COUNTEREXAMPLE-GUIDED SHORTCUT SUPPRESSION WITHOUT SHORTCUT LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spurious correlations (SC) are correlated features in datasets that allow for better benchmark predictions, but are not robust to domain change and so might harm generalisation when relied upon. SC can be unintended, such as background bias in object classification and diagnosis markings in medical images, but they have also been used intentionally, such as in dataset protection. Traditional methods for addressing SC often require expert knowledge and task-specific operations, leading to recent research to detect and correct SC without putting shortcut labels in the training objective. In this paper, we follow this line of research, both when SC is unintended and when it is deliberately injected. We propose a method that does not use shortcut labels during training; instead, we create label-conflicting counterexample mixtures so shortcut-following predictions become costly. We study this as a clean-counterexample supervision regime: biased training data and a small clean counterexample pool are available, but shortcut labels are not used in the training objective. In a controlled hidden-watermark setting, the method suppresses watermark-only prediction to chance, validating the intended mechanism. On augmented Waterbirds, where the shortcut is a replaceable background cue, we compare against previous methods, attaining higher worst-group accuracy without domain labels in the objective. Additional CelebA results test a visually entangled attribute shortcut and show weaker, more variable gains. These results suggest that clean counterexamples can provide useful weak supervision for shortcut attenuation, with effectiveness governed by shortcut separability and counterexample quality.

1 INTRODUCTION

Spurious correlations (SC) are predictive features, readily learned from training data by deep learning models but generalise poorly to Out-of-Distribution (OOD) or test settings (Beery et al., 2018). This tendency to associate spurious patterns to labels is called shortcut learning (Geirhos et al., 2020). A rich literature aims to mitigate the effects of SC (Kim et al., 2019; Alvi et al., 2019), while some works exploit this trait, e.g. for adversarial attacks (Szegedy et al., 2014; Jang et al., 2024), or to improve performance via overfitting (Kent, 2021). We believe it important for machine learning developers to be able to choose what their models learn in addition to achieving good performance. For this paper, we investigate when a small clean counterexample set without shortcut labels can attenuate shortcut reliance. Our goal is not to replace group-supervised debiasing under stronger assumptions, but to isolate whether clean counterexamples provide a useful shortcut-conflicting signal under the same information budget. (Yao et al., 2022) proposes LISA, which uses domain labels on the training set to select MixUp pairs. (Bai et al., 2025) proposes NCM, which uses a small set of same-class, shortcut-swapped pairs to constrain the classifier away from pair-difference directions. We instead assume a small class-labeled pool of shortcut-free images and create label-conflicting mixtures, without using shortcut labels in the training objective.

As opposed to spurious correlations, non-spurious or invariant correlations are those that remain predictive of the target even after a shift in distribution over input variables. In this work, we use causal reasoning as a design lens for constructing counterexamples that make shortcut-following predictions costly.

Firstly, we analyse a simple case of spurious correlations: a hidden watermark dataset Jang et al. (2024). A class of dataset watermarking techniques exists relying on inserting controlled SC between

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

data and label. We analyse the interplay between these fields, showing how in such a simple case, there is a simple yet effective method allowing a model to avoid its reliance on spurious watermark signals. We find that clean data alone may not break shortcut reliance; the key training signal is that the mixed example makes a shortcut-following prediction incorrect.

Previous efforts to mitigate the effects of SC often require expert knowledge, such as access to labelled SC data (Alvi et al., 2019). However in practice, there are circumstances when the presence of SC is known, but further information on SC is difficult to identify; this is a key fact in dataset watermarking, which this work aims to address. We ask: 'Can clean counterexamples be used to debias unwanted correlations from a model without explicit labels of the bias?' To do so, we present a novel method that uses a *counterexample dataset*, a small clean dataset absent of a bias present in the larger dataset. Our method allows a model to train on biased data, while reducing reliance on SC absent from counterexamples. This is done via mixing counterexamples with biased data of a different class, balancing the statistical occurrences of biases such that they are no longer predictive. In the terminology of this paper, these are label-conflicting counterexample mixtures: the shortcut cue points toward one label while the training target points toward another.

Then we extend the principle to more realistic settings. Here, clean data can take the shape of a simple form of data augmentation, or performing more demanding data collection processes on a smaller scale. For example, collecting specimen photographs for birds in blank backgrounds. We use augmented Waterbirds as the main realistic setting, where the shortcut is a replaceable background cue, and CelebA as an entangled-attribute stress test. For "clean" data on waterbirds, we extend the Waterbirds dataset by background augmentation, creating counterexamples that are not shortcut-aligned. We find that the method improves worst-group accuracy and shortcut-conflicting accuracy most clearly when the counterexamples preserve task evidence while breaking the shortcut-label link.

Contribution. In this work our contribution is three-fold:

1. We formulate a weak-supervision paradigm for shortcut attenuation using counterexamples without using shortcut labels in the training objective.
2. We define label-conflicting counterexample mixing, where clean counterexamples from different classes are mixed with biased samples so shortcut-following predictions are penalised.
3. We provide a regime analysis across hidden watermarks, augmented Waterbirds, and CelebA. Under fixed epoch-15 evaluation on Waterbirds, label-conflicting MixUp improves average accuracy by 5.10 points and WGA by 17.47 points over simply adding the same clean counterexample pool. And outperforms LISA and NCM on these metrics.

2 Related Work

Shortcut learning and benchmarks. Shortcut learning occurs when models rely on predictive but non-robust cues that fail under domain or group shift (Geirhos et al., 2020; Beery et al., 2018). This motivates evaluation beyond average accuracy, especially on benchmarks where worst-group performance exposes reliance on background, demographic, or annotation artefacts. Waterbirds and CelebA are common image settings for this question, while GroupDRO and WILDS formalise group and distribution-shift evaluation protocols (Sagawa et al., 2020; Koh et al., 2021; Liu et al., 2015).

Dataset watermarking as controlled shortcut injection. Dataset watermarking aims to protect datasets against misuse, e.g. for intellectual property reasons, by artificially injecting SC into datasets. That is, they alter the data points conditioned on the corresponding class, such that learning models exploit these patterns to make predictions. Dataset owners can detect these learned relationships, allowing them to detect unauthorised data use. For our work, we focus on Jang et al. (2024), an architecture agnostic watermarking method. Jang et al. (2024) is designed to be well hidden, and the end-user has no access to the details of the introduced bias or labels for it. This controlled injection allows SC to be compact, hidden, and class-conditional, making it a useful controlled mechanism case for our work. It serves as a starting point before we jump into more realistic settings where SC is less structurally separable.

Mixing, augmentation, and counterfactual data. Data augmentation is a well explored avenue for tackling generalisation. Related literature uses data augmentation to improve OOD robustness (Ilse

et al., 2021; Balashankar et al., 2023), and works like Zhang et al. (2021) investigate how popular data augmentation techniques improve generalisation. Mixup (Zhang et al., 2018) and Cutmix (Yun et al., 2019) are related image augmentation methods that mix information from two images into one. Counterfactual data augmentation methods such as (Balashankar et al., 2023; Sauer & Geiger, 2021) also follow a line similar to ours, generating OOD augmentations to encourage causal learning.

