# Discovering and Mitigating Visual Biases through Keyword Explanation

Younghyun Kim*, Sangwoo Mo*, Minkyu Kim, Kyungmin Lee, Jaeho Lee, Jinwoo Shin

CVPR 2024 Highlight

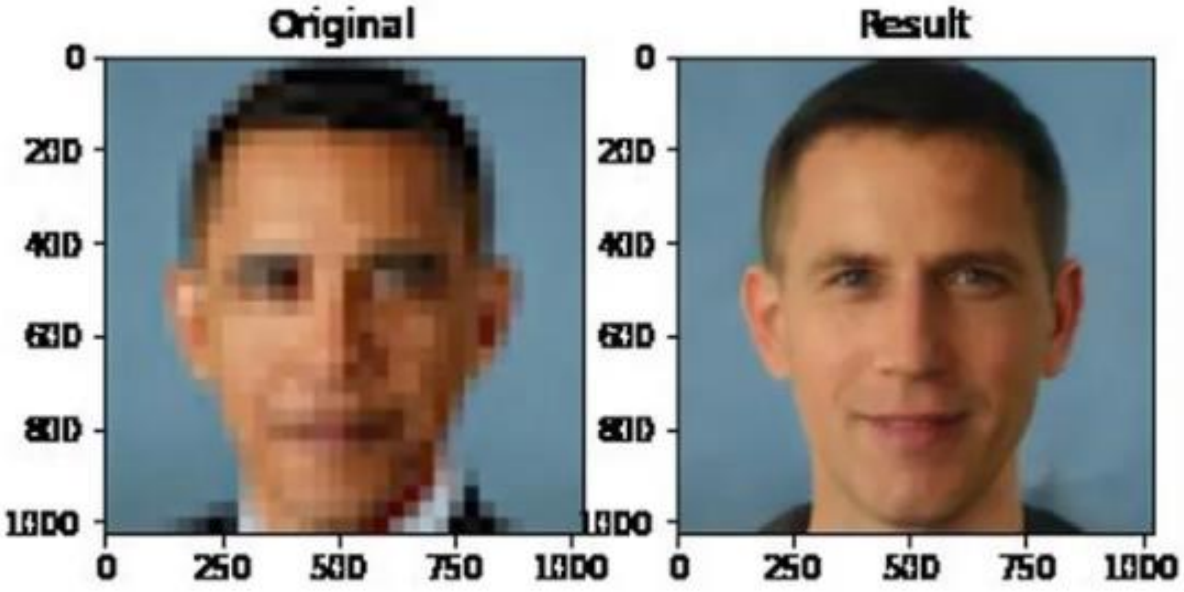
2024.06.04.

https://arxiv.org/pdf/2301.11104

# Biases are everywhere in ML domain

- There exist visual biases inherited from ML algorithm in real-world application


Google Photos automatic tagging




PULSE algorithm: low pixel image to high resolution image

https://www.bbc.com/news/technology-33347866
https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

# These visual biases pose several critical problems

- Biases may cause fairness issue
- Biases may harm model performance

Classifier **mispredicts** blond male !

**Rare** in training examples
(bias)

[Sagawa et al., 2019] Distributionally Robust Neural Networks for Group Shifts

# However, visual biases are not interpretable

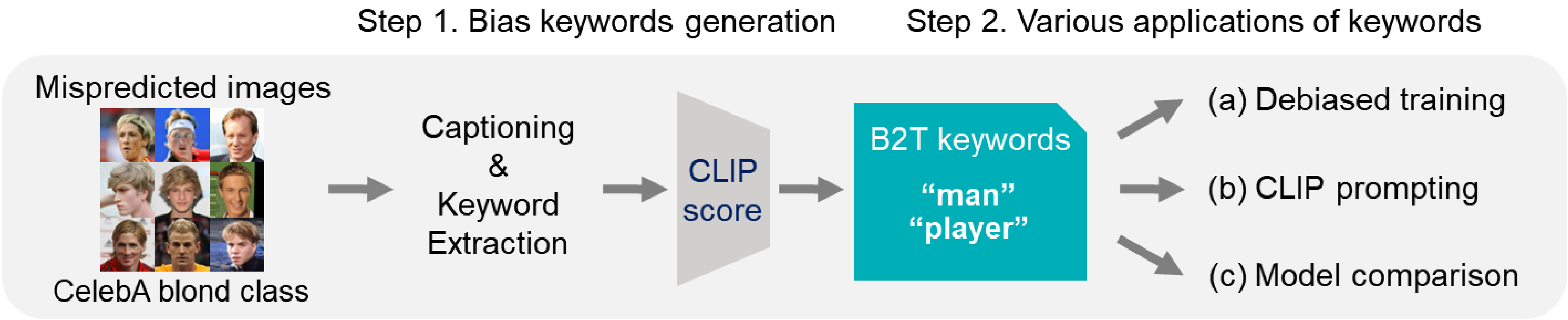
- Prior works visualized spurious features that are not human-readable
- Thus, they are hard to be directly utilized for debiasing

