CRAYON: Mitigating Spurious Correlations in Image Classifiers with Simple Yes-No Annotations

Seongmin Lee
Georgia Tech
Atlanta, Georgia, USA
seongmin@gatech.edu

Ali Payani
Cisco Systems Inc.
San Jose, California, USA
apayani@cisco.com

Duen Horng (Polo) Chau
Georgia Tech
Atlanta, Georgia, USA
polo@gatech.edu

Abstract

Modern deep learning models often rely on spurious correlations between data and labels that solely present in the training data, resulting in biased performance and limited generalization. Existing methods aimed at mitigating spurious correlations often make the impractical assumption that developers have full knowledge of which attributes are spuriously correlated in the training data. We present CRAYON *(**C**orrecting **R**easoning with **A**nnotations of **Y**es **O**r **N**o), which offers effective, scalable, and practical solutions to refine models with spurious correlations using simple yes-no annotations.* CRAYON *empowers both classical and modern model interpretation techniques to not only identify but also guide model reasoning:* CRAYON-ATTENTION *guides classic interpretations based on saliency maps to focus on relevant image regions, and* CRAYON-PRUNING *prunes irrelevant neurons identified by modern concept-based methods to remove their influence. Evaluation of* CRAYON *with the annotations collected from 2,875 participants highlights its remarkable ability to effectively mitigate spurious correlations in practice, boosting the worst and mean group accuracy of a smile classifier by 54.88pp and 16.72pp, respectively. Showcased through extensive evaluation on three benchmark image datasets against six state-of-the-art methods,* CRAYON *achieves performance comparable or even superior to approaches that require more complex human annotations, and vastly outperforms methods that do not use human annotations.*

1. Introduction

Deep learning models have achieved remarkable performance, surpassing humans in image classification tasks [13]. However, recent advancements in deep learning interpretation have discovered that these models often make predictions based on irrelevant attributes [20, 51], resulting in biased performance [48, 53, 58, 59], poor generalization [34, 39], and reduced trustworthiness [16]. For example, a model trained on an imbalanced dataset to classify *smiling* and *non-smiling* faces bases its predictions on hair, which is irrelevant to smiling [20]. It is crucial to enhance these models so that their predictions rely on pertinent data features [34, 39]. Researchers address irrelevant attributions by mitigating spurious correlations among data attributes and labels in the training data through methods such as balancing the training dataset [26, 33, 63] or introducing new loss terms [17, 29, 33]. However, these methods make the **impractical assumption that developers have full knowledge of which attributes are spuriously correlated in advance** [2, 28]. Recent attempts to relax this assumption have yielded less satisfactory results [3, 28, 64].

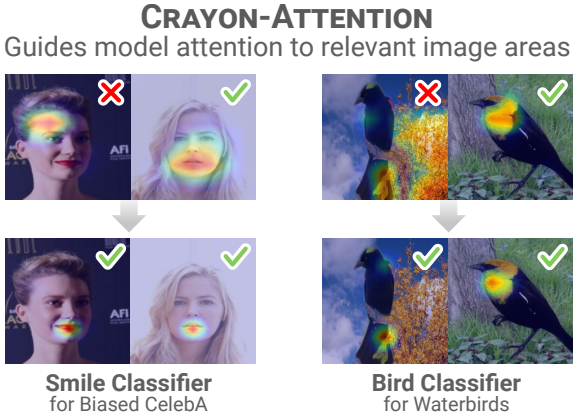


Figure 1. CRAYON-ATTENTION guides a model to attend to relevant image areas. CRAYON-ATTENTION redirects a smile classifier's occasional incorrect focus from a person's forehead to the mouth. Similarly, it shifts a bird classifier's attention away from irrelevant background towards the relevant bird body.

Therefore, it is essential to incorporate humans to effectively rectify a model's attention [45]. Some researchers have collected ground truth attention maps, which indicate where the model should or should not focus, and guide the models' saliency maps to resemble these ground truth

maps [9, 10, 40]. However, these methods require humans to provide an accurate ground truth map for every training data point, which can be extremely time-consuming and labor-intensive. Moreover, human-provided attention maps have their inherent limitations, such as imperfect annotation boundaries and inconsistency in data types between real-valued model-generated maps and binary human-provided maps with values of 0 or 1 [9].

To address the above research gaps, we present CRAYON (**C**orrecting **R**easoning with **A**nnotations of **Y**es **O**r **N**o), which makes the following contributions:

- **Yes-No Annotations as a Simple, Scalable and Practical Strategy to Guide Model Reasoning.** We introduce the major idea that simple yes-no annotations on model interpretations can offer an effective, scalable, and practical solution to address critical limitations of existing methods that heavily rely on laborious annotations. Our strategy empowers both classical and modern interpretation techniques to not only identify but also *rectify* model reasoning:
  1. **Guiding classic saliency map-based interpretations to highlight relevant image regions**. We propose CRAYON-ATTENTION to guide a model to attend to the relevant regions of images by using yes-no annotations on the relevance of saliency maps of each image. CRAYON-ATTENTION guides the model attention away from the areas highlighted in irrelevant maps, while preserving attention in the relevant saliency maps (Sec. 3.2).
  2. **Pruning irrelevant neurons identified by modern concept-based interpretations to remove their impact.** CRAYON-PRUNING identifies irrelevant neurons in the penultimate layer of a model by presenting the visual concepts responsible for highly activating each neuron. These irrelevant neurons are then pruned so that the model's predictions are not influenced by irrelevant concepts (Sec. 3.3).
- **CRAYON effectively mitigates spurious correlations with human annotations.** We showcase CRAYON's effectiveness, scalability, and practicality in refining model attention through yes-no annotations from 2,875 participants. Remarkably, CRAYON achieves near-peak performance with annotations for just 5% of the training data. CRAYON significantly outperforms methods that do not use human annotations, achieving up to 50.77pp higher worst group accuracy and 15.78pp higher mean group accuracy in a smile classifier. (Sec. 4).
- **Extensive evaluation on three benchmark image datasets against six state-of-the-art methods.** CRAYON achieves performance comparable to or even surpassing existing approaches that require complex annotations, and vastly outperforms methods that do not use human annotations (Sec. 5).