LISA, NCM, and label-conflicting mixing. We identify closest existing methods to ours, which we provide experimental comparisons to: LISA (Yao et al., 2022) uses domain labels on the training set to select MixUp pairs and supervises with soft interpolated labels. NCM (Bai et al., 2025) does not mix: it uses same-class, shortcut-swapped pairs to estimate spurious directions and constrains the classifier to be orthogonal to them. However, group specific labels can be expensive to obtain, and are not always available. We instead mix a biased image with an other-class off-axis clean image and keep a hard clean label, without using shortcut labels in the training objective.

Group-aware and group-inference debiasing. Many strong spurious-correlation methods use group labels, infer latent groups, or modify a trained model. GroupDRO optimises worst-group risk with group annotations (Sagawa et al., 2020); JTT, EIIL, LfF, CNC, and AFR use model errors, inferred environments, biased classifiers, contrastive pairs, or feature reweighting to improve robustness without explicit group labels in the main training objective (Liu et al., 2021; Creager et al., 2021; Nam et al., 2020; Zhang et al., 2022; Qiu et al., 2023). Recently, there has also been non-counterfactual-based methods such as Zheng et al. (2025), which aims to identify neuron activations linked to SC and suppress them, and Mulchandani & Kim (2025), which uses data pruning to clean the dataset itself. DFR instead retrains the last layer after standard training (Kirichenko et al., 2023). These approaches are important references, but generally operate under different information budgets. Our paper studies a same-information question: given the same biased data and the same clean counterexample pool, does label-conflicting counterexample mixing provide a stronger shortcut-suppression signal than clean data alone?

Causal and intervention-inspired robustness. Causal reasoning is a growing field in computer vision (Zhang et al., 2024). Castro et al. (2020) focuses on the causality of medical images, and Ilse et al. (2021) explores the use of data augmentation as a simulation for causal interventions. Schrouff et al. (2024) points out the importance of causal graph structure when balancing data for robustness. While Schrouff et al. (2024) focuses on data balancing and not augmentation, it provides argument for the importance of causal methods in dealing with SC via data modification. We use causal language as a design lens for constructing shortcut-label conflicts, we do not perform causal discovery.

3 Problem Setting

Problem layout. We study a weak-supervision setting with biased training data, a small clean counterexample pool, and no shortcut labels in the training objective. We first describe the idealised hidden-watermark setting, where the shortcut cue can be written explicitly as N, then use this notation as intuition for less structured natural shortcuts.

$\{(X_i, y_i)\}_{i=1}^K$ is a clean dataset of K distinct images in $\mathbb{R}^{r\times c}$ and with class labels $y \in [I] \coloneqq \{1, \dots, I\}$. We assume all images of the same class are sampled from $\mathcal{D}_i \coloneqq \mathbb{P}(X|y = i)$, and $I < K$, so there are more images than classes. Next, we introduce a noise dataset $\{(N_i, y_i)\}_{i=1}^K$ of images $N \in \mathbb{R}^{r\times c}$ from I classes. The two datasets share the same K. Each (N_i, y_i) shares the same class label y_i with the corresponding (X_i, y_i). Note that $N_1, \dots, N_K$ can be duplicated.

From the clean and noise datasets, the adversary constructs the augmented dataset $\{N_i, X_i, X_i^{N_i}, y_i\}_{i=1}^K$ by **pixel-wise addition:** given two images X, N, we get $X^N \coloneqq X + N$. A subset of the augmented dataset is the **observed dataset**[1]:

$$\text{observe } \left\{X_i^{N_i}, y_i\right\}_{i=1}^K \subset \{N_i, X_i, X_i^{N_i}, y_i\}_{i=1}^K \qquad \text{(challenge)}$$

$$\text{Learn } f_w \text{ s.t. } f_w(X) = y \neq f_w(N). \qquad \text{(goal)}$$

[1]As N_i is not observed, we cannot apply adversarial training on N_i.

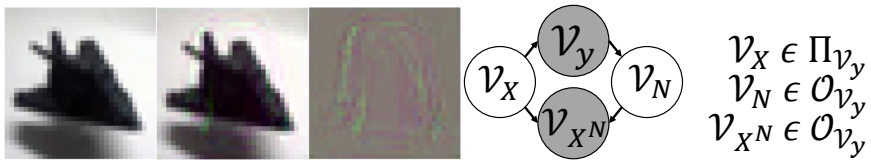


Figure 1: Examples of watermarking results (Jang et al., 2024) and SCM representation in $\{\mathcal{V}, \Pi_{\mathcal{V}}, \mathcal{O}_{\mathcal{V}}\}$ notation (Pearl, 2009). Left-1: clean data $\mathcal{V}_X$, Left-2: observed data $\mathcal{V}_{X^N}$, Left-3: noise $\mathcal{V}_N$. Right: SCM representation for the idealised hidden-watermark setting used as the design lens for Section 4.1.

Counterexamples. We introduce a counterexample dataset $\{(C_j, y_j)\}_{j=1}^{L}$ of clean data points close to X. The images $C_j \in \mathbb{R}^{r \times c}$ have the same number of classes as X_i; for any (X_i, y_i), (C_j, y_j), if $y_j = y_i$, then C_j is sampled from $\mathcal{D}_i$. We also assume $L \ll K$ for the practicality of data collection. $\{(C_j, y_j)\}_{j=1}^{L}$ is used to break SC, as below. Shortcut labels are not used in training; in Waterbirds and CelebA, group labels are used only to define evaluation splits and compute metrics such as worst-group accuracy.

For the hidden-watermark calibration experiment, we note that in practice the observed dataset $\{(X_i^{N_i}, y_i)\}_{i=1}^{K}$ may contain noise-free data points. Our paradigm requires counteracting SC with counterexamples, so it is crucial to know the amount of counterexamples to use. For this, we define the proportion of noisy data points as $SC_{\text{strength}} \coloneqq \#\text{noisy data}/\#\text{data}$. We use this quantity only in the idealised watermark setting to calibrate the amount of counterexample data.

Generalisation step. For less structured settings, N should be read as shorthand for any shortcut cue that is correlated with the label, not necessarily additive noise. In watermarking, N_i is the hidden class-wise signal; in CelebA and Waterbirds, N_i represents a natural attribute such as hair colour, or a background cue. In these more realistic settings, it is often impossible to fully remove or isolate N. We therefore use these more complex datasets to analyse the practical effects of this fact.

4 METHOD

4.1 SHORTCUT-CONFLICTING TRAINING SIGNAL

We use an idealised SCM in Figure 1 as a design lens. In this notation, $\Pi_{\mathcal{V}_y}$ denotes the direct causes of the label node, while $\mathcal{O}_{\mathcal{V}_y}$ denotes other variables that may be predictive without being direct causes. In Figure 1, $\mathcal{V}_X \in \Pi_{\mathcal{V}_y}$ is the image content we want the model to use, while $\mathcal{V}_N \in \mathcal{O}_{\mathcal{V}_y}$ is the shortcut cue we want to suppress. We use the SCM as an intuition: a useful counterexample should combine task evidence and shortcut evidence so that they no longer support the same label, breaking statistical correlation between the shortcut and the label; creating a training example where shortcut-following predictions are penalised. Under this shortcut-conflicting construction:

- If f_w uses the causal path $\mathcal{V}_X \to \mathcal{V}_y$: it correctly predicts y_j, and we get low loss.
- If f_w uses the spurious path $\mathcal{V}_y \to \mathcal{V}_N$: it incorrectly predicts y_i, and we get high loss.