(a) **class:** band aid, **spurious feature:** fingers, **-41.54%** (b) **class:** space bar, **spurious feature:** keys, **-46.15%** (c) **class:** plate, **spurious feature:** food, **-32.31%** (d) **class:** butterfly, **spurious feature:** flowers, **-21.54%** (e) **class:** potter's wheel, **spurious feature:** vase, **-21.54%**

[Singla et al., 2022] Salient ImageNet: How to Discover Spurious Features in Deep Learning

# B2T: Bias-to-text

- We use language to interpret visual biases
- We first **extract** B2T keywords, then use them to **various applications**: (a) debiased training, (b) CLIP prompting, and (c) model comparison



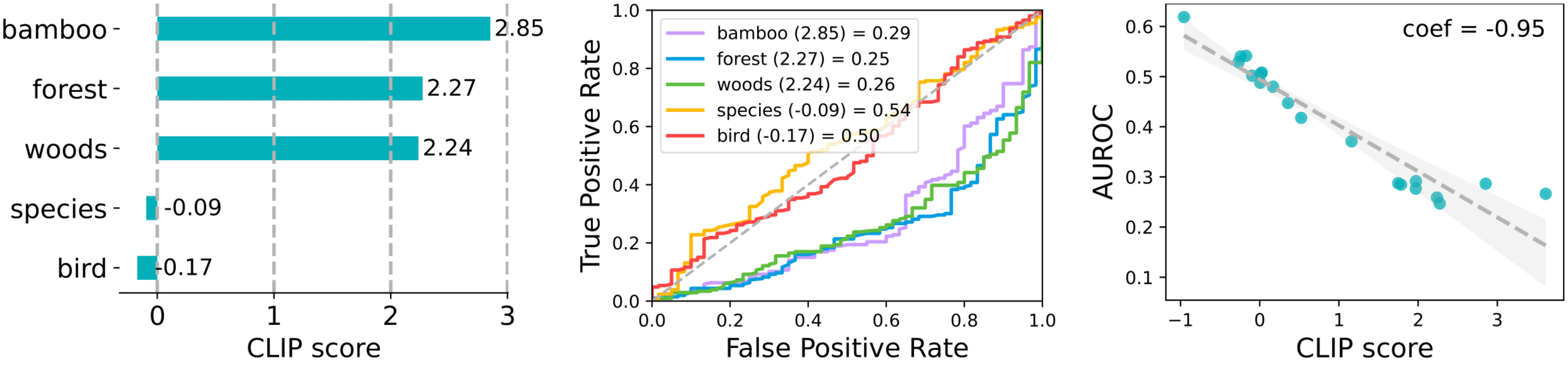
# CLIP score

- CLIP score measures the similarity between keyword $a$ and correctly or incorrectly classified images $x$ from a validation set $\mathcal{D}$

$$s_{\text{CLIP}}(a; \mathcal{D}) := \text{sim}(a, \mathcal{D}_{\text{wrong}}) - \text{sim}(a, \mathcal{D}_{\text{correct}}).$$

# Validation of the CLIP score

- CLIP score effectively identifies incorrect bias keywords
- e.g.) waterbird class in the Waterbirds dataset



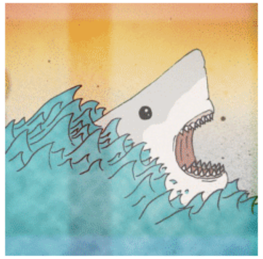

(a)CLIP score

(b) ROC curve of subgroup acc.

