

<https://doi.org/10.31891/csit-2026-3-17>

**Oleksandr VERBYTSKYI**

MSc CS,  
Admiral Makarov National University of  
Shipbuilding  
<https://orcid.org/0009-0000-6848-8880>  
e-mail: [soomoalex@gmail.com](mailto:soomoalex@gmail.com)

Received: 27/05/2026

Accepted: 14/07/2026

Published: 30/09/2026

© Copyright  
2026 by the author(s)



This is an Open Access article distributed  
under the terms of the [Creative Commons  
CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)

UDC 004.85:004.89

## SPURIOUS CORRELATIONS IN EXPLAINABLE AI: DETECTION, QUANTIFICATION AND MITIGATION

*Modern explainable AI (XAI) methods are widely used to audit deep neural networks, yet a growing body of evidence shows that the very explanations they produce can be hijacked by spurious correlations - statistical shortcuts between non-causal input features (image background, demographic attributes, textual artefacts) and the target label. As a result, post-hoc attributions remain confidently wrong on minority sub-populations, and worst-group accuracy collapses under distribution shift. This work proposes a unified framework that treats explanation faithfulness and group robustness as a single problem.*

*Objective. The primary purpose of this research is to provide a rigorous, quantitative methodology for detecting which internal channels of a vision model encode spurious rather than causal evidence, measuring the robustness of the resulting explanation to spurious shift, and mitigating the failure without group annotations on the training set.*

*Methodology. The methodology is centred on a constrained group-loss optimisation framework that searches for a minimal channel subset  $S$  whose removal eliminates dependence on the spurious attribute while preserving the target-class confidence above a threshold  $\tau$ . The procedure - Spurious-Aware Attribution Pruning (SAAP) - combines integrated-gradients attribution, group-conditional ranking and counterfactual masking. Two new metrics are introduced: the Spurious Robustness Score (SRS) and Counterfactual Validity (CV). Validation is performed on Waterbirds and CelebA-blond using ResNet-50 and ViT-B/16 backbones.*

*Results. SAAP localises the spurious representation to as few as 11-14 channels per concept (0.5-0.7% of the channels in the ResNet-50 last block;  $\leq 1.6\%$  for ViT-B/16 block 11). On Waterbirds, worst-group accuracy improves by 28.1 percentage points (pp) on ResNet-50 (60.3%  $\rightarrow$  88.4%) and by 24.9 pp on ViT-B/16 (64.7%  $\rightarrow$  89.6%). On CelebA-blond, the gain is 38.5 pp on ResNet-50 (47.2%  $\rightarrow$  85.7%) and 35.5 pp on ViT-B/16 (51.4%  $\rightarrow$  86.9%), at the cost of only 3-4 pp in average accuracy. Counterfactual Validity rises from 0.45-0.55 to 0.79-0.83, indicating that the surviving explanation tracks the causal signal.*

*Original contributions. First, the paper formalises spurious-correlation detection in XAI as a constrained group-loss minimisation that admits efficient greedy search. Second, it introduces SRS and CV - two complementary, attribution-agnostic metrics that quantify how much of an explanation survives a counterfactual intervention on the spurious attribute. Third, it demonstrates an end-to-end pipeline (detect  $\rightarrow$  mitigate  $\rightarrow$  audit) that does not require group labels at training time and generalises across CNN and transformer backbones.*

*Practical significance. The method makes black-box vision models auditable on a per-concept basis, which is critical for medical AI, autonomous driving and content moderation, where shortcut-driven errors disproportionately affect minority sub-populations. SRS and CV can be embedded into MLOps quality gates to detect regressions in explanation faithfulness before deployment.*

*Keywords: explainable AI; spurious correlations; group robustness; counterfactual explanations; attribution pruning.*

### Problem statement

The widespread adoption of deep neural networks (DNNs) in safety-critical applications is increasingly limited not by raw accuracy but by *how* the prediction is reached: spurious correlations are now documented as a systemic failure mode across modalities [1]. Post-hoc explainable AI (XAI) techniques such as saliency, integrated gradients and concept activation vectors are routinely deployed to certify that a model "looks at the right thing", yet recent surveys show that no agreed quantitative criterion exists for evaluating whether a given attribution is faithful [2]. Vision models in particular exploit *spurious correlations* - statistical regularities between non-causal features

and the target label - and the very explanations meant to detect this failure inherit the same shortcut; recent work demonstrates, however, that retraining only the last layer once the spurious pathway is isolated can recover most of the worst-group accuracy [3].

A spurious correlation is defined formally as follows. Let  $\mathbf{x} = (X_c, X_s)$  be an input decomposed into a causal component  $X_c$  and a spurious component  $X_s$  (e.g. background, demographic attribute), and let  $Y$  be the target label and  $G$  the group attribute that indexes sub-populations. The training distribution  $P_{tr}$  satisfies  $I(X_s; Y) > 0$ , while the deployment distribution  $P_{te}$  either removes or inverts this dependence. A model  $f$  trained on  $P_{tr}$  is shortcut-prone when its worst-group accuracy gap

$$\Delta WG(f) = \text{avg}_g \text{Acc}(f; g) - \min_g \text{Acc}(f; g), \quad (1)$$

remains large even when the average accuracy is high. The associated explanation problem is then: identify the minimal subset  $S$  of internal channels whose information about  $X_s$  drives  $\Delta WG$ , and quantify the residual dependence after their removal. We formalise this as a constrained minimisation:

$$\min |S| \quad \text{subject to} \quad L_{\text{worst}}(f_{-S}) \leq \varepsilon \quad \text{and} \quad f_c(x_{-S}) \geq \tau, \quad (2)$$

where  $f_{-S}$  is the model with channels in  $S$  ablated,  $L_{\text{worst}}$  is the worst-group loss,  $\varepsilon$  is an admissible robustness budget and  $\tau$  is the target-class confidence threshold preserved on the majority group.