In practice, we make shortcut-following costly by mixing a biased sample with a clean counterexample from another class and supervising the mixture with the counterexample label.

4.2 DEBIASING BY $\{C_j, y_j\}_{j=1}^{L}$

Our method uses a clean dataset $\{C_j, y_j\}_{j=1}^{L}$ to *debias* $\{X_i^{N_i}, y_i\}_{i=1}^{K}$. Here, debiasing refers to reducing the model's reliance on the shortcut cue N. In particular, for each observed data point $(X_i^{N_i}, y_i)$, we construct a counter-example $(X_i^{N_i} \oplus C_t, y_t)$, where $\oplus$ is a mixing of $X_i^{N_i}$ and C_t, y_t, mixing counterfactual noise data with clean data. (C_t, y_t) takes from:

$$(C_t, y_t) \in \{(C_j, y_j) \mid y_j \neq y_i, j = 1, \dots L\}.$$

We repeat this $I - 1$ times for any particular $(X_i^{N_i}, y_i)$, with a *different* y_t for each repeat. We emphasise using multiple C_t with a different y_t to provide contrasting target labels for suppressing



Figure 2: Examples of mixed counterexamples produced by replacing part of a biased image with clean counterexample content from another class. The mixed image contains shortcut evidence associated with one class while being supervised with the clean counterexample label.

SC. Thus, the mixed example contains the shortcut cue associated with y_i but is supervised with target $y_t \neq y_i$, without using shortcut labels.

We train a model f_w using the three following datasets, where $\mathcal{C}$ is the *counterexample dataset*.

$$\mathcal{C} \coloneqq \left\{\left(X_i^{N_i} \oplus C_t, y_t\right)\right\}_{i=1,\ t=1}^{K,\ I-1}, \quad \{(X_i^{N_i}, y_i)\}_{i=1}^{K}, \quad \{(C_i, y_i)\}_{i=1}^{L} \tag{1}$$

As in Section 3, N_i denotes the shortcut cue in the idealised watermark case and serves as shorthand for natural shortcut evidence, such as an attribute or background, in the realistic settings.

4.3 Mixing operation

We give an explicit example of $\oplus$. We define the mixing $X \oplus C$ which masks a portion of X and replaces the masked region with a part of C under a mask B:

$$X \oplus C = B \odot X + (1 - B) \odot C, \ \ B \in \{0,1\}^{r \times c}.$$

Here, $\odot$ is element-wise multiplication and B is a mask matrix. This is the Cutmix operation. More generally: we mix a biased image from class y_i, a clean counterexample from class y_t, and assign the result the label y_t. Another example is the MixUp operation, where $\oplus$ is a linear interpolation of two images. The contribution is not the mixing operator itself: standard CutMix/MixUp preserves the usual label semantics, whereas our construction deliberately creates label conflict, so that shortcut evidence inherited from the biased image points to one class while the clean counterexample label supervises another.

4.4 Training objective and shortcut-conflicting signal

Our training of f_w contains three datasets, instead of one as in standard training. Our training objective combines three terms:

$$\underbrace{\mathbb{E}_{(X_i^{N_i}, y_i)} \mathcal{L}(f_w(X_i^{N_i}), y_i)}_{\text{biased dataset: learns both causal and spurious}} + \underbrace{\lambda_1 \mathbb{E}_{(C_j, y_j)} \mathcal{L}(f_w(C_j), y_j)}_{\text{clean dataset: learns only causal}} + \underbrace{\lambda_2 \mathbb{E}_{(X_i^{N_i} \oplus C_j, y_j)} \mathcal{L}(f_w(X_i^{N_i} \oplus C_j), y_j)}_{\text{counterexamples: penalises spurious}}. \tag{2}$$

This objective suppresses SC via the third term following the logic below: Consider the mixed image $\left(X_i^{N_i} \oplus C_j, y_j\right)$ where $y_j \neq y_i$, it contains:

- Region from $X_i^{N_i}$: contains both X_i (class y_i) and N_i (watermark for class y_i)
- Region from C_j: contains C_j (class y_j)

If f_w relies on the SC $\mathbb{P}(\mathcal{V}_y|\mathcal{V}_N)$, it would predict:

$$f_w\left(X_i^{N_i} \oplus C_j\right) \approx y_i \quad \text{(following the watermark)}. \tag{3}$$

With the given label y_j (see (1)), it creates a training signal that *penalises* reliance on $\mathbb{P}(\mathcal{V}_y|\mathcal{V}_N)$. When N is a natural background or attribute cue, the same logic applies, albeit more approximately: the mixed example penalises predictions utilising SC by providing contrasting target labels. This approximation is strongest when the clean counterexamples preserve task-relevant evidence while breaking the shortcut-label link.

5 EXPERIMENTS

We perform main experiments on two settings: a watermarking dataset following Jang et al. (2024) and augmented Waterbirds (Sagawa et al., 2020). CelebA is included as a stress test for visually entangled shortcut attributes. For the main experiments, we use same-information baselines: biased-only empirical risk minimisation (ERM), biased+clean data, and counterexample CutMix/MixUp, and comparisons to standard MixUp, LISA, and NCM using similar information budgets. The same-information methods except biased-only ERM receive the same clean counterexample pool; LISA and standard MixUp train on the vanilla Waterbirds set, and NCM uses counterfactual pairs taken from the original waterbirds dataset. Shortcut information may be used outside the learning loop to collect or curate this pool, but shortcut labels are not used as training supervision, loss terms, group weights, adversarial targets, or per-sample annotations during model optimisation. Group labels are used only to construct evaluation views and compute worst-group accuracy. We then perform ablation studies to understand the contribution of the clean counterexample set. For watermarks, we report the biased accuracy, clean accuracy, and watermark-only accuracy. We report additional accuracy metrics such as worst-group accuracy (WGA), minority, and clean/context-OOD for augmented Waterbirds, with CelebA details deferred to the appendix.

5.1 HIDDEN WATERMARK SHORTCUT

Application and goals. We first analyse the idealised hidden-watermark setting, where the shortcut cue is cleanly represented by N. Undercover Bias (Jang et al., 2024) is a dataset watermarking method that aims at detecting misuse of image classification datasets, as mentioned in Section 2. This method adversarially introduces watermark N correlated to label y, so we know that the observed dataset $\{X_i^{N_i}, y_i\}_{i=1}^K$ contains SC, but the nature of this SC is difficult to extract. Our goal here is to:

1. Calibrate the watermark shortcut strength in this idealised setting.
2. Train a model f_w that ignores the bias, following the description in Section 4.2.

This is our clean mechanism case: the shortcut is compact, hidden, class-conditional, and separable.

