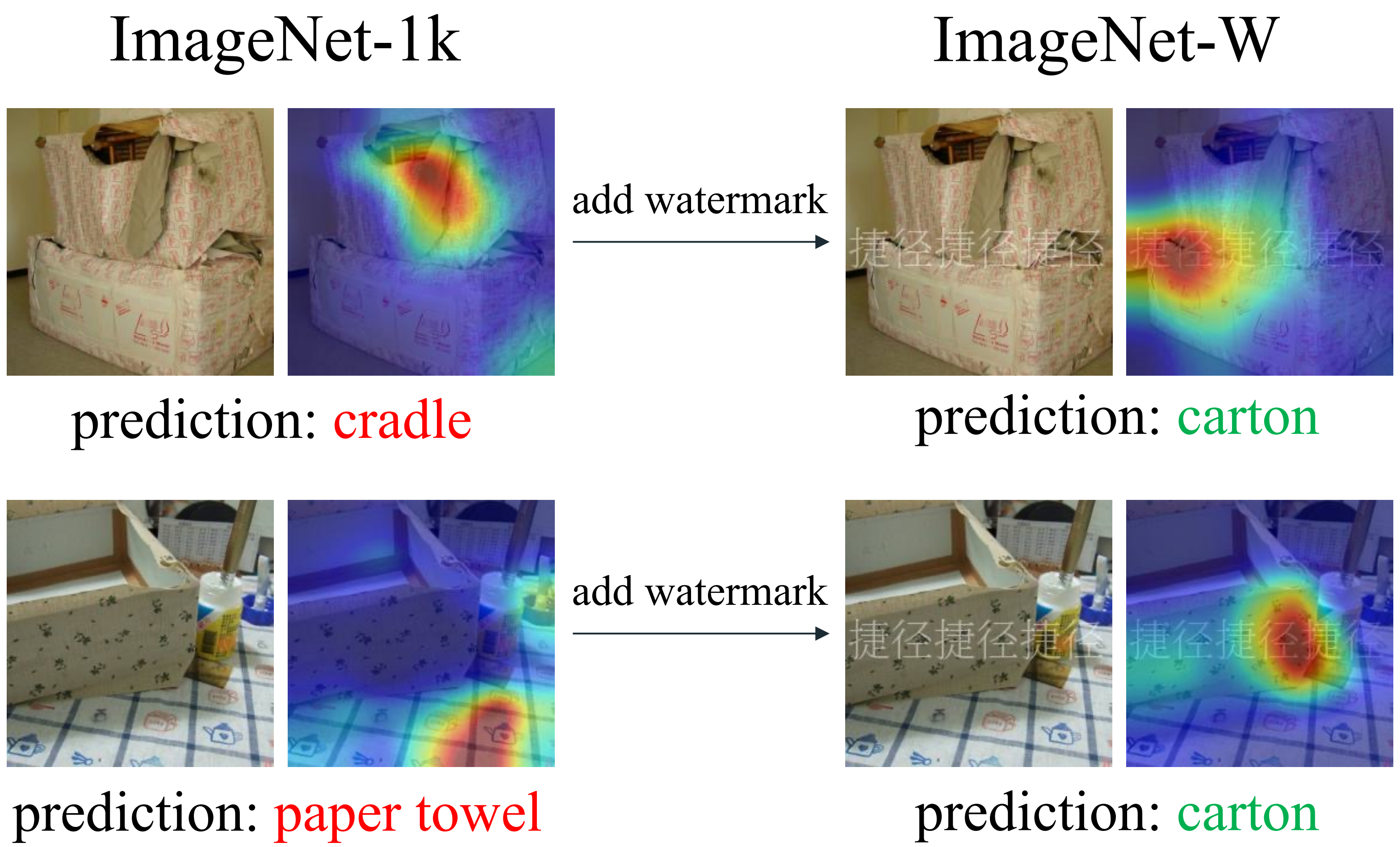


- ImageNet **validation** set: **none** of them have watermark

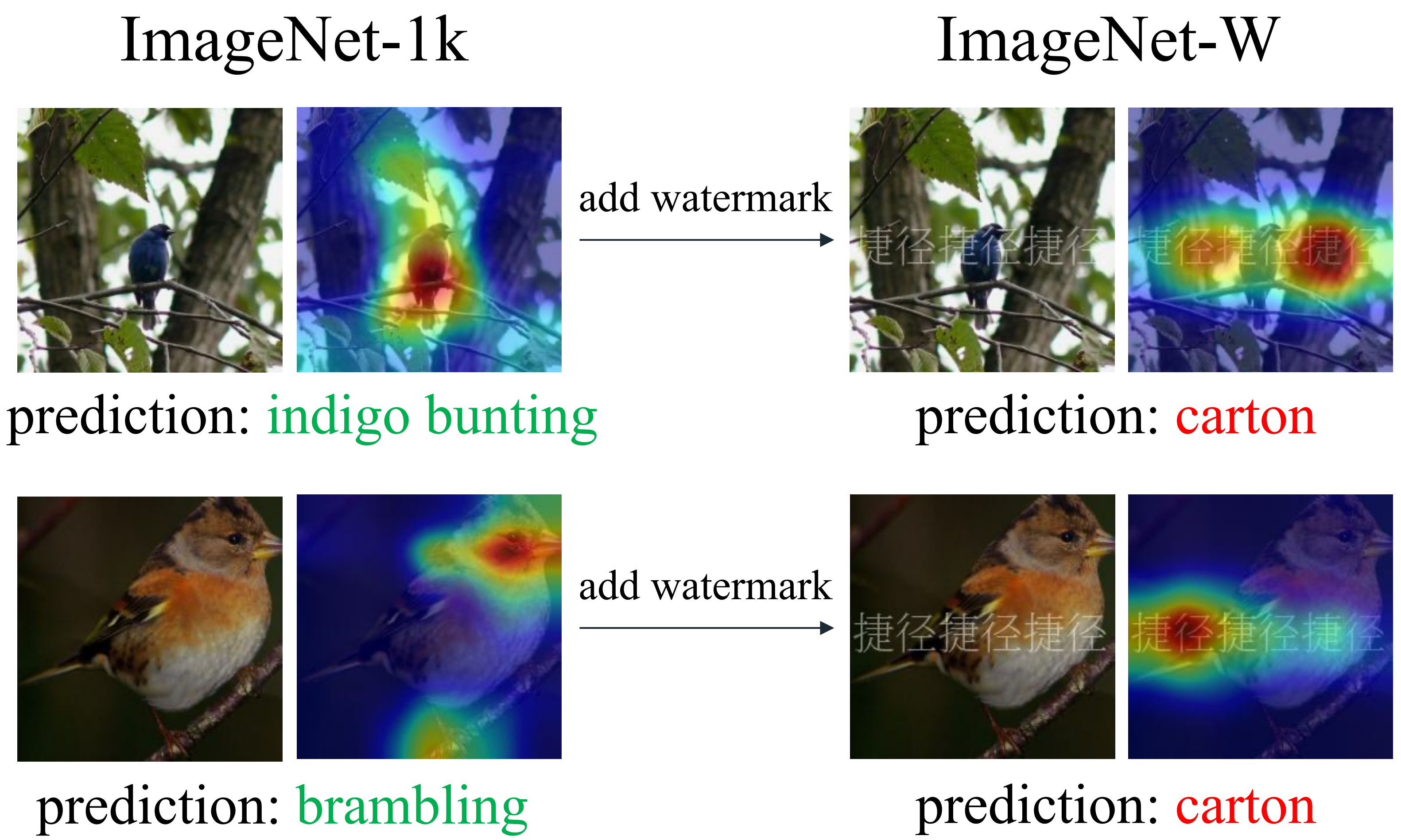
ImageNet-Watermark (ImageNet-W)

- **ImageNet-W**: A new ImageNet test set to evaluate the watermark shortcut reliance
  - Overlay transparent watermark “捷径捷径捷径”
    - 捷径 (jié jìng): *shortcut* in Chinese