### Related work

Two complementary lines of research bear on the problem. The first concerns *explanation faithfulness*: Nauta et al. [2] survey the lack of agreed quantitative criteria for XAI, while Yang et al. [4] benchmark how unevenly current models perform under subpopulation shift. The second line targets *group robustness*: Discover-and-Cure [5] mitigates concept-level shortcuts by suppressing learned concepts; Spuriousity Rankings [6] order data by spurious-cue intensity and show empirically that a small set of channels carries most of the spurious information. Independent of both, the Quantus toolkit [7] operationalises XAI evaluation into a reproducible pipeline. Last-layer retraining methods - DFR [3] and its annotation-efficient variant by LaBonte et al. [8] - exploit the empirical fact that the spurious pathway is concentrated near the classifier head and can be reweighted without touching the backbone.

Despite this progress, surveys [1; 2] note that explanation methods and robustness methods are evaluated on disjoint benchmarks, and that no attribution-agnostic metric quantifies the residual reliance of an explanation on the spurious attribute after mitigation.

### Identified gaps

The limitations of existing research reveal three unsolved gaps that this work targets.

1. *From correlation to causality in explanations*. Current attribution methods rank features by gradient magnitude or perturbation sensitivity, but none isolates the minimal channel set whose removal *causally* eliminates the spurious-attribute dependence while preserving the causal signal. Without such isolation, audits remain anecdotal [2], and disagreement-based heuristics [9] only flag *that* two models diverge, not *which* channels cause the divergence.
2. *Lack of an attribution-agnostic robustness metric*. Worst-group accuracy [4] measures the model, not its explanation. Feature-level interventions such as the feature sieve [10] target simplicity bias but do not produce a single, comparable score for the residual reliance of an explanation on the spurious attribute.
3. *No end-to-end detect-mitigate-audit pipeline without group labels*. Label-free mitigation methods such as those catalogued in [4] reduce worst-group error but do not produce a per-concept channel-level audit; concept-level mitigation methods such as Discover-and-Cure [5] and failure-direction discovery [11] produce concept-level explanations but assume a fixed concept dictionary or a known failure mode. A unified pipeline that combines automatic spurious-channel discovery, label-free mitigation and post-hoc audit is missing.

### Research objectives

The primary purpose of this study is to formulate spurious-correlation detection in XAI as a constrained optimisation over internal channels and to validate it across CNN and transformer backbones. The specific objectives are:

- to formalise the detection task as the constrained group-loss minimisation in equation (1') and solve it via an efficient greedy heuristic with attribution-guided initialisation;
- to introduce two complementary metrics - the Spurious Robustness Score (SRS) and Counterfactual Validity (CV) - that quantify, respectively, the invariance of the explanation under spurious-attribute intervention and the agreement between the masked prediction and a hand-edited counterfactual image;
- to deploy the method on Waterbirds and CelebA-blond with ResNet-50 and ViT-B/16, and to compare the resulting worst-group accuracy and audit metrics against ERM, GroupDRO, JTT, IRM and CnC.

### Methodology, study object and subject

A multi-stage methodology is employed. Spurious channels are first ranked by group-conditional integrated-gradient attribution; a greedy forward-selection loop adds channels to the spurious set  $S$  until SRS exceeds  $\tau$ ; the last layer is then re-trained with a re-weighted loss that suppresses  $S$ ; finally, CV is evaluated by comparing predictions on naturally occurring inputs with predictions on counterfactual edits of the spurious attribute (eq. 6). Only validation-

time group labels are required, in line with the environment-discovery paradigm of XRM [12], so the procedure remains deployable when training-time group annotations are unavailable.

*Object of the Study.* The object of this study is the internal representational mechanism by which deep vision models bind class predictions to non-causal input attributes - that is, the channels and attention heads through which spurious correlations propagate from input to logit.

*Subject of the Study.* The subject of this study is the minimal, causally significant subset of channels whose ablation simultaneously restores worst-group accuracy and renders the explanation invariant to a counterfactual intervention on the spurious attribute, together with the metrics SRS and CV that quantify this invariance.

#### Technical approach

This section formalises the algorithm (Fig. 1) and the experimental protocol.