Datasets and construction. Recall datasets $\{(X_i, y_i)\}_{i=1}^K$ and $\{(N_i, y_i)\}_{i=1}^K$, both of which are determined by the adversary as mentioned in Section 3. We follow Jang et al. (2024) and use CIFAR-10 (Krizhevsky & Hinton, 2009) as $\{(X_i, y_i)\}_{i=1}^K$ and FashionMNIST (Xiao et al., 2017) to generate $\{(N_i, y_i)\}_{i=1}^K$. They both consist of 10 classes, so we have $\{(N_i, y_i)\}_{i=1}^K$ in addition to $\{(X_i, y_i)\}_{i=1}^K$. Our adversary uses $\{(N_i, y_i)\}_{i=1}^K$ as the dataset for the *auxiliary task*, to determine if we have used the dataset $\{(X_i^{N_i}, y_i)\}_{i=1}^K$. The auxiliary task is done by measuring the accuracy of our model on $\{(N_i, y_i)\}_{i=1}^K$. If accuracy is high, it is determined that we have used the dataset, following Jang et al. (2024).

Using $\{(X_i, y_i)\}_{i=1}^K$ and $\{(N_i, y_i)\}_{i=1}^K$ we construct $\{(X_i^{N_i}, y_i)\}_{i=1}^K$ using the methods described, see appendix for details. To construct $\{(C_j, y_j)\}_{j=1}^L$, we remove and use 10% of data points from $\{(X_i, y_i)\}_{i=1}^K$. We test our methods with a ResNet (He et al., 2016) following the original work. We provide details of further training parameters in the appendix.

Our method vs the naive method. Fig.3A,B show our method (Ours) compared with the naive method (Naive). In our method, we observe boosted performance in clean data (blue violin plot goes up), a drop in performance in watermarked data (orange violin plot goes down), closing the gap between the two. Most importantly, prediction on watermarks dropped to random chance (green violin plot goes to 0.1), whereas the naive method is reliably predictive on watermarks (green violin plot is high). This provides evidence for our method suppressing SC. Full watermark ablations, including counterexample-size sensitivity, partial watermark strength, structural checks, and shifted-watermark tests, are deferred to Section A.1 of the appendix, where we discuss columns D,E,F in Fig. 3.

Our method vs purely including clean data in training. We now explain the importance of contribution 1 in Section 1. In Fig.3A,C, we replace the portion of $\mathcal{C}$ that underwent our mixing process by a portion of $\{C_j, y_j\}_{j=1}^L$ (that is, f_w now has access to only clean and biased data). We see an improvement in model performance when testing on $\{X_i, y_i\}_{i=1}^K$ (blue goes up), but no lowering of accuracy on $\{N_i, y_i\}_{i=1}^K$ (green does not drop). This fits our theoretical expectation,

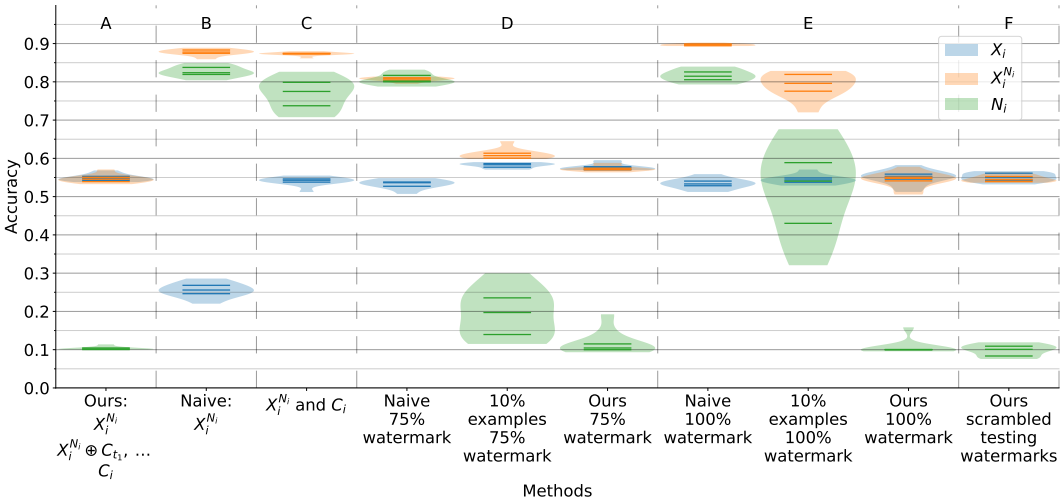


Figure 3: Violin plots showing the performance of f_w when trained and tested with different methods in the hidden-watermark setting. Blue is f_w's performance on $\{X_i, y_i\}_{i=1}^K$, orange on $\{X_i^{N_i}, y_i\}_{i=1}^K$, and green on $\{N_i, y_i\}_{i=1}^K$. The green watermark-only accuracy is the key shortcut proxy: random chance performance indicates that watermark features no longer provide a useful prediction path.

since providing clean data provides a stronger link between $\mathcal{V}_X$ and $\mathcal{V}_y$ (Figure 1), but because it does not prevent information flow from $\mathcal{V}_N$, it fails to remove unwanted influence.

This comparison highlights the problem in simply increasing the amount of data. While this leads to increased performance in a clean environment, it does not guarantee ignoring SC, which are difficult to detect and can imply risks in critical tasks.

Link to natural shortcut experiments. Note that the task was chosen for its specific characteristics: a global, prominent, and perfectly predictive watermark feature. Other tasks might present features that do not match our assumptions, so other mixture methods will be required for different task media. For example, by using linear interpolation instead of patch-based mixing for local features. The following experiments test this boundary: augmented Waterbirds replaces the hidden watermark with a background cue and provides evidence for the importance of high-quality clean counterexamples. Then, CelebA tests follow for an even more entangled shortcut.

5.2 Augmented Waterbirds Background Shortcut

Application and goals. We next analyse augmented Waterbirds (Sagawa et al., 2020), where the shortcut is a natural background cue rather than a hidden watermark. The target is bird type, and the SC is background type: waterbirds are often shown on water and landbirds are often shown on land. This setting tests our counterexample method for a more complex form of shortcut than the hidden watermark above. It also lets us test a more specific hypothesis: better clean counterexamples should provide a stronger debiasing signal. Waterbirds models are also ResNet-18 trained from scratch for 15 epochs with fixed seeds.

Datasets and construction. We use the biased Waterbirds dataset as the main biased training data. The clean counterexample pool is built from strict minority-source examples, so the auxiliary examples do not duplicate the biased training source images. Here, minority refers to when waterbirds appear on land and landbirds appear on water. We evaluate accuracies on biased examples, minority examples, clean foreground-preserving examples, and context-OOD examples. Worst-group accuracy is computed using bird label and background type during evaluation.

Counterexample mixing improves worst-group robustness. Table 1 reports the gain over biased-only ERM for counterexample MixUp. WGA improves from 15.28% for biased-only ERM to 34.88% for MixUp. Biased+clean data gives a smaller mean increase, from 15.28% to 17.41%. Thus, as in

Method	Extra data	Domain labels in train?	Avg. acc. (%)	WGA (%)
Biased-only ERM	none	no	47.85 ± 1.64	15.28 ± 3.98
Biased+clean	240 off-axis clean	no	48.47 ± 4.57	17.41 ± 5.52
Standard MixUp	vanilla set (incl. 240 minority)	no	59.18 ± 4.90	30.78 ± 5.81
LISA	vanilla set (incl. 240 minority)	yes (pairing)	59.46 ± 8.20	27.79 ± 5.52
NCM (matched ResNet-18)	240 on-axis CF pairs	no	40.71 ± 6.50	9.33 ± 6.01
Counterexample MixUp	240 off-axis clean donors	no	53.57 ± 0.81	$\mathbf{34.88 \pm 6.30}$

Table 1: Augmented Waterbirds, OOD-background as clean counterexample pool. NCM is a matched ResNet-18 port of Bai et al. (2025).