2. Related Work

2.1. Aligning Model Reasoning with Humans

Various approaches have emerged to extend model interpretation techniques beyond mere identification, aiming to rectify model attributions as well [8, 15, 25, 43, 44, 65]. To better align model attributions with human intuition, researchers have introduced interactive learning frameworks that incorporate human revisions of models [21, 45, 55]. The efficacy of human feedback in guiding model attributions has been showcased in diverse domains, such as natural language processing [23, 24, 52, 62, 68] and visual question answering [6, 36].

Concurrently, researchers have undertaken efforts to refine vision models by collecting human annotations for model attention. The RRR loss [40] was proposed to redirect MLP models away from regions annotated by humans as irrelevant, later extending its applicability to deeper CNN models [9–11]. CDEP [38] and SPIRE [35] aim to reduce the impact of irrelevant pixels by leveraging contextual decomposition and masking specific objects in images, respectively. Stammer et al. [54] refine models at both pixel and concept levels by disentangling concepts within an image. However, all these methods require humans to supply ground truth attention maps for each image, which can be prohibitively costly to obtain. To address this challenge, some progress has been made by introducing simpler alternatives such as scribble maps [49] and bounding boxes [37], which yet do not resolve the inherent limitations of human-provided attention maps [9]. Building upon these advancements, we further simplify human feedback to yes-no annotations on model interpretation results.

2.2. Mitigating Spurious Correlations

A lot of efforts have been dedicated to mitigating spurious correlations in deep learning models to enhance fairness [48, 53, 58], reliability [27], and generalizability [34]. Attributing spurious correlations to imbalances in training data [42], some researchers alleviate such issues by reweighting or subsampling training data [19, 26, 28, 33, 61]. However, challenges arise when a training dataset lacks spurious-free data. In response, some researchers opt to create balanced training datasets by collecting or generating additional instances [5, 12, 18, 22, 32, 63]. Yet, these approaches can be costly or impractical in real-world scenarios [46]. Various loss terms have been introduced to counter the impact of spurious correlations [17, 29, 51, 66]. However, most of these methods require prior knowledge about the attributes responsible for the correlations. Several methods have been proposed to address such limitation and have shown efficacy [3, 28, 64]. To achieve even higher

performance while overcoming all the aforementioned limitations, we incorporate simple yes-no human annotations.

3. Methods

3.1. Overview

CRAYON guides a trained model to base its predictions on relevant data areas by harnessing yes-no annotations, which pertain to the relevance of the rationale behind the model's predictions, as revealed through model interpretations. In this section, we describe (1) how simple yes-no annotations for classic interpretations based on saliency maps guide the model's attention to the relevant regions — we call this CRAYON-ATTENTION (Sec. 3.2) and (2) how we extend our idea to modern concept-based interpretations to prune the neurons activated by irrelevant visual concepts — we call this CRAYON-PRUNING (Sec. 3.3).

3.2. CRAYON-Attention: Guide Saliency Maps

Generating saliency maps stands as one of the most commonly employed model interpretation techniques [47, 50]. For a given model and its training data $\mathbf{x}_1, \ldots, \mathbf{x}_N$, the saliency map $M_{\mathbf{x}_n}$ highlights the regions within the image $\mathbf{x}_n$ that the model focuses on for its prediction. Once saliency maps are generated for all N training data points, we proceed to gather yes-no annotations regarding the relevance of each map to the prediction task. We denote the set of indices corresponding to training data with relevant and irrelevant maps as R and I, respectively.

To refine the model using the yes-no annotations, we introduce a loss function based on the energy loss [57]. For the data point $\mathbf{x}_n$ whose saliency map $M_{\mathbf{x}_n}$ highlights the relevant regions (i.e., $n \in R$), the model should generate similar saliency maps following the refinement. Hence, we formulate the loss function $\mathcal{L}_{rel,n}$ as follows:

$$\mathcal{L}_{rel,n} = \sum_{h=1}^{H} \sum_{w=1}^{W} [M'_{\mathbf{x}_n}]_{hw} (1 - [M_{\mathbf{x}_n}]_{hw}) \quad (1)$$

where H and W represent the height and width of the saliency maps, respectively, and $M'_{\mathbf{x}_n}$ is the saliency map for the model being trained and the data point $\mathbf{x}_n$. We clarify that $M_{\mathbf{x}_n}$ is the saliency map for the original model before refining, and $M'_{\mathbf{x}_n}$ is for the model being trained. For better stability of the loss function, we normalize both $M_{\mathbf{x}_n}$ and $M'_{\mathbf{x}_n}$, scaling their values between 0 and 1 by dividing each map by its maximum value.