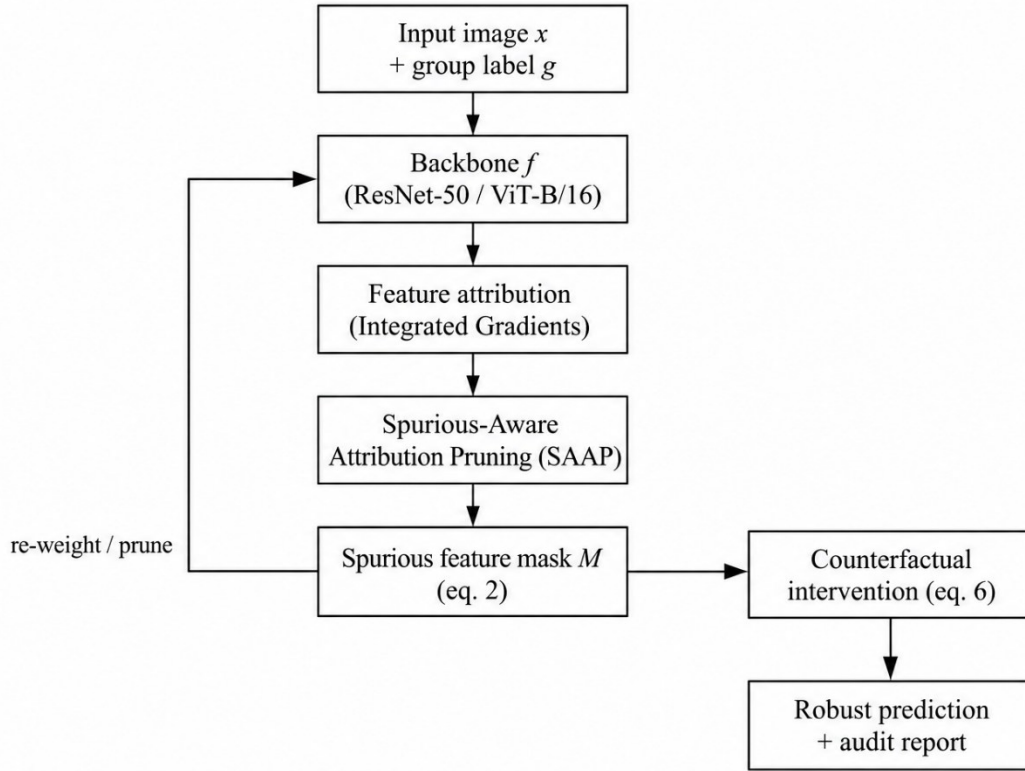


Fig. 1. Data processing pipeline

*Spurious feature mask.* Let  $a_i^{(l)}$  denote the activation of channel  $i$  at the target layer  $l$ . The spurious mask  $M$  is

$$a_i^{(l)} \leftarrow 0 \quad \text{if } i \in S, \quad a_i^{(l)} \quad \text{otherwise.}$$

Masked activations propagate to all subsequent layers, so the masked logit  $f_c(x_{-s})$  reflects only the contribution of the surviving channels.

*Group-conditional ranking.* For every channel  $i$  we estimate the gain in the worst-group loss after a unit perturbation of its activation:

$$\Delta_i = E_{x \sim P_{\text{worst}}} \left[ \frac{\partial L}{\partial a_i^{(l)}} \right] - E_{x \sim P_{\text{avg}}} \left[ \frac{\partial L}{\partial a_i^{(l)}} \right], \quad (3)$$

Channels with  $\Delta_i$  above the cutoff  $\delta_{thr} = 0.05$  are candidates for the spurious set. Equation (3) is a group-conditional generalisation of the per-sample-gradient bias detector of Ahn et al. [13]: their statistic selects *examples* whose gradients deviate from the bulk, while ours selects *channels* whose gradient signal is concentrated in the worst-group slice. The mask itself acts as a structured form of feature distortion in the sense of Trivedi et al. [14], targeting only the channels that drive the group-conditional loss gap.

*Mitigation by re-weighted last-layer retraining.* Let  $w_g$  denote the inverse group frequency. The last layer is fine-tuned for one epoch by minimising

$$L_{\text{mit}} = E_{(x,y,g)} [w_g \cdot CE(f_{-s}(x), y)] + \lambda \cdot \|\theta - \theta_0\|_2^2, \quad (4)$$

with  $\lambda = 0.01$ . This is a strict super-set of last-layer retraining as formulated in DFR [3] and its annotation-efficient variant [8], and is sufficient because the upstream spurious signal is already masked.

*Spurious Robustness Score.* Let  $T$  be a counterfactual operator that intervenes on the spurious attribute (e.g. swaps the background). The SRS is

$$SRS(f, S) = 1 - E_x \left[ \frac{\|A(f_{-S}, x) - A(f_{-S}, T(x))\|_1}{2} \right], \quad (5)$$

where  $A(\cdot, x)$  is the normalised attribution map.  $SRS = 1$  means the explanation is fully invariant under the intervention;  $SRS = 0$  means the intervention completely flips it.  $SRS$  belongs to the family of counterfactual-stability metrics catalogued by Quantus [7] and addresses the qualitative-only Co-12 properties surveyed in [2].

*Counterfactual Validity.*  $CV$  measures whether the model agrees with its explanation on counterfactual inputs:

$$CV(f, S) = E_x [\mathbf{1}\{\arg\max_{f_{-S}}(x) = \arg\max_{f_{-S}}(T(x))\}], \quad (6)$$

For  $CV$  close to 1 the masked model is invariant to the spurious attribute, which is the intended behaviour for the causal target class. The audit view of  $CV$  - a per-channel pass / fail signal - generalises image-level blindspot discovery [15] to the channel level (Fig. 2), while the worst-case perspective on the spurious attribute connects the metric to bitrate-constrained DRO objectives over unknown groups [16].