Clean stream	Δavg clean	Δavg CutMix	Δavg MixUp	ΔWGA clean	ΔWGA CutMix	ΔWGA MixUp
Black/blank	-5.38	-4.56	$+4.51$	-1.00	$+1.78$	$+3.04$
Blurred background	$+10.36$	$+10.16$	$+5.33$	$+9.22$	$+9.93$	$+10.46$
OOD background	$+0.62$	$+3.65$	$+5.72$	$+2.13$	$+13.18$	$\mathbf{+19.60}$

Table 2: Augmented Waterbirds improvements over biased-only ERM baseline, in percentage points.

the watermark setting, label-conflicting counterexample mixing gives a stronger robustness signal than simply adding the same clean data.

Comparing clean counterexample variants of increasing quality. Black backgrounds remove background information but create visually artificial examples. Blurred backgrounds keep more natural image statistics while weakening the background-label link. OOD backgrounds go further by keeping the bird foreground and placing it into realistic backgrounds outside the water/land shortcut axis. We use Places365-1M as the OOD background dataset (Zhou et al., 2018). Our hypothesis is that the better the clean counterexamples preserve bird evidence while breaking the background correlation, the stronger our method should be.

Table 2 compares the three clean streams using fixed epoch-15 weights. For black or blank backgrounds, the WGA gains are small: 1.78 points for CutMix and 3.04 points for MixUp. Blurred backgrounds give larger gains of 9.93 and 10.46 points, respectively. OOD background replacement gives the strongest WGA gains, 13.18 points for CutMix and 19.60 points for MixUp. Average-accuracy gains are less monotonic across constructions, but these WGA results suggest that realistic, foreground-preserving counterexamples provide a stronger shortcut-suppression signal.

Minority performance shows SC suppression. Minority examples are the Waterbirds analogue of the watermark-only diagnostic: they test whether the model can work when SC is reversed from training. For OOD background replacement, Table 3 shows that minority average accuracy rises from 17.28% for biased-only ERM and 16.93% for biased+clean to 33.52% for counterexample CutMix and 29.87% for counterexample MixUp. Minority WGA likewise rises from 8.11% for biased-only ERM and 11.89% for biased+clean to 19.63% for CutMix and 16.51% for MixUp. This suggests that the model is less dependent on background SC.

Comparison to MixUp, LISA, and NCM. Table 1 reports test average accuracy and WGA. Counterexample MixUp achieves the highest mean WGA among the displayed mixing and constraint methods (34.88%), exceeding standard MixUp by 4.10 points and LISA by 7.09 points, without using domain labels in its training objective. This robustness gain trades off against average accuracy, which is 5.61 points below standard MixUp and 5.89 points below LISA. However, across the three seeds, Counterexample MixUp also has the smallest observed average-accuracy standard deviation (0.81 points versus 4.90, 8.20, and 6.50 for standard MixUp, LISA, and matched NCM, respectively).

5.3 CelebA Attribute Shortcut

Finally, we keep CelebA (Liu et al., 2015) as an entangled-attribute stress test. Here, the target and shortcut are visual face attributes rather than separable image layers, so input-space counterexample mixing is expected to be noisier than in hidden watermarks or replaceable-background Waterbirds. We use gender as the target attribute and blond hair as the shortcut attribute, and construct the clean counterexample pool from examples where the shortcut attribute is off. Fig. 4 shows the resulting baseline trio: average, clean, and biased accuracies remain high, while WGA is more volatile. This

Method	Pooled avg. acc. (%)	Pooled WGA (%)	Minority avg. acc. (%)	Minority WGA (%)
Biased-only ERM	47.85 ± 1.64	15.28 ± 3.98	17.28 ± 3.35	8.11 ± 3.26
Biased+clean	48.47 ± 4.57	17.41 ± 5.52	16.93 ± 2.16	11.89 ± 0.75
Counterexample CutMix	51.51 ± 4.54	28.46 ± 2.08	$\mathbf{33.52 \pm 0.84}$	$\mathbf{19.63 \pm 4.67}$
Counterexample MixUp	$\mathbf{53.57 \pm 0.81}$	$\mathbf{34.88 \pm 6.30}$	29.87 ± 3.45	16.51 ± 6.85

Table 3: Augmented Waterbirds results for the OOD-background clean stream. The pooled headline split excludes the context-OOD view.

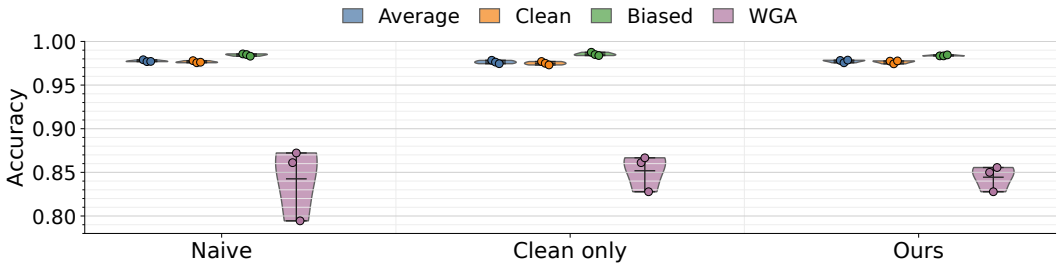


Figure 4: CelebA performance over three seeds. Note WGA volatility: Our method suppresses WGA variation and increases performance, without sacrificing average, clean, and biased accuracy.

supports the regime interpretation that clean counterexamples help most when they preserve task-relevant evidence while breaking the shortcut-label link; more discussion on CelebA is provided in Section A.3 of the appendix.

6 Limitations and Future Work

Our method is scoped to image classification and input-space counterexample mixing; extending it to other modalities (e.g., text) or to regression likely requires different counterexample constructions (e.g., feature- or representation-space mixing). Our evaluation spans a controlled hidden-watermark mechanism case and more natural visual shortcuts, but augmented Waterbirds is a synthetic dataset, like the original Waterbirds dataset, and CelebA shows more modest gains when shortcut and target evidence are visually entangled. Performance also depends on counterexample quality: weak or artificial counterexamples can provide a much weaker shortcut-conflicting signal than realistic, foreground-preserving ones. Finally, we study a different information regime than group-supervised debiasing: for Waterbirds and CelebA, group/shortcut labels define evaluation splits and worst-group metrics, but are not used as training supervision in the counterexample objective. Curating the extra pool may require shortcut knowledge, even if we do not require explicit shortcut or groups labels. The NCM row is a matched-backbone ResNet-18 port, not Bai's frozen-CLIP linear probe. We hope the method's success on the watermark setting inspires future work on finding better dataset protection regimes.