Carton class images:



Non-carton class images:



Pervasive watermark shortcut reliance in 32 models

regardless of:

- **Architectures and Sizes**: ResNet, RegNet, ViT(-B, -L, -H), …
- **Training data:** ImageNet, IG, WIT, LAION, …
- **Augmentation and Regularization:** Mixup, CutMix, Cutout, AugMix, …
- **Supervision**: (fully, weakly, self, text)-supervised
- **Transfer learning**: no transfer, linear probing, fine-tune, zero-shot

- **Self-supervised and foundation models**: MoCov3, MAE, SWAG, CLIP (zero-shot)…

Large accuracy drop from ImageNet to ImageNet-W:

Max drop: **-26.7%** Average drop: **-10.7%**

Pervasive watermark shortcut reliance in 32 models

regardless of:

- **Architectures and Sizes**: ResNet, RegNet, ViT(-B, -L, -H), …
- **Training data:** ImageNet, IG, WIT, LAION, …
- **Augmentation and Regularization:** Mixup, CutMix, Cutout, AugMix, …
- **Supervisions**: (fully, weakly, self, text)-supervised
- **Transfer learning**: no transfer, linear probing, fine-tune, zero-shot

- **Self-supervised and foundation models**: MoCov3, MAE, SWAG, **CLIP (zero-shot)…**

i.e., not trained on ImageNet

[1] Alec Radford, et al., "Learning Transferable Visual Models From Natural Language Supervision," in ICML, 2021.

Why does CLIP (zero-shot) also rely on watermark shortcut?

LAION (size: 400M to 2B) also contains watermarks.

[1] C. Schuhmann et al., “LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs,” in NeurIPS Workshops, 2021.
[2] C. Schuhmann et al., “LAION-5B: An open large-scale dataset for training next generation image-text models,” in NeurIPS Datasets and Benchmarks Track, 2022.

Evaluation Metrics: Acc Gap between I.D. and O.O.D.

UrbanCars

Background shortcut

accuracy

In-distribution Accuracy

BG Gap

Accuracy of **uncommon** BG

Co-occurring Object shortcut

accuracy

In-distribution Accuracy

CoObj Gap

Accuracy of **uncommon** CoObj

ImageNet

Texture shortcut

accuracy

ImageNet

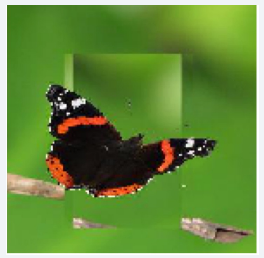
SIN Gap

Stylized ImageNet

Background shortcut

accuracy

Mixed-Same



IN-9 Gap

Mixed-Rand

Watermark shortcut

accuracy

ImageNet

IN-W Gap

ImageNet-W

Methods for Multi-Shortcut Benchmark

Empirical Risk Minimization (ERM)

I.e.,
standard supervised training

Shortcut Mitigating Methods

Category 1: Standard Augmentation and Regularization

Example method:

CutMix
(Yun et al., ICCV'19)

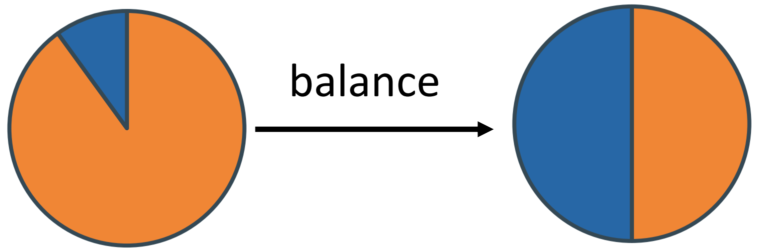
Category 2: Targeted Augmentation

Example method:

Style Transfer
(Geirhos et al., ICLR'19)

Category 3: Using Shortcut Labels

Example method:



SUBG
(Idrissi et al., CLeaR'22)