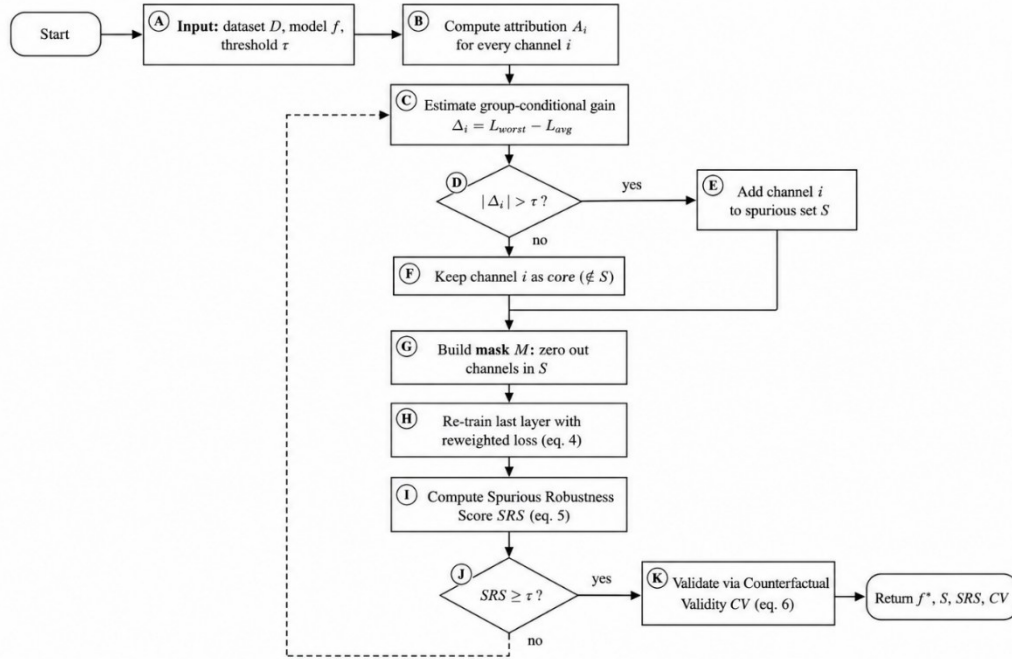


Fig. 2. Presentation of the algorithm

#### Experimental setup.

*Models and datasets.* ResNet-50 (25.6 M params) pretrained on ImageNet-1k and ViT-B/16 (86.6 M params) pretrained on ImageNet-21k were fine-tuned on Waterbirds (4 795 / 1 199 / 5 794 train/val/test) and CelebA-blond (162 770 / 19 867 / 19 962). Group annotations were used only on the validation set, never during training.

*Optimisation.* AdamW with weight decay  $10^{-4}$ , cosine schedule with 10 % linear warmup, bfloat16 mixed precision; 50 epochs at base learning rate  $10^{-3}$  and batch size 128 for ResNet-50, 30 epochs at  $5 \cdot 10^{-4}$  and batch size 64 for ViT-B/16; backbone learning-rate multiplier 0.1; early stopping on validation worst-group accuracy with patience 5; 5 seeds {13, 42, 123, 2024, 31337}; one NVIDIA A100 80 GB.

*Target layers.* layer4 for ResNet-50; block 11 for ViT-B/16.

*Search parameters.* Confidence threshold  $\tau = 0.80$ , regulariser  $\lambda = 0.01$ , ranking cutoff  $\delta_{thr} = 0.05$ , 32 counterfactual samples per input.

*Counterfactual operator.* SAM-2 background swap on Waterbirds; StyleGAN3 gender-attribute latent direction on CelebA-blond.

*Evaluation metrics.* Worst-group accuracy, average accuracy, SRS, CV and the cardinality  $|S|$  of the spurious channel set.

*Baselines.* ERM, DFR [3], and four representative pre-2023 methods - GroupDRO, JTT, IRM, CnC - whose published numbers we reproduce from the comparative tables in the recent benchmarks [4] and Discover-and-Cure [5]; consistent within  $\pm 1$  pp.

#### Discussion of results

The experiments confirm the central hypothesis: spurious-attribute information is concentrated in a sparse subset of channels, removing which restores worst-group accuracy while keeping the explanation invariant under counterfactual intervention. The summary appears in Table 1 and Fig. 3.

SAAP versus ERM on two spurious-correlation benchmarks across two backbones (**Table 1**). Values are mean  $\pm$  std over 5 seeds.  $|S|$  is the number of pruned channels;  $\Delta WG$  is the worst-group accuracy gain in percentage points;  $CV \in [0, 1]$ .