7 Conclusion

We study whether a small clean counterexample pool can attenuate shortcut reliance without using shortcut labels in the training objective. Label-conflicting counterexample mixtures suppress watermark-only prediction to chance in the controlled hidden-watermark setting. On augmented Waterbirds, counterexample mixing improves both average accuracy and WGA over simply adding the same clean pool, and occupies a robustness-oriented trade-off relative to standard MixUp and LISA. CelebA shows weaker and more variable effects when target and shortcut evidence are visually entangled. Together, these results position counterexample mixing as a useful weak-supervision strategy when group labels are unavailable, while showing that its effectiveness depends on the structure of the shortcut and the available counterexamples.

AI USE STATEMENT

Generative AI tools were used for polishing structures and sentences, code editing, and literature lookup during preparation of this manuscript. All scientific claims, experimental numbers, and citations were reviewed by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

We treat hidden-watermark experiments as a controlled mechanism case, not a recipe for removing watermarks from third-party data.

REPRODUCIBILITY STATEMENT

Dataset constructions, mixing rules, and training hyperparameters are described in Section 5 and Section A.5. Matched LISA / NCM / MixUp comparisons use the same backbone, epoch budget, and seed set; protocol details are given in the appendix.

REFERENCES

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In Laura Leal-Taixé and Stefan Roth (eds.), *Computer Vision – ECCV 2018 Workshops*, pp. 556–572, Cham, 2019. Springer International Publishing.

Ruqi Bai, Yao Ji, Zeyu Zhou, and David I. Inouye. From invariant representations to invariant data: Provable robustness to spurious correlations via noisy counterfactual matching. *arXiv preprint arXiv:2505.24843*, 2025. URL https://arxiv.org/abs/2505.24843.

Ananth Balashankar, Xuezhi Wang, Yao Qin, Ben Packer, Nithum Thain, Ed Chi, Jilin Chen, and Alex Beutel. Improving classifier robustness through active generative counterfactual data augmentation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 127–139, Singapore, dec 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.10. URL https://aclanthology.org/2023.findings-emnlp.10/.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, Jul 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17478-w. URL https://doi.org/10.1038/s41467-020-17478-w.

Elliot Creager, Joern-Henrik Jacobsen, Richard Zemel, and Polina Kirichenko. Environment inference for invariant learning. In *International Conference on Machine Learning*, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Maximilian Ilse, Jakub M. Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4555–4562. PMLR, 2021.

Jinhyeok Jang, ByungOk Han, Jaehong Kim, and Chan-Hyun Youn. Rethinking data bias: Dataset copyright protection via embedding class-wise hidden bias. In *European Conference on Computer Vision (ECCV)*, 2024.

Ben Kent. Foundry and machine learning, 2021. Conference on Visual Media Production.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2306.08514*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Philippe, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Evan Z. Liu, Babak Haghgoo, Annie 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*, 2021.

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)*, December 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Varun Mulchandani and Jung-Eun Kim. Severing spurious correlations with data pruning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Bk13Qfu8Ru.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: De-Biasing Classifier from Biased Classifier. In *Neural Information Processing Systems*, 2020. URL https://mlanthology.org/neurips/2020/nam2020neurips-learning/.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28448–28467. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/qiu23c.html.

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. In *International Conference on Learning Representations*, 2020.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021.

Jessica Schrouff, Alexis Bellot, Amal Rannen-Triki, Alan Malek, Isabela Albuquerque, Arthur Gretton, Alexander D'Amour, and Silvia Chiappa. Mind the graph when balancing data for fairness or robustness. In *Advances in Neural Information Processing Systems*, volume 37, pp. 29913–29947, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25407–25437. PMLR, 2022.

Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

Kexuan Zhang, Qiyu Sun, Chaoqiang Zhao, and Yang Tang. Causal reasoning in typical computer vision tasks. *Science China Technological Sciences*, 67(1):105–120, jan 2024.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=8yKEo06dKNo.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Re. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26484–26516. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhang22z.html.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. NeuronTune: Towards self-guided spurious bias mitigation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78454–78477. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zheng25s.html.

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, 2018. doi: 10.1109/TPAMI.2017.2723009.

A Appendix

A.1 Full Hidden Watermark Experiments

This appendix section gives the hidden-watermark details not kept in the main text. See Figure 3 for the corresponding violin plots.

Difference of accuracies as measure of debiasing completeness. Fig. 3D,E show the results of an experiment supporting contribution 2 in Section 1. The columns show accuracies of f_w trained on: the naive method, our method but with $\mathcal{C}$ at a tenth of its required size, and our full method. We verify our method on two datasets, one with 75% of the images watermarked and the other fully watermarked. We see in both examples that accuracy gaps begin large (blue and orange violins are far apart), but the gap closes until they match when $\mathcal{C}$ meets its required size. Note that for the stronger watermark correlation Fig. 3E, the gap closes much more slowly, indicating that the weaker the SC, the more effective our method.

Structural difference between $\{X_i^{N_i}, y_i\}_{i=1}^K$ and $\mathcal{C}$. The watermarks N exist in $\mathcal{C}$ in a different form than within $\{X_i^{N_i}, y_i\}_{i=1}^K$. In the following, we discuss a key concern of this form of mixing:

1. There is a structural difference between the watermarks in $\mathcal{C}$ and $\{X_i^{N_i}, y_i\}_{i=1}^K$ - watermarks in $\mathcal{C}$ do not cover the whole image, while in $\{X_i^{N_i}, y_i\}_{i=1}^K$ they do.
2. Watermarks are predictive on themselves.

Because of 1 and 2, we have the concern that the network might learn to identify and exploit the existence of watermarks only when it detects the absence of our mixing process. To address this concern, we look at the performance between clean and watermarked test data. If f_w learns to recognise and exploit the existence of the mixing process, it should learn to predict well on all of $\{N_i, y_i\}_{i=1}^K$, $\{X_i^{N_i}, y_i\}_{i=1}^K$ and $\{C_j, y_j\}_{j=1}^L$. However, the comparable clean and watermarked performance together with chance-level watermark-only accuracy counters this concern. This supports our claim that f_w is not able to circumvent our method by noticing when mixing occurs.

Our method vs purely including clean data in training. In Fig. 3A,C, we replace the portion of $\mathcal{C}$ that underwent our mixing process by a portion of $\{C_j, y_j\}_{j=1}^L$ (that is, f_w now has access to only clean and biased data). We see an improvement in model performance when testing on $\{X_i, y_i\}_{i=1}^K$ (blue goes up), but no lowering of accuracy on $\{N_i, y_i\}_{i=1}^K$ (green does not drop). This supports the main-text point that clean data alone can improve clean accuracy without suppressing the watermark shortcut.

Different watermark correlation in test scenario. Fig. 3A,F provide more evidence that our model does not perform prediction based on watermark feature. We generate two datasets which are identical, except that in the second dataset testing watermarks are all moved 1 class forwards (i.e. class 1 is now labelled class 2, etc.), wrapping around in a modulo fashion for the 10th class. That is, the data points for the test watermarks take the form $(N_i, y_i + 1)$. As Fig. 3 shows, our method performs equivalently on both datasets (all three violins from both experiments are approximately the same). Note that watermark data of N_i, marked green are both at 0.1 accuracy. This suggests that watermark features do not impact our model's prediction process.