For the data point $\mathbf{x}_n$ with irrelevant saliency map (i.e., $n \in I$), the model should attend to the regions that are not highlighted in the map $M_{\mathbf{x}_n}$. In this regard, we construct the loss function $\mathcal{L}_{irrel,n}$ as follows:

$$\mathcal{L}_{irrel,n} = \sum_{h=1}^{H} \sum_{w=1}^{W} [M'_{\mathbf{x}_n}]_{hw} [M_{\mathbf{x}_n}]_{hw} \quad (2)$$

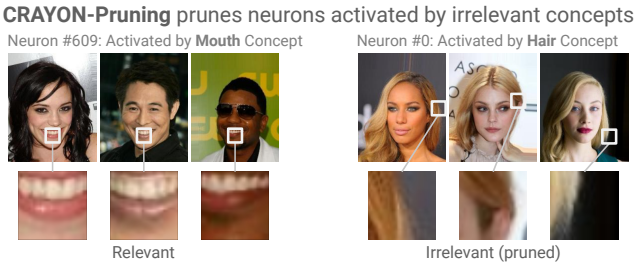


Figure 2. CRAYON-PRUNING prunes the neurons activated by irrelevant concepts in the penultimate layer and fine-tunes the last layer. For each neuron in the penultimate layer of a smile classifier, we generate image patches that summarize the visual concepts responsible for the activation of the neuron. **Left:** Among these neurons, neuron #609 is activated by *mouth* concept, which is relevant to smile classification. **Right:** Neuron #0, on the other hand, is activated by irrelevant *hair* concept and is pruned by CRAYON-PRUNING.

While guiding the model to attend to the right regions, we need to preserve the accuracy of the model's predictions. Therefore, we incorporate the prediction loss $\mathcal{L}_{pred,n}$ for the data point $\mathbf{x}_n$:

$$\mathcal{L}_{pred,n} = \sum_{k=1}^{K} -y_{nk} \log \hat{y}_{nk} \quad (3)$$

where y_{nk} is 1 if the label of the data $\mathbf{x}_n$ is k and 0 otherwise and $\hat{y}_{nk}$ is the probability of the data $\mathbf{x}_n$ being labeled as k computed by the model being trained.

Summing up the loss functions, we obtain the loss $\mathcal{L}_{att}$ that guides a model with yes-no annotations on saliency maps,

$$\mathcal{L}_{att} = \sum_{n=1}^{N} \mathcal{L}_{pred,n} + \alpha \sum_{n \in R} \mathcal{L}_{rel,n} + \beta \sum_{n \in I} \mathcal{L}_{irrel,n} \quad (4)$$

where α and β are the hyperparameters that control the weights of the loss terms.

3.3. CRAYON-Pruning: Prune Irrelevant Neurons

Neurons, also referred to as *channels*, in the penultimate layer of CNN models are known to be activated by specific high-level visual concepts in the input data [4, 31]. Based on this finding, a model interpretation method that summarizes the concepts responsible for a neuron's activation as a collection of image patches has recently been proposed [14]. These patches are generated by selecting the images that most highly activate the neuron and cropping out the corresponding region. For example, a neuron in the penultimate layer of a smile classifier would have patches corresponding to the *mouth* concept, indicating that the neuron's activation is attributed to the presence of a mouth

(Fig. 2, left). On the other hand, the patches of another neuron in the same model might indicate that its activation is attributed to *hair* (Fig. 2, right).

CRAYON-PRUNING identifies the neurons in the penultimate layer that are activated by irrelevant visual concepts by presenting the image patches of each neuron and collecting yes-no annotations on their relevance. For instance, in the smile classifier shown in Fig. 2, the neuron activated by the *mouth* concept is relevant while the neuron activated by the *hair* concept is irrelevant. We prune the irrelevant neurons and fine-tune the last fully-connected layer of the model to remove the effect of the irrelevant concept on the model's prediction. For this fine-tuning process, we use the prediction loss in Equation 3.

4. Evaluation with Human Annotations

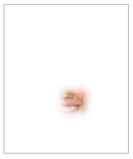
To demonstrate that CRAYON provides effective, scalable, and practical solutions for rectifying model attention, we collect yes-no human annotations with 2,875 participants on Amazon Mechanical Turk (MTurk).

4.1. Experimental Setup

4.1.1 Dataset

We train a smile classifier using the Biased CelebA dataset [20]. Its training set demonstrates a spurious correlation between *hair color* and *smiling* attributes. Specifically, most images of individuals with *black hair* exhibit *smiling* expressions, while those with *blond hair* are mostly *not smiling*. This correlation causes the smile classifier to occasionally make incorrect associations between the predictions and hair color. We provide details about the dataset and model training in the Appendix.

4.1.2 Human Annotations



For CRAYON-ATTENTION, we visualize Grad-CAM of each training image, where regions that receive higher Grad-CAM attention are depicted more visible, while regions with lower attention appear more transparent. Our visualization design improves over the conventional method that highlights model-attended areas in red [1]. From our pilot study, we discovered that the conventional red highlights could obscure image contents, making it hard for participants to assess relevance.

We show each visualization to three participants and ask yes-no questions about whether they can determine if the person in the image is smiling. An image that receives unanimous "*yes*" responses from all three participants (i.e., can determine smiling) is annotated as having a *relevant* saliency map. Conversely, if at least two participants respond with "*no*," the image is annotated as having an *irrelevant* map. Images that do not meet these criteria are excluded from guiding model attention due to the ambiguity of the relevance of their Grad-CAMs. We opted not to require three "no" responses to annotate a map as *irrelevant* due to our observation that some participants consider a map relevant even when irrelevant areas are prominently highlighted, as long as there is some attention on any facial part, such as the nose, forehead, or ear. Hence, we relaxed the requirement to a minimum of two "*no*" responses.