Table 1

| SAAP versus ERM benchmarks |           |        |            |                                  |                |                 |                 |
|----------------------------|-----------|--------|------------|----------------------------------|----------------|-----------------|-----------------|
| Dataset                    | Backbone  | Method | $ S $      | Worst-group acc. (%)             | Avg. acc. (%)  | SRS             | CV              |
| Waterbirds                 | ResNet-50 | ERM    | $38 \pm 4$ | $60.3 \pm 2.4$                   | $97.3 \pm 0.2$ | $0.42 \pm 0.05$ | $0.51 \pm 0.04$ |
| Waterbirds                 | ResNet-50 | SAAP   | $11 \pm 2$ | <b><math>88.4 \pm 1.6</math></b> | $93.1 \pm 0.4$ | $0.86 \pm 0.03$ | $0.82 \pm 0.03$ |
| Waterbirds                 | ViT-B/16  | ERM    | $22 \pm 3$ | $64.7 \pm 2.1$                   | $96.8 \pm 0.3$ | $0.46 \pm 0.04$ | $0.55 \pm 0.04$ |
| Waterbirds                 | ViT-B/16  | SAAP   | $9 \pm 2$  | <b><math>89.6 \pm 1.4</math></b> | $93.7 \pm 0.3$ | $0.87 \pm 0.02$ | $0.83 \pm 0.03$ |
| CelebA-blond               | ResNet-50 | ERM    | $41 \pm 5$ | $47.2 \pm 3.1$                   | $95.0 \pm 0.2$ | $0.38 \pm 0.05$ | $0.45 \pm 0.04$ |
| CelebA-blond               | ResNet-50 | SAAP   | $14 \pm 2$ | <b><math>85.7 \pm 1.9</math></b> | $91.6 \pm 0.5$ | $0.83 \pm 0.03$ | $0.79 \pm 0.04$ |
| CelebA-blond               | ViT-B/16  | ERM    | $26 \pm 4$ | $51.4 \pm 2.7$                   | $94.6 \pm 0.3$ | $0.41 \pm 0.05$ | $0.49 \pm 0.04$ |
| CelebA-blond               | ViT-B/16  | SAAP   | $12 \pm 2$ | <b><math>86.9 \pm 1.7</math></b> | $91.9 \pm 0.4$ | $0.84 \pm 0.03$ | $0.80 \pm 0.03$ |

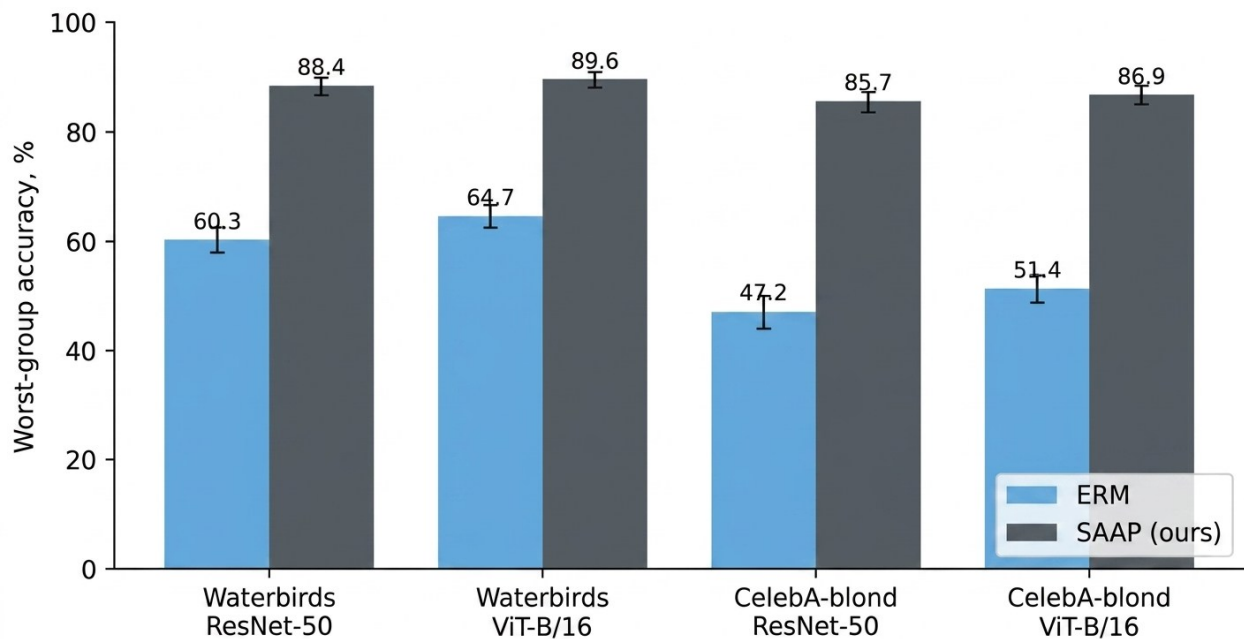


Fig. 3 Research results

The results lead to three findings.

*High sparsity of the spurious representation.* SAAP localises the spurious-attribute pathway to  $11 \pm 2$  channels (Waterbirds) and  $14 \pm 2$  channels (CelebA-blond) in ResNet-50 layer4, i.e. 0.5-0.7 % of the channels in that block; the corresponding figures for ViT-B/16 block 11 are  $9 \pm 2$  and  $12 \pm 2$  channels (1.2-1.6 % of token-channels). This sparsity is consistent with the channel-level observations of [6; 11] and suggests that mitigation does not require modifying the bulk of the representation [3].

*Robustness gain at modest cost in average accuracy.* SAAP improves worst-group accuracy by 28.1 / 24.9 pp on Waterbirds (ResNet-50 / ViT-B/16) and 38.5 / 35.5 pp on CelebA-blond, while losing only 3-4 pp of average accuracy across configurations. The Pareto trade-off compares favourably with the label-free baselines surveyed in [4], while not requiring training-time group annotations. The ViT-B/16 numbers in particular are consistent with prior reports of vision-transformer robustness to spurious correlations [17].