Category 4: Inferring Pseudo Shortcut Labels

Example method:

error set ≈ shortcut absent

JTT
(Liu et al., ICML'21)

Results of Multi-Shortcut Benchmark

Shortcut Mitigating Methods

Empirical Risk Minimization (ERM)

I.e., standard supervised training

Category 1: Standard Augmentation and Regularization

Category 2: Targeted Augmentation

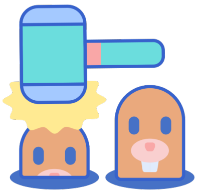
Category 3: Using Shortcut Labels

Category 4: Inferring Pseudo Shortcut Labels

All methods:

Mitigate a shortcut: closing the accuracy gap relative to **ERM** baseline

Amplify a shortcut: enlarge the accuracy gap relative to **ERM** baseline



mitigate one shortcut but amplify other shortcuts

Results of Multi-Shortcut Benchmark

Shortcut Mitigating Methods

Empirical Risk Minimization (ERM)

I.e.,
standard supervised training

Category 1:

Standard Augmentation and Regularization

E.g., CutMix
x2.94 BG Gap
Background

Category 2:

Targeted Augmentation

targeted
untargeted

E.g., Style Transfer
X1.20 IN-W Gap
Watermark

Category 3:

Using Shortcut Labels

labeled
unlabeled

E.g., SUBG (BG)
X3.24 CoObj Gap
Co-occurring Object

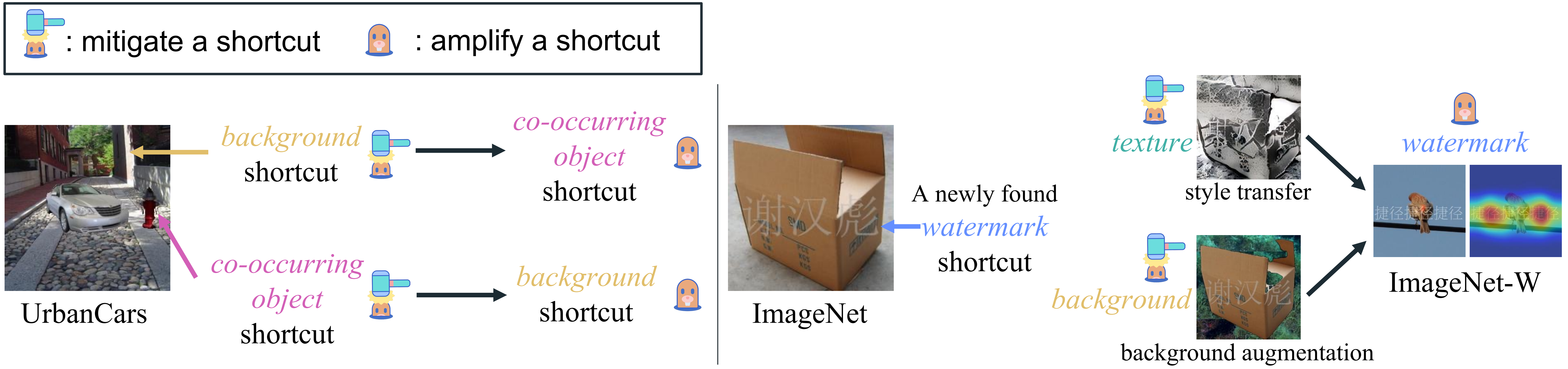
Category 4:

Inferring Pseudo Shortcut Labels

inferred
missed

E.g., EIIL (epoch=1)
X2.21 CoObj Gap
Co-occurring Object

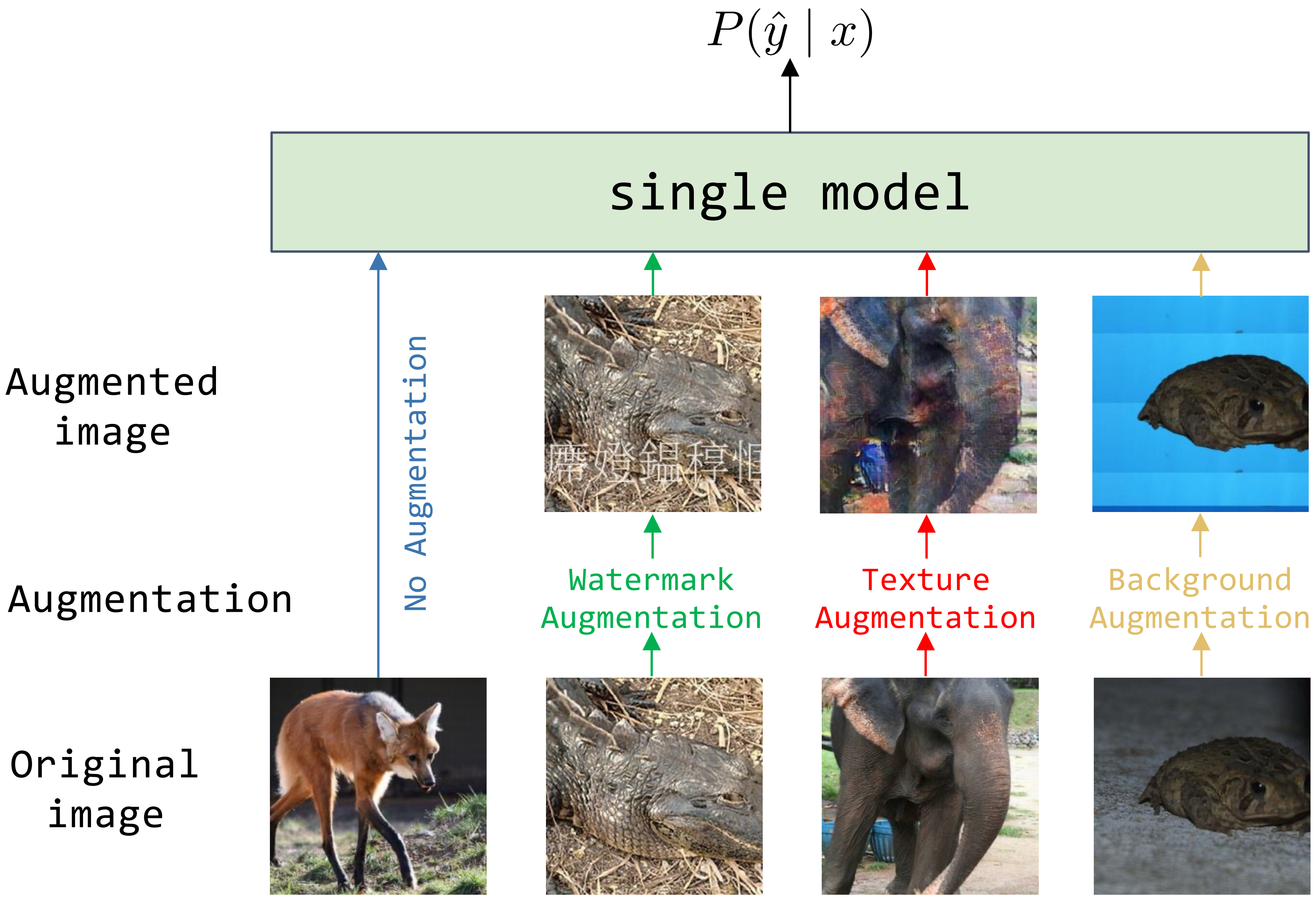
Takeaway



Takeaway: Multi-shortcut mitigation resembles a Whac-A-Mole game, i.e., mitigating one shortcut amplifies reliance on others .