(c) Correlation of CLIP score and AUROC

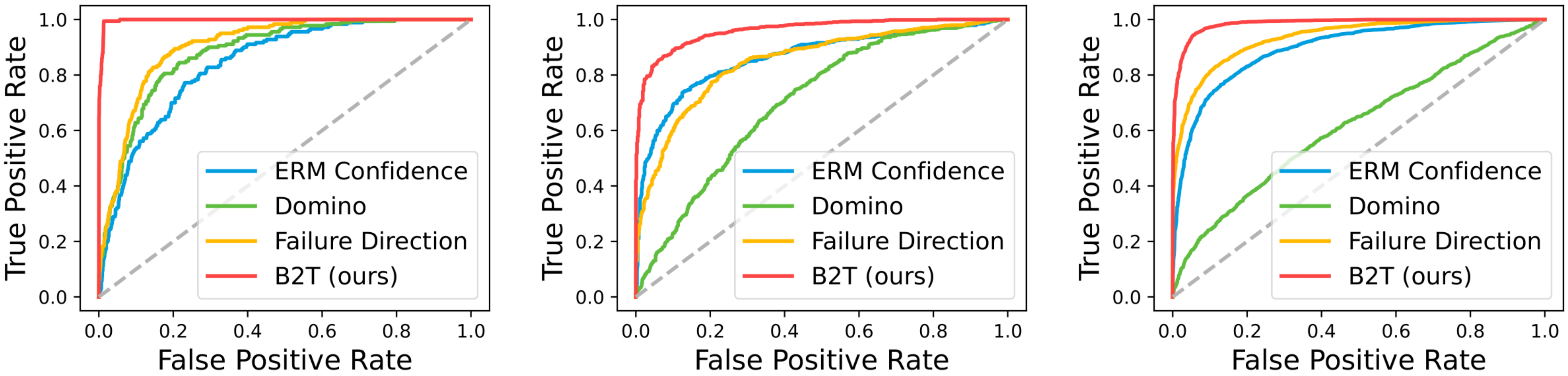
# Can B2T identify the known biases?

- B2T discovers **spurious correlations** and **distributions shifts**
- e.g.) "man" for CelebA blond / "forest" and "ocean" for Waterbirds
  "illustration" and "drawing" for IN-R / "snow" and "window" for IN-C

| | (a) CelebA blond | | (b) Waterbirds | | (c) ImageNet-R | | (d) ImageNet-C snow / frost | |
|---|---|---|---|---|---|---|---|---|
| **Keyword** | Man | | Forest | Ocean | Illustration | Drawing | Snow | Window |
| **Samples** |  |  |  |  |  |  |  |  |
| **Actual** | blond | blond | waterbird | landbird | backpack | white shark | airliner | American egret |
| **Pred.** | not blond | not blond | landbird | waterbird | maze | envelope | damselfly | quill |
| **Caption** | person, a **man** with a beard. | actor as a young **man**. | a bird in the **forest**. | a bird in the **ocean**. | hand drawn **illustration** of a backpack. | a **drawing** of a shark attacking […] | airliner in the **snow**, photo. | a bird on a frozen **window**. |

# Sample-wise bias labeling

- B2T successfully infers **sample-wise bias (or group) labels**




(a) CelebA blond (b) Waterbirds waterbird (c) Waterbirds landbird

# Novel real-world biases

- B2T explores **novel biases** in larger datasets
- e.g.) "cave," "fire," "bucket," and "hole" for Dollar Street
  "flower," "playground," "baby," and "interior" for ImageNet

| | (e) Dollar Street | | | | (f) ImageNet | | | |
|---|---|---|---|---|---|---|---|---|
| **Keyword** | Cave | Fire | Bucket | Hole | Flower | Playground | Baby | Interior |
| **Samples** |  |  |  |  |  |  |  |  |
| **Actual** | wardrobe | stove | plate rack | toilet seat | ant | horizontal bar | stethoscope | monastery |
| **Pred.** | poncho | caldron | oil filter | wheelbarrow | bee | swing | baby pacifier | arched ceiling |
| **Caption** | the **cave** is full of surprises. | a **fire** in the kitchen. | a **bucket** of water and a few tools. | the **hole** in the ground. | a yellow **flower** with a black head. | person on a swing in the **playground**. | a newborn **baby** boy in a stethoscope. | the **interior** of the church. |