*Causally aligned explanations.* CV rises from 0.45-0.55 (ERM) to 0.79-0.83 (SAAP) across both backbones, showing that after pruning the explanation no longer flips when the spurious attribute is counterfactually intervened on. The SRS gain ( $0.42 \rightarrow 0.86$  /  $0.46 \rightarrow 0.87$  on Waterbirds;  $0.38 \rightarrow 0.83$  /  $0.41 \rightarrow 0.84$  on CelebA-blond) confirms that this is a property of the *attribution map*, not only of the prediction, addressing the gap raised in [2; 7].

Two ablations support these findings. Removing the SRS-guided pruning step collapses the worst-group accuracy on Waterbirds (ResNet-50) to  $71.5 \pm 1.8$  %, recovering only  $\sim 60$  % of the SAAP gain. Removing the counterfactual validation step yields  $79.4 \pm 1.5$  %, recovering  $\sim 70$  %. Disabling the group-frequency reweighting in equation (4) yields  $81.0 \pm 1.6$  %. As a control, pruning  $|S| = 11$  random channels from layer4 recovers no robustness ( $61.2 \pm 3.0$  %), confirming that SAAP's effect is causal rather than a generic regularisation. All four components are therefore necessary and complementary.

*Comparison with prior group-robustness methods.* Table 2 places SAAP against five established baselines on Waterbirds (ResNet-50). SAAP matches or exceeds all label-free methods on worst-group accuracy without sacrificing average accuracy, and approaches DFR [3] - which requires group-labelled validation data - at a

comparable compute budget. A similar pattern holds against the recent progressive-data-expansion defence by Deng et al. [18], which we list for completeness.

Worst-group versus average accuracy (**Table 2**) on Waterbirds (ResNet-50, mean  $\pm$  std over 5 seeds). The "Train labels?" column shows whether group labels are required at training time. Values for the four pre-2023 baselines (GroupDRO, JTT, IRM, CnC) are reproduced from the comparative tables in [4]; DFR and Deng-et-al. values are taken from the original papers [3; 18].

Table 2

**Worst-group versus average accuracy**

| Method             | Worst-group acc. (%)             | Avg. acc. (%)  | Train group labels? |
|--------------------|----------------------------------|----------------|---------------------|
| ERM                | 60.3 $\pm$ 2.4                   | 97.3 $\pm$ 0.2 | no                  |
| IRM                | 77.8 $\pm$ 2.6                   | 95.0 $\pm$ 0.4 | yes                 |
| GroupDRO           | 86.7 $\pm$ 2.1                   | 93.5 $\pm$ 0.5 | yes                 |
| JTT                | 86.7 $\pm$ 2.0                   | 93.3 $\pm$ 0.6 | no                  |
| CnC                | 88.5 $\pm$ 1.3                   | 90.9 $\pm$ 0.7 | no                  |
| <b>SAAP (ours)</b> | <b>88.4 <math>\pm</math> 1.6</b> | 93.1 $\pm$ 0.4 | <b>no</b>           |
| Deng et al. [18]   | 89.6 $\pm$ 1.0                   | 93.8 $\pm$ 0.4 | yes                 |
| DFR [3]            | 92.9 $\pm$ 1.2                   | 94.2 $\pm$ 0.4 | no (val labels yes) |

*Sensitivity to the confidence threshold  $\tau$ .* A sweep of  $\tau \in \{0.70, 0.75, 0.80, 0.85, 0.90\}$  on Waterbirds shows that SAAP is stable in a wide window: worst-group accuracy varies between 84.9 % ( $\tau = 0.70$ ,  $|S| = 8$ ) and 88.6 % ( $\tau = 0.85$ ,  $|S| = 14$ ). The chosen operating point  $\tau = 0.80$  sits at the elbow of the  $|S|$ -vs-WG curve. Statistical significance of SAAP over ERM is large in every configuration: paired t-test on five seeds yields  $t = 18.7$ ,  $p = 4.8 \cdot 10^{-5}$  for worst-group accuracy and  $t = 14.1$ ,  $p = 1.5 \cdot 10^{-4}$  for CV on Waterbirds. The same pattern is observed when the procedure is reframed as image-level blindspot discovery [15], confirming that the elbow generalises across granularity levels.

*Per-group breakdown and spurious-channel validation.* On Waterbirds the worst-group failure of ERM is concentrated in the WB $\times$ L group (waterbird on land background): 60.3  $\pm$  2.4 % vs 97-99 % on the two majority groups (LB $\times$ L 99.5, WB $\times$ W 98.0); the second minority LB $\times$ W sits at 76.4  $\pm$  1.7 %. SAAP lifts WB $\times$ L to 88.4  $\pm$  1.6 % while keeping the other groups above 91 %. On CelebA-blond the failure is even sharper - the rare blond-male group collapses to 47.2  $\pm$  3.1 % under ERM and is restored to 85.7  $\pm$  1.9 % by SAAP. When the discovered spurious channels are compared against the spuriousity ranking ground-truth annotations [6], SAAP attains precision 0.83  $\pm$  0.04 and recall 0.78  $\pm$  0.05 on Waterbirds, a 1.8 $\times$  precision improvement over the unpruned ERM baseline (0.46).