For CRAYON-PRUNING, for each neuron in the model's penultimate layer, we generate three image patches representing the visual concepts responsible for its activation. Each patch is shown to an individual participant, and we inquire whether the patch implies the presence of a smile. We annotate a neuron as *relevant* if all three patches are labeled to be indicative of smiling; otherwise, we annotate it as *irrelevant*.

4.1.3 Settings for CRAYON

CRAYON-ATTENTION fine-tunes the smile classifier for 10 epochs with a learning rate of 1e-5 and the hyperparameters α and β in Equation 4 of 10000 and 100, respectively. CRAYON-PRUNING prunes irrelevant neurons in the penultimate layer and trains the last fully connected layer for 50 epochs with a learning rate of 5e-6. Additionally, we evaluate CRAYON-PRUNING+ATTENTION, which prunes irrelevant neurons in the penultimate layer and then fine-tunes the model using the loss function in Equation 4[1]. CRAYON-PRUNING+ATTENTION trains the model for 10 epochs with a learning rate of 1e-5, α of 1000, and β of 20. In all methods, we use a batch size of 64 and the Adam optimizer with a weight decay of 1e-4.

4.1.4 Compared Methods

To underscore the significance of human annotations, we compare CRAYON with three state-of-the-art methods addressing spurious correlations without human annotations:

- *JtT* [28] upweights the loss of training data points misclassified by the original model.
- *MaskTune* [3] guides the model to learn diverse features by masking regions highly attended to by the original model.
- *CnC* [64] leverages contrastive learning to closely locate representations of data points with the same class labels but different spurious features, which are identified using the original model.

[1]Neuron annotations must be applied before using attention annotations, as the neuron annotations stem from the original non-fine-tuned model. If we fine-tune the model with CRAYON-ATTENTION, the concepts detected by each neuron will change and invalidate the existing neuron annotations.

Table 1. CRAYON effectively mitigates spurious correlations in a smile classifier using yes-no human annotations. Both CRAYON-ATTENTION and CRAYON-PRUNING+ATTENTION with attention annotations for just 1,000 data points nearly reach their peak *mean group accuracy* (MGA) performance. We run each method five times with different random seeds and report the average values of the WGA (*worst group accuracy*) and MGA.

Method	#Annot.	WGA	MGA
Original	-	32.60	73.71
CRAYON-ATTENTION	20,200	82.27	89.77
CRAYON-PRUNING	2,048	68.82	86.66
CRAYON-PRUNING+ATTENTION	22,248	**87.48**	**90.43**
CRAYON-ATTENTION	1,000	80.64	88.98
CRAYON-PRUNING+ATTENTION	3,048	82.91	89.80
JtT [28]	0	36.71	74.65
MaskTune [3]	0	37.72	78.85
CnC [64]	0	37.76	75.34
ERM	0	32.86	73.01

To ensure that the mitigation results are not simply due to extended training, we also compare with the naive *empirical risk minimization (ERM)* [41] approach that minimizes the classification loss by training the model for more epochs.

The Appendix describes the methods' training details. We could not compare CRAYON with the methods that leverage complex human annotations (attention maps, bounding boxes) because they either require proprietary apparatus [9, 10] or have not been evaluated with actual human annotators [37, 40].

4.2. Results: Practical Effectiveness of CRAYON

Table 1 shows the performance of our methods using the collected human annotations. In accordance with the established practice in the literature of spurious correlation research [41], we employ *worst group accuracy* (WGA) and *mean group accuracy* (MGA) as the evaluation metrics. Specifically, we first evaluate the model accuracy for each of the four *attribute groups*: *black hair + smiling*, *blond hair + not smiling*, *black hair + not smiling*, and *blond hair + smiling*. We then compute the minimum and mean accuracy values across these groups and denote them as *worst group accuracy* (WGA) and *mean group accuracy* (MGA), respectively. To ensure robustness, we run each method five times with different random seeds and report the average of the WGA and MGA values in Table 1; the Appendix provides the standard deviations.

Overall, all CRAYON approaches effectively guide the smile classifier to rely on relevant regions, not spurious correlations. Comparing to the unrefined original model (Row 1), CRAYON-ATTENTION (Row 2) substantially enhances WGA by 49.67 percentage points (pp), raising it from 32.60% to 82.27%; and MGA by 16.06pp, increasing it from 73.71% to 89.77%. CRAYON-PRUNING (Row 3) also demonstrates improvements, achieving a 36.22pp increase in WGA and a 12.95pp increase in MGA. CRAYON-PRUNING+ATTENTION (Row 4), which combines both *attention* and *pruning* approaches, elevates both WGA and MGA beyond the capabilities of each individual approach, elevating WGA to 87.48% and MGA to 90.43%.

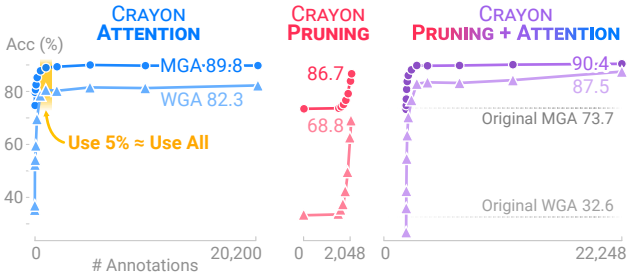


Figure 3. Even with annotations available for just 5% of the training data points, both CRAYON-ATTENTION and CRAYON-PRUNING+ATTENTION nearly reach their peak performance. CRAYON-PRUNING realizes its full effectiveness when annotations are provided for most neurons in the penultimate layer.