# Debiased DRO training

- Bias keywords can be used as group names for **debiased distributionally robust optimization (DRO) training**

| | | CelebA blond | | Waterbirds | |
|---|---|---|---|---|---|
| Method | GT | Worst | Avg. | Worst | Avg. |
| ERM | - | 47.7±2.1 | 94.9 | 62.6±0.3 | 97.3 |
| LfF [55] | - | 77.2 | 85.1 | 78.0 | 91.2 |
| GEORGE [74] | - | 54.9±1.9 | 94.6 | 76.2±2.0 | 95.7 |
| JTT [44] | - | 81.5±1.7 | 88.1 | 83.8±1.2 | 89.3 |
| CNC [86] | - | 88.8±0.9 | 89.9 | 88.5±0.3 | 90.9 |
| DRO-B2T (ours) | - | **90.4**±0.9 | 93.2 | **90.7**±0.3 | 92.1 |
| DRO [66] | ✓ | 90.0±1.5 | 93.3 | 89.9±1.3 | 91.5 |

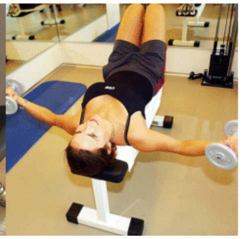
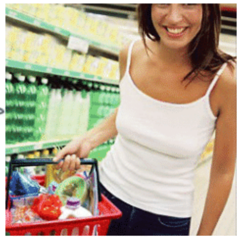
# CLIP zero-shot prompting

- Bias keywords can **improve the CLIP zero-shot classifier** by integrating them into prompt

| | CelebA blond | | Waterbirds | |
|---|---|---|---|---|
| | Worst | Avg. | Worst | Avg. |
| CLIP zero-shot | 76.2 | 85.2 | 50.3 | 72.7 |
| + Group prompt [85] | 76.7 | 87.0 | 53.7 | 78.0 |
| + B2T-neg prompt | 72.9 | 88.0 | 45.4 | 70.8 |
| + B2T-pos prompt (ours) | **80.0** | 87.2 | **61.7** | 76.9 |

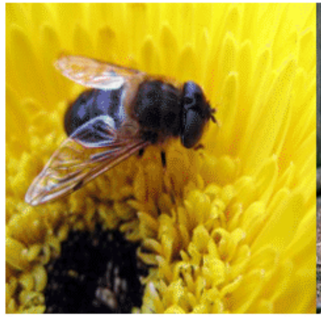
# Model comparison

- Bias keywords can be used to **analyze and compare different classifiers** based on their keywords
- e.g.) architecture: ResNet vs. ViT

| **Keyword** | Work | | Supermarket | |
|---|---|---|---|---|
| **Samples** |  |  |  |  |
| **ViT-B** | O | O | O | O |
| **RN50** | O | X | O | X |
| **Actual (RN50)** | dumbbell | dumbbell | shopping basket | shopping basket |
| **Pred (RN50)** | dumbbell | horizontal bar | shopping basket | grocery store |
| **Caption** | a set of dumbbells with weights. | person **work**s out in the gym. | a basket full of food. | woman shopping in a **supermarket**. |

# Label diagnosis

- B2T can diagnose **common labeling errors**, such as mislabeling and label ambiguities

| **Keyword** | Bee | Boar | Desk | Market |
|---|---|---|---|---|
| **Samples** |  |  |  |  |
| **Label** | fly | pig | computer mouse | custard apple |
| **Pred.** | bee | wild boar | desktop computer | grocery store |
| **Caption** | a **bee** on a yellow flower. | wild **boar** in the forest. | the **desk** in the office. | fruit and vegetables at the **market**. |

# B2T: Bias-to-Text

- We interpret visual biases as **keywords** that enables the **discovery of novel biases** and the **effective debiasing of vision models**