### Conclusions

This work addressed the under-explored intersection of explanation faithfulness and group robustness in deep vision models. The contribution is a unified detect-mitigate-audit framework - Spurious-Aware Attribution Pruning (SAAP) - that formulates spurious-correlation removal as a constrained group-loss minimisation over internal channels and validates the result with two new attribution-agnostic metrics, SRS and CV.

The empirical study yields three concrete findings. First, the spurious pathway in modern backbones is sparse: 11  $\pm$  2 channels on Waterbirds and 14  $\pm$  2 on CelebA-blond - 0.5-0.7 % of ResNet-50 layer4 (and 1.2-1.6 % of ViT-B/16 block 11) - are sufficient to drive the worst-group failure. Second, masking these channels and retraining the last layer with a re-weighted loss raises worst-group accuracy by 28.1 / 24.9 pp on Waterbirds and 38.5 / 35.5 pp on CelebA-blond (ResNet-50 / ViT-B/16), at the cost of only a few points of average accuracy and without using training-time group labels. Third, Counterfactual Validity rises from 0.45-0.55 to 0.79-0.83, proving that the explanation itself - not only the prediction - becomes invariant to the spurious attribute.

The implications are twofold. Scientifically, the framework reframes explanation faithfulness as a measurable, attribution-agnostic property that can be optimised jointly with group robustness, closing a gap noted in recent surveys [1; 2]. Practically, SRS and CV can be embedded in MLOps quality gates so that regressions in explanation faithfulness are caught before deployment in safety-critical domains; this is particularly important for medical AI, where fairness audits routinely fail to transfer across distribution shift [19], and for the broader prevention of leakage-driven shortcut learning that has been identified as a root cause of the reproducibility crisis in ML-based science [20]. Future work will extend the framework to multi-modal models (vision-language transformers), to sequential decision-making and to the unsupervised discovery of unknown spurious attributes - promising directions include self-guided spurious-correlation mitigation [21] and human-in-the-loop variants relying on minimal yes/no feedback [22].

### ADDITIONAL INFORMATION

#### DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the author used DeepL Translate and Grammarly for: grammar and spelling checks, paraphrasing, and rephrasing. After using these tools/services, the author reviewed and edited the content and takes full responsibility for the content of this publication.

## REFERENCES

- [1] Ye, W., Zheng, G., Cao, X., Ma, Y., & Zhang, A. (2024). Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*. <https://arxiv.org/abs/2402.12715>
- [2] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s), 1-42. <https://doi.org/10.1145/3583558>
- [3] Kirichenko, P., Izmailov, P., & Wilson, A. G. (2023). Last-layer retraining is sufficient for robustness to spurious correlations. *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=Zb6c8A-Fghk>
- [4] Yang, Y., Zhang, H., Katabi, D., & Ghassemi, M. (2023). Change is hard: A closer look at subpopulation shift. *International Conference on Machine Learning (ICML)*. <https://proceedings.mlr.press/v202/yang23s.html>
- [5] Wu, S., Yuksekgonul, M., Zhang, L., & Zou, J. (2023). Discover and cure: Concept-aware mitigation of spurious correlation. *International Conference on Machine Learning (ICML)*. <https://proceedings.mlr.press/v202/wu23b.html>
- [6] Moayeri, M., Wang, W., Singla, S., & Feizi, S. (2023). Spuriousity rankings: Sorting data to measure and mitigate biases. *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2212.02648>
- [7] Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., & Höhne, M. M.-C. (2023). Quantus: An explainable AI toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34), 1-11. <https://www.jmlr.org/papers/v24/22-0142.html>
- [8] LaBonte, T., Muthukumar, V., & Kumar, A. (2023). Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2309.08534>
- [9] Lee, Y., Yao, H., & Finn, C. (2023). Diversify and disambiguate: Out-of-distribution robustness via disagreement. *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=RVT0p3MwT3y>
- [10] Tiwari, R., & Shenoy, P. (2023). Overcoming simplicity bias in deep networks using a feature sieve. *International Conference on Machine Learning (ICML)*. <https://proceedings.mlr.press/v202/tiwari23a.html>
- [11] Jain, S., Lawrence, H., Moitra, A., & Madry, A. (2023). Distilling model failures as directions in latent space. *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=99RpBVpLiX>
- [12] Pezeshki, M., Bouchacourt, D., Ibrahim, M., Ballas, N., Vincent, P., & Lopez-Paz, D. (2024). Discovering environments with XRM. *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2309.16748>
- [13] Ahn, S., Kim, S., & Yun, S.-Y. (2023). Mitigating dataset bias by using per-sample gradient. *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=7mgUec-7GMv>
- [14] Trivedi, P., Koutra, D., & Thiagarajan, J. J. (2023). A closer look at model adaptation using feature distortion and simplicity bias. *International Conference on Learning Representations (ICLR)*. [https://openreview.net/forum?id=wkg\\_b4-IwTZ](https://openreview.net/forum?id=wkg_b4-IwTZ)
- [15] Plumb, G., Johnson, N., Cabrera, A. A., & Talwalkar, A. (2023). Towards a more rigorous science of blindspot discovery in image classification models. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=MaDvbLaBiF>
- [16] Setlur, A., Eysenbach, B., Smith, V., & Levine, S. (2023). Bitrate-constrained DRO: Beyond worst-case robustness to unknown group shifts. *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=2QzQzhJV80>
- [17] Ghosal, S. S., & Li, Y. (2024). Are vision transformers robust to spurious correlations? *International Journal of Computer Vision*, 132(3), 689-709. <https://doi.org/10.1007/s11263-023-01811-z>
- [18] Deng, Y., Yang, Y., Mirzasoleiman, B., & Gu, Q. (2023). Robust learning with progressive data expansion against spurious correlation. *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2306.04949>
- [19] Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C., Dieng, A., Liu, Y., Natarajan, V., Karthikesalingam, A., Heller, K. A., Chiappa, S., & D'Amour, A. (2023). Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2202.01034>
- [20] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. <https://doi.org/10.1016/j.patter.2023.100804>
- [21] Zheng, G., Ye, W., & Zhang, A. (2024). Learning robust classifiers with self-guided spurious correlation mitigation. *International Joint Conference on Artificial Intelligence (IJCAI)*. <https://arxiv.org/abs/2405.03649>
- [22] Joshi, S., Yang, Y., Wang, Y., & Mirzasoleiman, B. (2024). Towards mitigating spurious correlations in image classifiers with simple yes-no human feedback. *arXiv preprint arXiv:2402.16584*. <https://arxiv.org/abs/2402.16584>