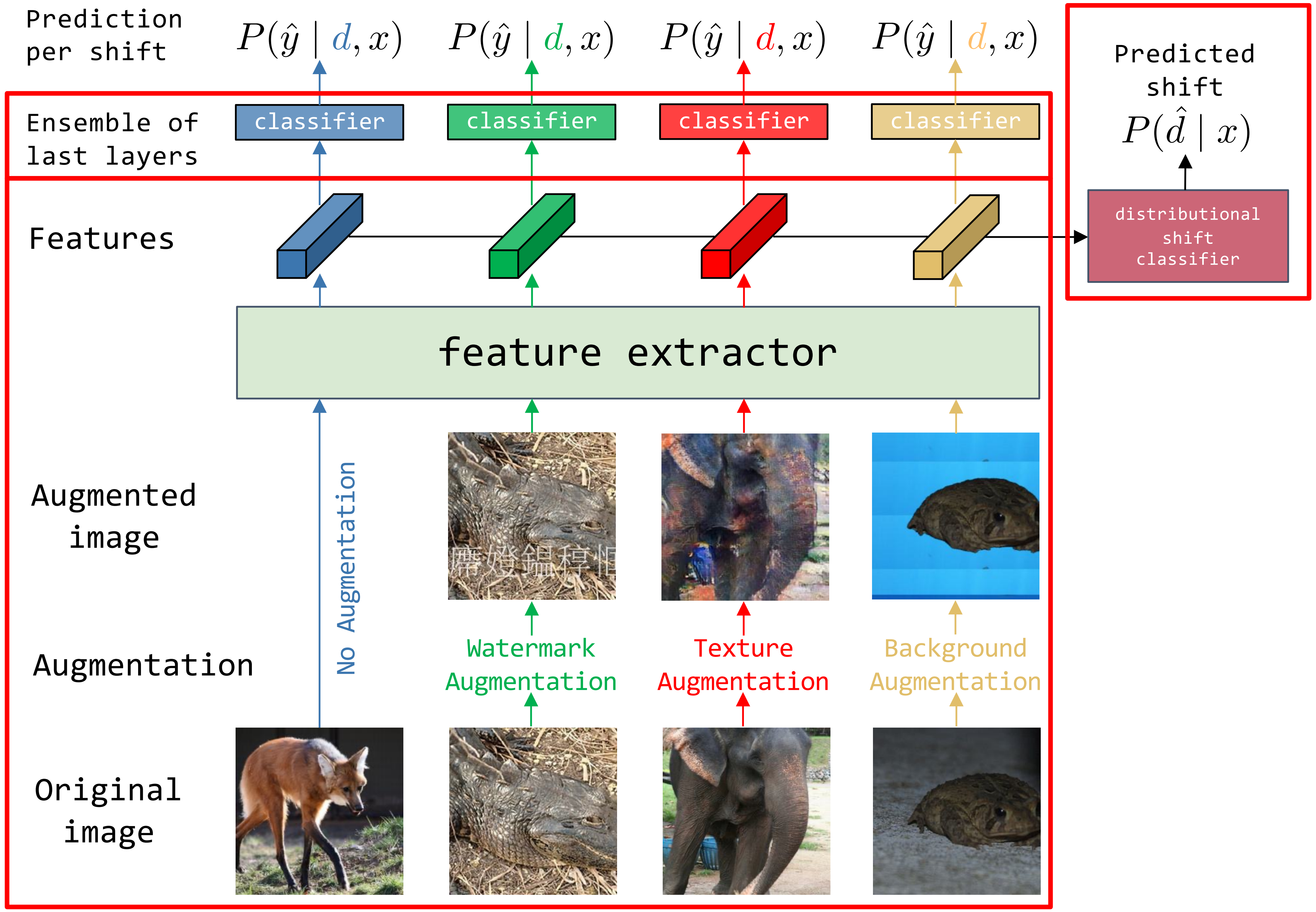
How to address this problem?

