This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Causal Feature Alignment: Learning to Ignore Spurious Background Features

Rahul Venkataramani[1,2] Parag Dutta[1] Vikram Melapudi[2]
Ambedkar Dukkipati [1]
[1]Indian Institute of Science, Bangalore
[2]GE HealthCare, Bangalore
{rahulv, paragdutta, ambedkar}@iisc.ac.in

Abstract

Deep neural networks are susceptible to spurious features strongly correlating with the target. This phenomenon leads to sub-optimal performance during real-world deployment where spurious correlations do not exist, leading to deployment challenges in safety-critical environments like healthcare. While spurious features can correlate with causal features in myriad ways, we propose a solution for a common manifestation in computer vision where the background corresponds to a spurious feature. In contrast to previous works, we do not require apriori knowledge of different groups in the data induced by the presence/absence of spurious features and corresponding access to samples. We propose a method, Causal Feature Alignment (CFA), to ignore the spurious background features by utilizing segmentations on a small subset of training data. To reduce the annotation burden, we reduce the pixel-wise annotation task of segmentation to a review task of selecting the best mask by utilizing the recently released foundation model and a feature attribution method. We demonstrate our method on a wide range of datasets, including the semi-synthetic ColoredMNIST, WaterBirds, and ImageNet Backgrounds Challenge, and obtain significant gains over state-of-the-art methods.

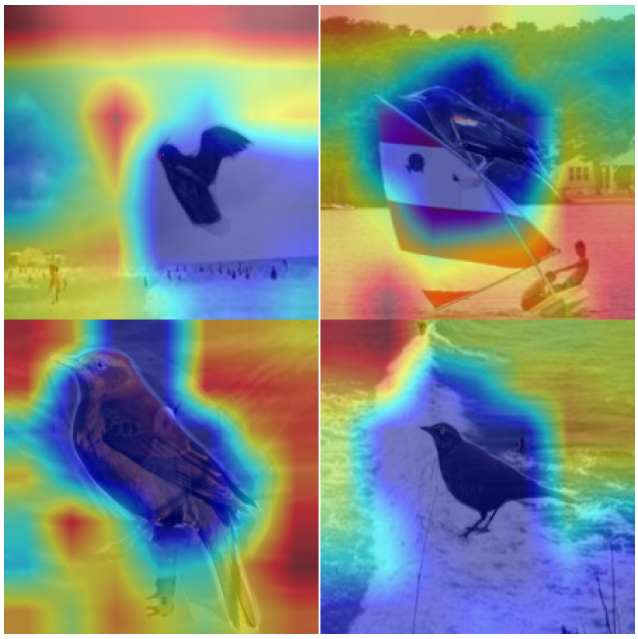

Figure 1. [Best viewed in color] Examples of GradCAM saliency map (with red corresponding to the model's notion of salient features) to demonstrate that a deep learning model trained using ERM on the waterbirds dataset often relies on the background (spurious, but easy to identify correlation) for making predictions.

1. Introduction

Deep neural networks trained based on the classical Empirical Risk Minimization (ERM) have attained expert-level performance across several tasks in recent years [2]. However, as applications of AI fuelled by deep learning continue to increase, recently, several works have pointed to the vulnerability of deep neural networks to rely on *spurious features* [17]. This leads to dramatic failures of AI during deployment in the real world. An illustration of this phenomenon was highlighted by [11] in the example of pneumothorax classification using chest X-ray. The authors discovered that the deep learning model utilized chest tubes, a treatment artefact found in the image's background, to predict whether the images belong to the pneumothorax class rather than the actual causal clinical features like a collapsed lung, ruptured air blisters, etc. Similar examples of vulnerabilities in dermatology [18] and biometrics [4] raise the broader question of the dangers these algorithms could pose on under-represented groups in training.

The seminal work [17] highlights the neural network's reliance on simpler spurious correlation, which attributes the propensity of networks to focus on weakly correlating but simpler to detect features in the presence of tougher but causal features, even though learning the causal features of-

ten leads the classifier to discriminate with better accuracy. This dependence of a classifier on non-causal features manifests as poor accuracy in the following couple of scenarios: **(i)** out-of-domain (OOD) generalization when spurious features vary across domains, and **(ii)** on samples from under-represented groups even within the same domain (quantified by worst-group accuracy - formally defined in 2.1). Typically, both of these scenarios are handled separately since OOD generalization requires access to domain information, whereas improving the worst-group accuracy requires predefined knowledge of groups and group label annotation on a subset of the data (either during training or validation). Additional latent information in the form of either group labels or domain labels is necessary to achieve competitive performance in these settings. Moreover, these methods are highly susceptible to hyper-parameter tuning as demonstrated in [3], where state-of-the-art worst-group performance is achieved using conventional ERM with proper hyper-parameter tuning.