Comparing CRAYON with methods that do not leverage human annotations (Table 1: Row 7-9) underscores the importance of incorporating humans in guiding model attention to relevant regions. JtT, which upweights the training data misclassified by the original model, exhibits marginal enhancement as the model misclassifies only 12 out of 20,200 training data points. Similarly, CnC, which uses misclassified data as positive and negative samples for contrastive learning, achieves only limited improvement. The constrained efficacy of MaskTune is attributed to the limitation of its strategy, which redirects model attention to alternative areas even for the data with relevant attention maps.

4.3. Varying Number of Annotations

We evaluate how the number of annotations n impacts CRAYON's performance. For CRAYON-ATTENTION, we randomly sample n images from the training set of the Biased CelebA dataset and compute both $\mathcal{L}_{rel}$ and $\mathcal{L}_{irrel}$ for these n images along with their annotations. For CRAYON-PRUNING, we randomly sample n neurons from the penultimate layer and prune the *irrelevant* neurons within this sampled group of n. We additionally investigate CRAYON-PRUNING+ATTENTION, where we use all 2,048 annotations for neuron relevance while varying the number of annotations for model attention[2]. For each n value, we run the

[2]We elect to focus on varying the number of attention annotations based on our observation that almost all pruning annotations need to be used for CRAYON-PRUNING to be fully effective.

method five times with different random seeds and report the average MGA and WGA values in Fig. 3.

Overall, the performance of all CRAYON methods improves as the number of annotations increases. Notably, CRAYON-ATTENTION and CRAYON-PRUNING+ATTENTION are effective even with a limited number of annotations, achieving nearly peak performance when annotations are available for only 5% of the training data points (Table 1: Row 5,6). Specifically, with yes-no annotations for 1,000 out of 20,200 training data points, CRAYON-ATTENTION enhances the WGA and MGA to 80.64% and 88.98%, respectively, while CRAYON-PRUNING+ATTENTION enhances the WGA and MGA to 82.91% and 89.80%, respectively. These values are only marginally lower than the performance achieved with annotations for all training data points. In contrast, the performance of CRAYON-PRUNING is constrained unless a substantial portion of neurons is annotated, underscoring the importance of acquiring annotations for all neurons in the penultimate layer.

5. Comparing CRAYON with Existing Methods

As most existing methods have not been evaluated with humans, we use machine-generated annotations to compare CRAYON with them.

5.1. Experimental Setup

5.1.1 Datasets

In addition to the Biased CelebA dataset (Sec. 4.1.1), we include two additional benchmark datasets, Waterbirds [41] and Backgrounds Challenge [60]. The Appendix describes more details about the datasets.

Waterbirds. The Waterbirds dataset combines bird photographs [56] with backgrounds [67] so that waterbirds and landbirds appear more frequently in water (e.g., lake) and land (e.g., forest) backgrounds, respectively. Bird classifiers trained on this dataset would classify waterbirds and landbirds based on the backgrounds rather than the bird bodies.

Backgrounds Challenge. Backgrounds Challenge addresses the problem that classifiers trained on the ImageNet [7] dataset often inappropriately base their predictions on image backgrounds, rather than foreground objects. Aiming to correct models to base their predictions on the foreground objects, the challenge introduces the ImageNet-9 (IN-9) dataset, a subset of ImageNet with nine coarse-grained classes. As it is hard to specify groups for image backgrounds, the Backgrounds Challenge assesses whether a model grounds its predictions in relevant foreground areas using datasets created by transforming image backgrounds:

- *Only-FG* dataset removes the image backgrounds by coloring them in black.
- *Mixed-Rand* dataset shuffles the background across all images in the IN-9 test dataset to decorrelate background and foreground.

Accordingly, in our work, we use accuracies on the Only-FG and Mixed-Rand datasets as our metrics; higher accuracy values on these datasets indicate that the model appropriately attends to foreground objects.

5.1.2 Compared Methods

Original Models. For consistent results, we train five models for the Biased CelebA and Waterbirds datasets; the five models differ only in the random seed, which would result in five different models. For the Backgrounds Challenge, we use the model released on the challenge repository. We refer to all these models as the *original* models, meaning no mitigation has been applied.

State-of-the-art Methods. In addition to JtT, MaskTune, CnC, and ERM (Sec. 4.1.4), we examine three state-of-the-art techniques that require complex annotations[3]:

- *RRR* [40] collects ground truth maps that annotate irrelevant pixels in the images and guides the model not to attend to the irrelevant pixels.
- *GRADIA* [10] identifies the images for which the original model generates irrelevant saliency maps or incorrectly predicts, collects ground truth maps of relevant pixels, and aligns the model's attention with the collected maps.
- *Bounding Box* [37] collects bounding boxes that cover the relevant regions of each image and guides the model to keep its attention within the boxes.

The Appendix describes training details for these methods.

5.1.3 Settings for CRAYON

For the Biased CelebA dataset, we run all our methods with the same setup described in Sec. 4.1.3. For the Waterbirds dataset, CRAYON-ATTENTION and CRAYON-PRUNING are trained for 10 epochs using the Adam optimizer with a learning rate of 5e-5 and a batch size of 128. CRAYON-PRUNING+ATTENTION uses the same settings as these methods, except for the learning rate, which is set to 5e-6. We set α and β for as 500 and 25, respectively. For the Backgrounds Challenge, we train all CRAYON methods for 10 epochs with a batch size of 256 using the SGD optimizer with a learning rate of 5e-6 and a weight decay of 1e-1. We set α as 0.1 and β as 0.25 for CRAYON-ATTENTION and α as 0.2 and β as 0.5 for CRAYON-PRUNING+ATTENTION.