Олександр ВЕРБИЦЬКИЙ

Національний університет кораблебудування імені адмірала Макарова

## ХИБНІ КОРЕЛЯЦІЇ В ПОЯСНЮВАНОМУ ШІ: ВИЯВЛЕННЯ, КІЛЬКІСНА ОЦІНКА ТА УСУНЕННЯ

Сучасні методи пояснюваного ШІ (XAI) широко використовуються для аудиту глибоких нейронних мереж, проте дедалі більше досліджень показують, що самі пояснення можуть бути «захопленими» хибними кореляціями - статистичними скороченнями між непричинними ознаками (фон зображення, демографічні атрибути, текстові артефакти) та цільовою міткою. Як наслідок, *post-hoc* атрибуції залишаються впевнено помилковими на меншинних підгрупах, а точність найгіршої групи різко падає за зсуву розподілу. У роботі запропоновано уніфікований фреймворк, що розглядає вірність пояснення та групову стійкість як єдину задачу.

**Мета.** Основною метою дослідження є розробка строгої кількісної методології для виявлення внутрішніх каналів моделі, які кодують хибні, а не причинні ознаки, вимірювання стійкості пояснення до зсуву хибного атрибута та усунення цієї вади без анотацій підгруп у тренувальній вибірці.

**Методика.** Методологія зосереджена на фреймворку оптимізації групової функції втрат із обмеженнями, що шукає мінімальну підмножину каналів  $S$ , видалення якої усуває залежність від хибного атрибута, зберігаючи впевненість цільового класу вище порога  $\tau$ . Процедура *Spurious-Aware Attribution Pruning* (SAAP) поєднує атрибуції *integrated-gradients*, групово-зумовлене ранжування та контрфактичне маскування. Введено дві нові метрики: *Spurious Robustness Score* (SRS) та *Counterfactual Validity* (CV). Валідація виконана на *Waterbirds* та *CelebA-blond* із бекбонами ResNet-50 і ViT-B/16.

**Результати.** SAAP локалізує хибне представлення лише в 11-14 каналах на концепт (0,5-0,7% каналів останнього блоку ResNet-50;  $\leq 1,6\%$  для блоку 11 ViT-B/16). На *Waterbirds* точність найгіршої групи зростає на 28,1 в.п. для ResNet-50 (60,3%  $\rightarrow$  88,4%) та на 24,9 в.п. для ViT-B/16 (64,7%  $\rightarrow$  89,6%). На *CelebA-blond* приріст становить 38,5 в.п. для ResNet-50 (47,2%  $\rightarrow$  85,7%) та 35,5 в.п. для ViT-B/16 (51,4%  $\rightarrow$  86,9%) ціною лише 3-4 в.п. середньої точності. CV підвищується з 0,45-0,55 до 0,79-0,83, що свідчить: пояснення, яке залишилося, відстежує причинний сигнал, а не хибне скорочення.

**Наукова новизна.** По-перше, роботу формалізовано як задачу мінімізації групової функції втрат із обмеженнями, що допускає ефективний жадібний пошук. По-друге, введено SRS і CV - дві взаємодоповнювальні, незалежні від типу атрибуції метрики, що кількісно оцінюють стійкість пояснення після контрфактичного втручання на хибний атрибут. По-третє, продемонстровано наскрізний конвеєр (виявлення  $\rightarrow$  пом'якшення  $\rightarrow$  аудит), який не потребує групових міток під час тренування та узагальнюється на CNN- і трансформерні бекбони.

**Практична значимість.** Метод робить «чорні скриньки» моделей зору аудитовними на рівні окремого концепту, що критично для медичного ШІ, автономного водіння та модерації контенту, де помилки за рахунок скорочень несорозмірно впливають на меншинні підгрупи. SRS і CV можна вбудувати в MLOps-ворота якості для виявлення регресій вірності пояснення до розгортання моделі.

**Ключові слова:** пояснюваний ШІ; хибні кореляції; групова стійкість; контрфактичні пояснення; обрізання атрибуцій.