A.2 Augmented Waterbirds Construction

Augmented Waterbirds starts from the Waterbirds benchmark (Sagawa et al., 2020), where the target label is bird type and the shortcut is the background type. We use the usual binary coding: landbirds have target label 0, waterbirds have target label 1, land backgrounds have background label 0, and water backgrounds have background label 1. Thus, the shortcut-aligned groups are landbird–land and waterbird–water, while the shortcut-conflicting minority groups are landbird–water and waterbird–land. The biased view keeps the shortcut-aligned source images; the minority view keeps the shortcut-conflicting source images.

For each source image, the augmented dataset also stores foreground-preserving variants based on a bird mask. The mask is used as the bird foreground, and only the background is changed. For example, an original waterbird-on-water image in the biased view remains a waterbird example, but its clean variants can replace the water background with a blurred version of the same scene, a black background, or an OOD background such as a forest or indoor scene from Places365-1M (Zhou et al., 2018). Similarly, an original landbird-on-water minority image remains a landbird example, but its foreground can be preserved while the water background is weakened or replaced. In all cases, the bird label is unchanged; the construction only changes how much background evidence is available.

The clean counterexample pool used for training is not simply the whole clean split. It is built from strict minority-source examples: source images whose original bird/background combination conflicts with the shortcut, with no source-ID overlap with the biased training images. This matters because the auxiliary clean examples should provide shortcut-conflicting evidence rather than duplicate shortcut-aligned training images. Concretely, if a biased training image is a waterbird on water, a label-conflicting mixed counterexample may pair that biased image with a clean landbird example derived from a landbird-on-water source image and supervise the mixture as landbird. The resulting mixed image contains background evidence that can still point toward the biased sample's label, but the training target points toward the clean counterexample label.

We evaluate four views. The biased view tests performance on shortcut-aligned examples such as waterbirds on water and landbirds on land. The minority view tests the direct shortcut-conflict cases, such as waterbirds on land and landbirds on water. The clean view tests foreground-preserving

Method	Context-OOD avg. acc. (%)	Context-OOD WGA (%)
Biased-only ERM	54.84 ± 4.10	31.69 ± 3.89
Biased+clean	57.65 ± 1.91	$\mathbf{42.07 \pm 2.11}$
Counterexample CutMix	$\mathbf{62.66 \pm 2.68}$	34.22 ± 2.80
Counterexample MixUp	61.22 ± 4.00	34.84 ± 10.10

Table 4: Context-OOD results for the OOD-background clean stream, using fixed epoch-15 weights. Values are mean $\pm$ population std over seeds $\{0, 1, 2\}$. This view uses the same background-generation mechanism as the clean counterexample stream and is therefore excluded from the pooled headline metric.

examples in which the background shortcut is weakened but the bird remains visible. The context-OOD view tests foreground-preserving examples placed on backgrounds outside the water/land axis. The main OOD-background table excludes context-OOD from the pooled headline row because that view is generated with the same OOD-background mechanism as the clean counterexample stream. We report it separately in Table 4.

On this same-mechanism view, counterexample mixing gives the highest mean average accuracy, whereas biased+clean gives the highest WGA. This differs from the minority view, where both mixing variants improve the mean WGA over biased+clean, and reinforces that the effect depends on the evaluation shift.

A.3 CelebA Attribute Stress Test

Application and goals. Finally, we test CelebA (Liu et al., 2015) as a setting where shortcuts are heavily entangled with the target. In this experiment, the target and shortcut are visual face attributes rather than separable image layers. This makes the task more difficult than both hidden watermarks and augmented Waterbirds. The purpose is to test whether the clean-counterexample signal carries beyond compact hidden watermarks and replaceable backgrounds.

Datasets and construction. We use CelebA attribute labels to define the target, shortcut, and evaluation groups. We construct the clean counterexample pool by choosing those examples where the shortcut attribute is off. For our experiments, we use gender as the target attribute and blond hair as the shortcut attribute. This is reversed from the common CelebA setup, where blond hair is the target and gender is the shortcut. We do this to facilitate ablation later on, where we perform perturbations to hair colour.

Naive training has volatile worst-group accuracy. Fig. 4 shows that the naive method can keep high average accuracy while WGA is volatile. This is the key CelebA insight: average accuracy alone suggests the task is solved, but worst-group performance exposes shortcut reliance and seed sensitivity.

Clean-only as a best-case control. Clean-only is used as a best-case control. As the spurious hair signal is absent, the model shows less WGA variation than biased data.

Counterexample mixing approximates the clean-only gain. Our method gives a good approximation to the clean-only robustness benefit. It does not clearly dominate clean-only, but approaches the same behaviour while using the same counterexample principle as in the watermark and Waterbirds experiments. Our insight is thus for complex tasks such as CelebA, the method remains useful, but the gain is weaker and more variable because the shortcut is visually entangled with the target.

Entanglement explains the weaker effect. This weaker effect is consistent with the structure of the dataset. In hidden watermarks, the shortcut is a compact added signal. In augmented Waterbirds, the background can be explicitly altered while preserving the bird. In CelebA, the target and shortcut live in the same face crop, so CutMix/MixUp can create noisy visual-label conflicts. CelebA therefore shows the limitation of the method: counterexample mixing remains competitive without hurting average accuracy, but entangled attributes make shortcut attenuation harder and more volatile.

Clean stream	Method	Avg. acc. (%)	WGA (%)
Black/blank	Biased-only ERM	53.35 ± 5.88	22.80 ± 4.70
Black/blank	Biased+clean	47.97 ± 2.26	21.81 ± 4.00
Black/blank	Counterexample CutMix	48.79 ± 4.32	24.58 ± 6.41
Black/blank	Counterexample MixUp	57.87 ± 5.36	25.84 ± 6.99
Blurred background	Biased-only ERM	51.97 ± 2.98	24.75 ± 0.82
Blurred background	Biased+clean	62.33 ± 2.33	33.97 ± 5.82
Blurred background	Counterexample CutMix	62.13 ± 3.21	34.68 ± 6.05
Blurred background	Counterexample MixUp	57.30 ± 7.53	35.21 ± 8.23
OOD background	Biased-only ERM	47.85 ± 1.64	15.28 ± 3.98
OOD background	Biased+clean	48.47 ± 4.57	17.41 ± 5.52
OOD background	Counterexample CutMix	51.51 ± 4.54	28.46 ± 2.08
OOD background	Counterexample MixUp	53.57 ± 0.81	34.88 ± 6.30

Table 5: Augmented Waterbirds clean-stream comparison using fixed epoch-15 weights. Values are mean $\pm$ population std over seeds $\{0, 1, 2\}$. Black/blank and blurred-background results pool all four test views; the OOD-background headline excludes context-OOD.

A.4 Analysis and Ablations

A.4.1 Full Clean-Stream Results

Table 5 gives the absolute results underlying the main-text clean-quality deltas.

A.4.2 Clean Data Alone Is Not Enough

Across the experiments, clean data is useful but does not by itself provide the same shortcut-conflicting signal as mixed counterexamples. In the hidden-watermark setting, adding clean examples improves clean-image accuracy, but the watermark-only accuracy remains high; this means the model can still use N as a predictive path. The mixed counterexample objective changes this signal because the watermark cue and the supervised label disagree. This is why the green watermark-only violin in Figure 3 is the central diagnostic: chance-level performance indicates that the watermark alone no longer supports the auxiliary prediction task.