[3] We tried RES [9] on our datasets and determined that it was computationally prohibitive. The algorithm did not finish one iteration even after 3 hours on an NVIDIA A6000 GPU; 2,205 iterations are needed for the Biased CelebA dataset.

Table 2. CRAYON uses simple yes-no annotations to effectively mitigate spurious correlations in the models for *Biased CelebA*, *Waterbirds*, and *Backgrounds Challenge* datasets (abbreviated as *Backgrounds*; comprising *Only-FG* (OF), and *Mixed-Rand* (MR) as described in Sec. 5.1.1.). CRAYON achieves performance comparable or superior to existing state-of-the-art approaches that require more complex annotations, securing either the **best** or ***second-best*** accuracies. WGA denotes *worst group accuracy*; MGA denotes *mean group accuracy*.

Method	Annotation	Biased CelebA		Waterbirds		Backgrounds	
		WGA	MGA	WGA	MGA	OF	MR
Original	-	28.29	71.73	25.31	67.84	85.75	78.27
CRAYON-ATTENTION	Yes-No	***74.41***	***87.48***	47.91	74.28	86.41	81.79
CRAYON-PRUNING	Yes-No	69.30	84.60	49.75	76.39	84.91	79.93
CRAYON-PRUNING+ATTENTION	Yes-No	**79.44**	**88.03**	**60.82**	***78.15***	**86.87**	***82.63***
RRR [40]	Map	50.86	79.47	41.03	76.05	86.67	82.12
GradIA [10]	Yes-No, Map	41.80	77.36	44.39	76.54	***86.78***	81.29
Bounding Box [37]	Bounding Box	74.28	86.97	***51.28***	**78.98**	86.66	**82.93**
JtT [28]	-	37.12	74.41	33.64	72.65	85.85	78.38
MaskTune [3]	-	32.11	77.34	32.11	73.63	84.62	78.25
CnC [64]	-	34.70	73.28	37.16	73.08	85.14	78.49
ERM	-	30.40	71.33	31.71	71.79	85.91	77.97

5.1.4 Machine-Generated Annotations

For a fair and scalable comparison, we algorithmically generate annotations for CRAYON and the compared methods. For the Biased CelebA dataset, in which all the images are aligned based on the positions of two eyes [30], we annotate the relevance of Grad-CAM of an image as *yes* if the highest attention value falls on the mouth or eyes areas, while less than half of this value is outside the central face, and *no* if the highest attention value is located elsewhere and the mouth area gets less than 80% of it; otherwise, we do not use the image for guiding model attention. To identify relevant neurons for CRAYON-PRUNING, we forward all training data through the model and collect 20 image patches for each neuron in the penultimate layer that summarize the visual concepts responsible for the neuron's activation. We annotate a neuron's relevance as *yes* if all its patches contain mouths and *no* otherwise. We generate the ground truth attention maps and bounding boxes required by the compared methods to cover mouths.

For the Waterbirds dataset, which provides segmentation maps of bird bodies for each image, the relevance of Grad-CAM of an image is labeled as *yes* if its segmentation map and Grad-CAM overlap more than 70% and *no* if they overlap less than 30%; otherwise, the image is not used for guiding model attention. For CRAYON-PRUNING, we collect 20 image patches for each neuron in the penultimate layer and label the neuron's relevance as *yes* if more than 70% of the patches pertain to bird bodies and *no* otherwise. The provided segmentation maps serve as the ground truth attention maps, and we draw boxes surrounding the maps to generate the bounding boxes for the competitors.

For the Backgrounds Challenge, we employ the provided segmentation maps for foreground objects. An image's Grad-CAM is labeled as *relevant* if its overlap with the image's segmentation map exceeds 50%, and *irrelevant* if the overlap is less than 20%. For CRAYON-PRUNING, a neuron is labeled as *relevant* if over 95% of its 20 image patches pertain to foreground objects, and *irrelevant* otherwise.

5.2. Results: Comparable or Outperforms SOTA

Table 2 compares our methods with state-of-the-art methods on the Biased CelebA, Waterbirds, and Backgrounds Challenge. For each original model, we conduct experiments with five different random seeds and report the average of the evaluation results.

Overall, our methods based on yes-no annotations achieve comparable or even better performance than other competitors that require more complex human intervention. For the Biased CelebA dataset, comparing with the original models (Row 1) demonstrates the effectiveness of both CRAYON-ATTENTION and CRAYON-PRUNING in mitigating spurious correlations, providing a significant boost to the *worst group accuracy* (WGA) by 46.12 percent points (pp) (74.41% for CRAYON-ATTENTION vs 28.29% for original) and the *mean group accuracy* (MGA) by 15.75pp (87.48% for CRAYON-ATTENTION vs 71.73% for original). Combining these two approaches into CRAYON-PRUNING+ATTENTION brings further improvement, achieving WGA and MGA values of 79.91% and 87.98%. CRAYON-ATTENTION and CRAYON-PRUNING+ATTENTION outperform all competitors that exploit ground truth maps and bounding boxes with richer in-

formation, in terms of both WGA and MGA. We attribute the superiority of our method to the limitations inherent in binary ground truth maps and boxes. To be specific, the maps and boxes are represented as binary values of 0 and 1, while the model-generated saliency maps consist of continuous real numbers. This inconsistency degrades the performance of the model attention guidance [9]. CRAYON resolves the challenge by using the saliency maps of the original model instead of binary ground truth.