A straightforward solution:
using multiple targeted augmentations



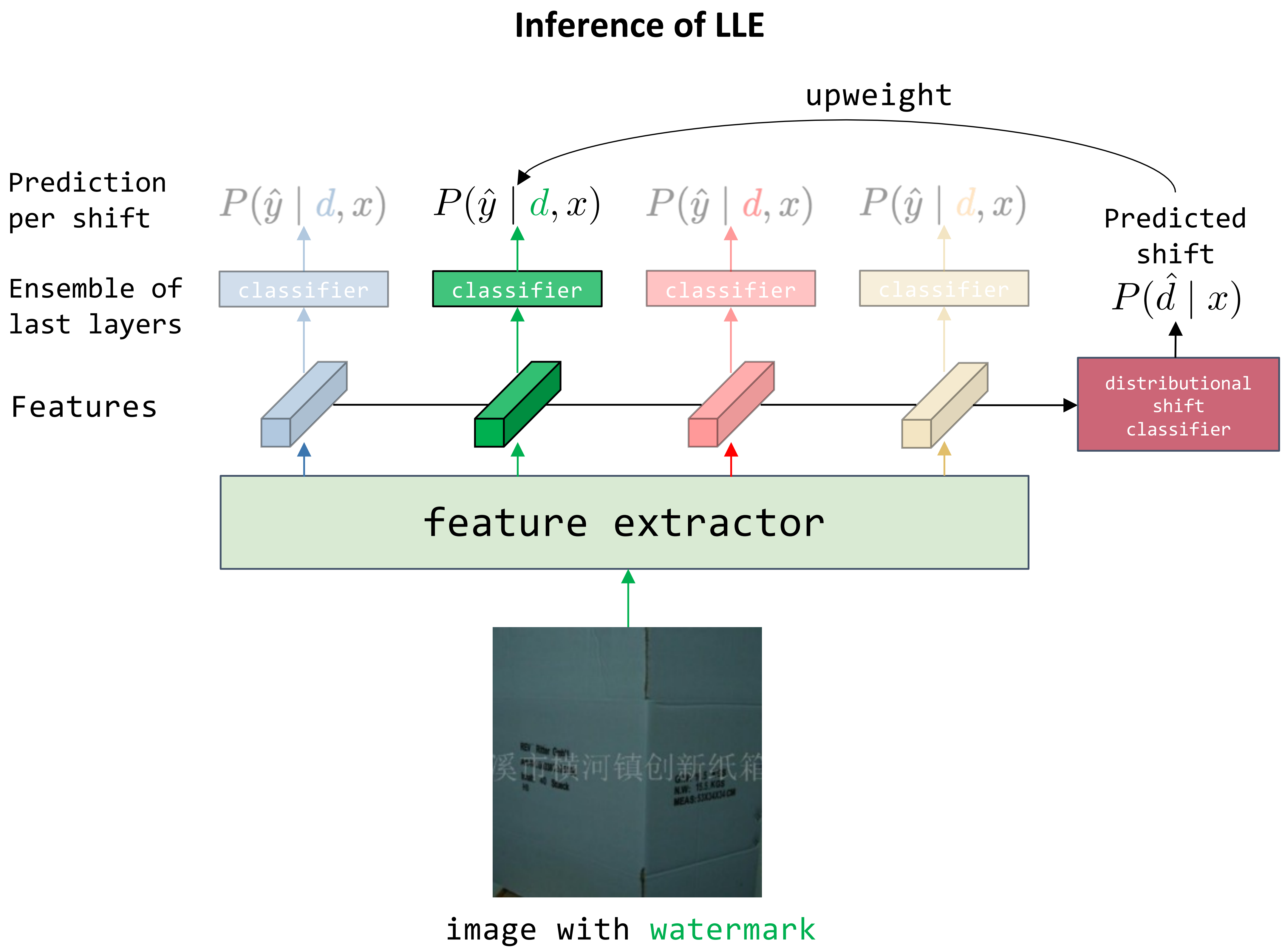
- **Suboptimal**
- **Reason**:
  - Incompatible <shortcut, augmentation>
    - E.g., <watermark, Texture Aug>

Our approach: Last Layer Ensemble (LLE)



- **Our approach**:
  - Ensemble of last classifier layers
  - Each layer is trained only with one type of augmentation
  - Aggregate prediction based on predicted distributional shift

Our approach: Last Layer Ensemble (LLE)



Code and Datasets

https://github.com/facebookresearch/Whac-A-Mole

One command to use **ImageNet-W**:

```
pip install imagenet-w
```