The same pattern appears more partially in augmented Waterbirds. For the OOD-background clean stream, biased+clean produces only a small mean WGA increase over biased-only ERM, while both counterexample mixing variants produce larger pooled and minority-view gains (main-text Table 3). Thus, clean examples reduce shortcut reliance when they are good counterexamples, but label-conflicting mixing can make the shortcut conflict more explicit.

A.4.3 Counterexample CutMix vs Counterexample MixUp

CutMix and MixUp instantiate the same supervision principle with different image-space assumptions. CutMix preserves local image regions and therefore works well when task evidence is spatially localised and can be pasted without destroying the label signal. In the OOD-background results, CutMix gives the highest minority average accuracy and minority WGA, while MixUp gives the highest pooled average accuracy and pooled WGA. MixUp is smoother and sometimes less disruptive, but can also dilute the foreground evidence and background conflict.

The comparison should therefore be read as regime-dependent rather than as a universal ranking. Under the epoch-15 protocol, counterexample MixUp has higher pooled WGA than CutMix (34.88% versus 28.46%), while CutMix is stronger on the minority view. For black or weakly realistic counterexamples, both methods can be weaker because the clean stream no longer preserves the same natural image evidence. The relevant conclusion is that the mixing operation must preserve enough task-relevant content while making shortcut-following predictions costly.

A.4.4 COUNTEREXAMPLE SET SIZE

The watermark ablations in Figure 3D,E vary the amount of counterexample signal by comparing a tenth-size $\mathcal{C}$ to the full counterexample construction. When $\mathcal{C}$ is too small, the clean and watermarked accuracies remain separated, indicating incomplete suppression of the watermark shortcut. As the counterexample set reaches the calibrated size, this gap closes and the watermark-only diagnostic approaches chance. The fully watermarked condition requires more counterexample signal than the 75% watermarked condition, matching the intuition that stronger shortcut correlations need more shortcut-conflicting evidence.

We interpret this size result only in the controlled watermark setting. There, SC_{strength} is explicitly defined as the proportion of watermarked examples and can be used to calibrate the amount of counterexample data. For natural shortcuts such as backgrounds or face attributes, there is no equivalent observable scalar in our method. The Waterbirds and CelebA experiments instead show the practical analogue: the quality and separability of the counterexamples matter at least as much as their count.

A.4.5 SHORTCUT SEPARABILITY

The experiments form a separability ladder. Hidden watermarks are compact, class-conditional, and additive, so a spatial mixture can create a clean conflict between shortcut evidence and target label. Augmented Waterbirds is harder because the shortcut is a broad visual context, but it is still replaceable: a bird foreground can be kept while the background-label link is broken. CelebA is harder again because target and shortcut attributes live in the same face crop, so input-space mixing can create noisy visual-label conflicts rather than clean counterexamples.

This ladder explains the empirical pattern. Watermark suppression is strongest because the shortcut cue is separable and the diagnostic is direct. Under fixed epoch-15 evaluation, the WGA gain from mixing increases from black/blank to blurred and OOD-background counterexamples, with OOD-background MixUp giving the largest gain over its corresponding biased-only baseline. CelebA shows weaker and more variable gains because blond hair and gender evidence are visually entangled; high average accuracy can coexist with volatile worst-group accuracy. The method should therefore be viewed as a clean-counterexample weak-supervision tool whose effectiveness depends on whether counterexamples preserve task evidence while breaking the shortcut-label link.

A.5 TRAINING DETAILS

For the hidden-watermark experiments, we follow the CIFAR-10 (Krizhevsky & Hinton, 2009)/FashionMNIST (Xiao et al., 2017) construction of Jang et al. (2024). The classifier is a small convolutional network with three convolutional blocks and dropout, followed by a 10-way linear classifier. We train with cross-entropy loss and Adam (Kingma & Ba, 2015) using its default learning rate 10^{-3} and batch size 128. The epoch budget is fixed by training condition and held constant across repeats; more demanding conditions are trained longer to reach convergence rather than applying one oversized budget to every condition. The clean counterexample pool is formed by reserving 500 clean examples per class from the ordered training set, leaving the remaining watermarked examples as biased training data. Mixed counterexamples are generated by pairing clean examples from one class with watermarked examples from other classes; the cached mixed set uses random-quarter CutMix boxes, and the mixed label is the clean counterexample label. The implementation trains one cross-entropy objective over the concatenated biased, clean, and mixed datasets, so their relative contributions are determined by the corresponding dataset sizes.

For augmented Waterbirds, all pilot runs use ResNet-18 (He et al., 2016) with a two-class output head, image size 224, ImageNet normalisation, and random horizontal flips during training. Unless otherwise stated, the model is trained from random initialisation with AdamW (Loshchilov & Hutter, 2019), learning rate 10^{-4}, weight decay 10^{-4}, batch size 64, and seeds $\{0, 1, 2\}$. The main Waterbirds results use the fixed epoch-15 weights; worst-group accuracy is test-only and is not used for checkpoint selection. We compare biased-only ERM, biased plus clean counterexamples, counterexample CutMix, and counterexample MixUp. For counterexample mixing, we use hard labels, $\alpha = 0.75$ for the mixing distribution, include the ordinary clean loss, and use a mixed-example multiplier of 1.0.

The Waterbirds clean-counterexample variants modify only the auxiliary clean stream. Black or blank backgrounds remove the context but create visually artificial images. Blurred-background examples weaken the background cue using colour jitter and Gaussian blur with radius 12. OOD-background examples paste the bird foreground onto Places365-1M backgrounds selected to avoid the water/land shortcut axis; the candidate scenes include forest, snow, mountain, desert, urban, and indoor categories, and exclude water, ocean, lake, river, wetland, grass, grassland, pasture, field, coast, and beach. OOD composites use a feather radius of 4 and colour matching.

Standard MixUp and LISA train on the 4,795-example union of the biased and minority training splits. Both use $\lambda \sim \mathrm{Beta}(2, 2)$ and soft interpolated labels. Standard MixUp pairs examples by a random minibatch permutation without using domain labels, whereas LISA uses background/domain labels to select different-class examples from the same domain.

The matched ResNet-18 NCM port trains its classification loss on the 4,555 biased examples and uses 240 same-class, opposite-background pairs only to estimate a spurious feature subspace. At the start of each epoch, it computes a rank-8 truncated SVD of pair differences in the 512-dimensional penultimate feature space. The classifier weight and its AdamW first-moment state are projected orthogonally to this subspace after each update. The feature subspace is refreshed each epoch because the ResNet-18 encoder is trained end-to-end. Background metadata is used offline to construct the pairs, but is not supplied as a loss target or per-example input during classifier training. This is a matched-backbone adaptation of hard NCM rather than the frozen-CLIP linear-probe architecture used in the original NCM experiments.

For CelebA, we use ResNet-18 (He et al., 2016) with image size 224, AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4} and weight decay 10^{-4}, batch size 64, and 5 epochs. The reported stress test uses gender as the target attribute and blond hair as the shortcut attribute. The clean counterexample pool uses 5% of the training data and the no-spurious clean-pool construction, where the shortcut attribute is absent from the auxiliary clean examples. The compared methods are ERM, clean-only, and counterexample CutMix with $\alpha = 0.75$, hard labels, clean loss enabled, and checkpoint selection by validation average accuracy.