CRAYON-PRUNING+ATTENTION demonstrates the effectiveness in mitigating spurious correlations also for the Waterbirds dataset, enhancing the original models' WGA from 25.31% to 60.82% and MGA from 67.84% to 78.15%. While the individual performances of CRAYON-PRUNING and CRAYON-ATTENTION are lower than some competitors, combining the two methods to CRAYON-PRUNING+ATTENTION complements their limitations, achieving the best WGA and the second-best MGA values among all the compared methods.

For the Backgrounds Challenge, CRAYON-PRUNING+ATTENTION achieves the best accuracy on the Only-FG dataset and the second-best accuracy on the Mixed-Rand dataset among all the compared methods, demonstrating its effectiveness in multi-label classification tasks. It boosts the accuracy on the Only-FG dataset from 85.75% to 86.87% and on the Mixed-Rand dataset from 78.27% to 82.63%. The underperformance of JtT, MaskTune, and CnC that do not use any annotations for attention guidance underscores the importance of incorporating human annotations in mitigating spurious correlations. The limited performance of JtT and CnC are attributed to the small number (only 2%) of the training data points misclassified by the original model. MaskTune's lower-than-original performance suggests its potential reliance on a large number of training data points, as a higher performance was reported only for the setting where four times of training data points were used [3].

We also qualitatively evaluate CRAYON-ATTENTION, as shown in Figure 1. For a model that irrelevantly attends to the *forehead* of an image from the Biased CelebA dataset, CRAYON-ATTENTION fixes its attention to the *mouth*. Similarly, CRAYON-ATTENTION rectifies the attention of a bird classifier that initially focuses on the background of a bird image to the *bird's body*.

5.3. Ablation Study

We conduct an ablation study to investigate the impact of the two proposed loss terms, $\mathcal{L}_{rel}$ and $\mathcal{L}_{irrel}$, on CRAYON-ATTENTION's performance. We deactivate one of the two loss terms by setting either α or β in Equation 4 to 0. The results are shown in Table 3.

Overall, removing either of the two loss terms degrades the performance of CRAYON-ATTENTION, underscoring the contributions of both $\mathcal{L}_{rel}$ and $\mathcal{L}_{irrel}$ in guiding model attention. When we deactivate $\mathcal{L}_{irrel}$ and rely solely on $\mathcal{L}_{rel}$ (Row 2), WGA and MGA experience declines of 12.32pp (from 74.41% to 62.09%) and 3.36pp (from 87.48% to 84.12%) for the Biased CelebA dataset and 11.43pp (from 47.91% to 36.48%) and 1.54pp (from 74.28% to 72.74%) for the Waterbirds dataset. For the Backgrounds Challenge, the accuracies on the Only-FG and Mixed-Rand datasets decrease from 86.41% to 86.27% and from 81.79% to 81.61%, respectively.

Table 3. Ablation study demonstrates the impact of two loss terms, $\mathcal{L}_{rel}$ and $\mathcal{L}_{irrel}$, on the performance of CRAYON-ATTENTION (row 1). $\mathcal{L}_{rel}$ notably enhances overall performance by offering substantial guidance on the model's attention (row 2). This guidance is significantly stronger compared to the effect of the loss term for irrelevant annotations, $\mathcal{L}_{irrel}$, which directs attention away from irrelevant image regions (row 3). BG stands for Backgrounds Challenge, OF for Only-FG, and MR for Mixed-Rand.

$\mathcal{L}_{rel}$	$\mathcal{L}_{irrel}$	**Biased CelebA**		**Waterbirds**		**BG**	
		WGA	MGA	WGA	MGA	OF	MR
✓	✓	74.41	87.48	47.91	74.28	86.41	81.79
✓		62.09	84.12	36.48	72.74	86.27	81.61
	✓	42.79	75.62	32.87	71.85	86.14	79.81

Excluding $\mathcal{L}_{rel}$ from $\mathcal{L}_{att}$ (Row 3) also significantly impairs the performance. For the Biased CelebA dataset, WGA decreases by 31.62pp (from 74.41% to 42.79%) and MGA by 11.86pp (from 87.48% to 75.62%), and for the Waterbirds dataset, WGA declines by 15.04pp (from 47.91% to 32.87%) and MGA by 2.43pp (from 74.28% to 71.85%). Likewise, for the Backgrounds Challenge, the accuracy on the Only-FG dataset decreases by 0.27pp (from 86.41% to 86.14%) and the accuracy on the Mixed-Rand dataset by 1.98pp (from 81.79% to 79.81%). These results show that $\mathcal{L}_{rel}$ plays a significant role in guiding model attention, while $\mathcal{L}_{irrel}$ provides additional guidance by directing attention away from irrelevant image areas.

6. Conclusion

We propose CRAYON, which mitigates spurious correlations in image classifiers using simple yes-no annotations. CRAYON-ATTENTION collects yes-no annotations on the relevance of saliency maps and refines models to attend to the relevant areas, while CRAYON-PRUNING identifies and prunes irrelevant neurons in the penultimate layer of the models based on the yes-no annotations on the relevance of the neuron activation. Evaluation with the large-scale human annotations demonstrate CRAYON's practicality, scaliability, and effectiveness in mitigating spurious correlations. Evaluation against state-of-the-art methods shows that CRAYON achieves comparable or even better perfor-

mance than the competitors that require much more complex annotations.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 4

[2] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International conference on learning representations*, 2021. 1

[3] Saeid Asgari, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Tran, Ali Mahdavi-Amiri, and Ghassan Hamarneh. Masktune: Mitigating spurious correlations by forcing to explore. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 4, 5, 7, 8

[4] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 3

[5] Ming-Chang Chiu, Pin-Yu Chen, and Xuezhe Ma. Better may not be fairer: Can data augmentation mitigate subgroup degradation? *arXiv preprint arXiv:2212.08649*, 2022. 2

[6] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163: 90–100, 2017. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Yuyang Gao, Tong Sun, Rishab Bhatt, Dazhou Yu, Sungsoo Hong, and Liang Zhao. Gnes: Learning to explain graph neural networks. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 131–140. IEEE, 2021. 2

[9] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Zhao Liang. Res: A robust framework for guiding visual explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 432–442, 2022. 2, 5, 6, 8

[10] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction*, 6 (CSCW2):1–28, 2022. 2, 5, 6, 7

[11] Misgina Tsighe Hagos, Kathleen M Curran, and Brian Mac Namee. Identifying spurious correlations and correcting them with an explanation-based learning. *arXiv preprint arXiv:2211.08285*, 2022. 2

[12] Zongbo Han, Zhipeng Liang, Fan Yang, Liu Liu, Lanqing Li, Yatao Bian, Peilin Zhao, Bingzhe Wu, Changqing Zhang, and Jianhua Yao. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *arXiv preprint arXiv:2209.08928*, 2022. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 1

[14] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2020. 3

[15] Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34:26726–26739, 2021. 2

[16] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, 55(2):1–38, 2022. 1

[17] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 1, 2

[18] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14992–15001, 2021. 2

[19] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 2

[20] Arvind Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In *British Machine Vision Conference (BMVC)*, 2021. 1, 4

[21] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015. 2

[22] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34:25123–25133, 2021. 2

[23] Piyawat Lertvittayakumjorn and Francesca Toni. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 2021. 2

[24] Piyawat Lertvittayakumjorn, Lucia Specia, and Francesca Toni. Find: human-in-the-loop debugging deep text classifiers. *arXiv preprint arXiv:2010.04987*, 2020. 2

[25] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9215–9223, 2018. 2

[26] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 9572–9581, 2019. 1, 2

[27] Weixin Liang, Girmaw Abebe Tadesse, Daniel Ho, L Fei-Fei, Matei Zaharia, Ce Zhang, and James Zou. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4(8):669–677, 2022. 2

[28] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 1, 2, 4, 5, 7

[29] Sheng Liu, Xu Zhang, Nitesh Sekhar, Yue Wu, Prateek Singhal, and Carlos Fernandez-Granda. Avoiding spurious correlations via logit correction. *arXiv preprint arXiv:2212.01433*, 2022. 1, 2

[30] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 7

[31] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120:233–255, 2016. 3

[32] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, pages 6927–6937. PMLR, 2020. 2

[33] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 1, 2

[34] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020. 1, 2

[35] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021. 2

[36] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[37] Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Using explanations to guide models. *arXiv preprint arXiv:2303.11932*, 2023. 2, 5, 6, 7

[38] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pages 8116–8126. PMLR, 2020. 2

[39] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10):896–904, 2021. 1

[40] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017. 2, 5, 6, 7

[41] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 5, 6

[42] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020. 2

[43] Gobinda Saha and Kaushik Roy. Saliency guided experience packing for replay in continual learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5273–5283, 2023. 2

[44] Johannes Schneider and Michalis Vlachos. Reflective-net: Learning from explanations. *Data Mining and Knowledge Discovery*, pages 1–22, 2023. 2

[45] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020. 1, 2

[46] Roy Schwartz and Gabriel Stanovsky. On the limitations of dataset balancing: The lost battle against spurious correlations. *arXiv preprint arXiv:2204.12708*, 2022. 2

[47] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[48] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2180–2188, 2022. 1, 2

[49] Haifeng Shen, Kewen Liao, Zhibin Liao, Job Doornberg, Maoying Qiao, Anton Van Den Hengel, and Johan W Verjans. Human-ai interactive and continuous sensemaking: A case study of image classification using scribble attention maps. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2021. 2

[50] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 3

[51] Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021. 1, 2

[52] Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. Improving natural language processing tasks with human gaze-guided neural attention. *Advances in Neural Information Processing Systems*, 33:6327–6341, 2020. 2

[53] Ramya Srinivasan and Ajay Chander. Biases in ai systems. *Communications of the ACM*, 64(8):44–49, 2021. 1, 2

[54] Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3619–3629, 2021. 2

[55] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019. 2

[56] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-UCSD Birds-200-2011 (CUB-200-2011). Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

[57] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 3

[58] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8919–8928, 2020. 1, 2

[59] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2022. 1

[60] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 6

[61] Yihao Xue, Ali Payani, Yu Yang, and Baharan Mirzasoleiman. Eliminating spurious correlations from pre-trained models via data mixing. *arXiv preprint arXiv:2305.14521*, 2023. 2

[62] Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. Refining language models with compositional explanations. *Advances in Neural Information Processing Systems*, 34:8954–8967, 2021. 2

[63] Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pages 25407–25437. PMLR, 2022. 1, 2

[64] Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022. 1, 2, 4, 5, 7

[65] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 349–357. IEEE, 2019. 2

[66] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Causaladv: Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021. 2

[67] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6

[68] Hugo Zylberajch, Piyawat Lertvittayakumjorn, and Francesca Toni. Hildif: Interactive debugging of nli models using influence functions. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 1–6, 2021. 2