We observe that the prior art for improving worst-group accuracy (compared in Sec. 3) attempts the general problem when spurious features can manifest in multiple ways and are therefore constrained to the choice of additional input obtained. In vision problems, we note that backgrounds frequently correspond to spurious features and hypothesize that knowledge of the background for a subset of examples can help overcome this problem. To illustrate the phenomenon, saliency maps on the popular Waterbirds dataset (Fig. 1) show that ERM-trained models are highly biased towards the background. Thus, an annotation like segmentation or bounding box is a proxy to delineate the causal feature in the image. Specifically, in this paper, we propose a method, Causal Feature Alignment (CFA), that does not utilize any knowledge of group information (i.e., the presence/absence of spurious information) or domain information and instead relies on the segmentations of the foreground on a small subset of the training data. The segmented region of the image spatially identifies the causal feature. We posit that a prediction made using merely this causal feature while ignoring the spurious background feature will improve performance in both the OOD and worst-group scenarios. This enforcement of utilizing only causal features can be considered an 'intervention' step in causal inference terminology [13].

Our method draws inspiration from a recent work [12] that demonstrates representations learned using ERM-trained deep learning models contain both causal and non-causal features, but the subsequent classifier is highly biased towards utilizing the spurious (non-causal) features for prediction. However, when only causal features are present in the representations, the classifier performs well across all groups. We build on this idea to enforce representation ns from original images to match representations of only causal features.

Contributions:

- We first demonstrate our method's ability to ignore spurious features on the Colored-MNIST dataset.
- We subsequently show that CFA can be extended to the much tougher Backgrounds Challenge [8], where we reduce background reliance on the ERM-trained model. This results in a *6*% improvement in accuracy (absolute metrics) over an ERM-trained model on a notorious variant of the Backgrounds challenge.
- We use the method without any modifications on the benchmark Waterbirds dataset, where we improve the worst group accuracy from *74*% to *93*%, setting a new state-of-the-art result in the process.
- We propose an algorithm to generate the causal feature (foreground mask) in an unsupervised fashion, with only review by humans, thereby reducing human effort.
- We also construct a more challenging Waterbirds test dataset to simulate a realistic scenario where not all groups are known during training. Our method sustains the impressive performance gain on this dataset, while methods reliant on group information see dramatic performance degradation.

To the best of our knowledge, this is the first work that identifies a need for utilizing spatial causal localization to improve worst-group accuracy by reducing the reliance of neural networks on spurious features and proposes a method (CFA) for the same.

2. Causal Feature Alignment (CFA)

CFA is a two-stage algorithm, with the first stage a classical ERM-trained algorithm that extracts both core and spurious features. We employ a second training on the ERM model to force the representation extracted from the original image to mimic the representation of a foreground-only image.

2.1. Setup

We consider the classical supervised learning setup for classification with data-target pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$, where $\mathcal{D}$ is the dataset containing N samples. We assume that these samples are derived from different groups $\mathbf{g} \in \mathbf{G}$, where groups are defined by a combination of presence/absence of spurious feature $\mathbf{s} \in \mathbf{S}$ and target label $\mathbf{y}$ ($\mathbf{G} \equiv \mathbf{Y} \times \mathbf{S}$). For instance, in the Waterbirds dataset, four groups correspond to different combinations of bird type (landbird v/s waterbird) and background type (land v/s water). Typically, these groups are not equally represented in the dataset due to a naturally high correlation between attributes. This results in inferior accuracy on under-represented groups in the training dataset quantified through *worst-group accuracy*, defined as the minimum accuracy among all the groups evalu-

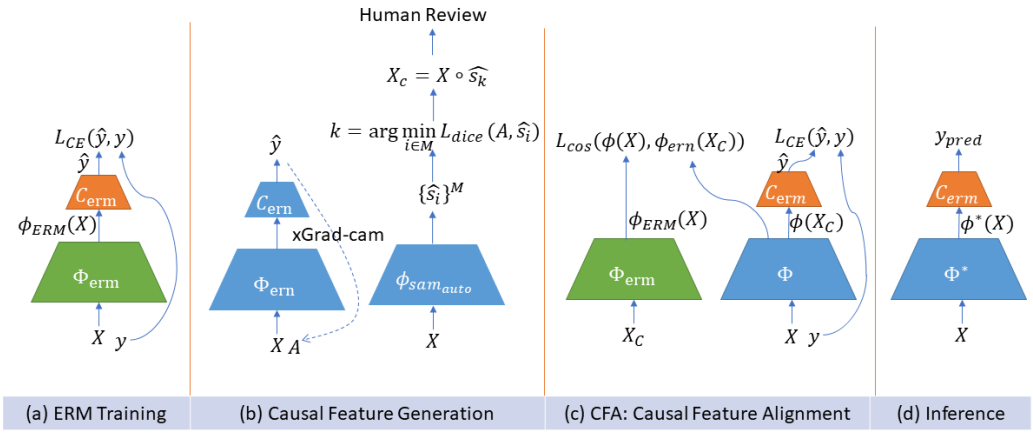

Figure 2. [Best viewed in color] (a) ERM Training: The first step is conventional ERM training with the entire training data. (b) We utilize a foundation model to generate images with only causal features that are reviewed by humans (c) CFA forces the network to output representations invariant to the background and focus only on the task of interest. (d) Inference: The representations learned during CFA are used for making predictions. Note that no additional supervision is used in the inference phase.

ated separately:

$$Acc_{WG} = \min_{g \in \mathbf{G}} Acc(\mathbf{f}((\mathbf{x}), \mathbf{y}) | \mathbf{x} \in \mathbf{g}) \quad (1)$$

Typically, groups with fewer training examples have worse group accuracy.

2.2. Method

Stage 1: ERM Training: A deep learning model can be represented as a composition of a feature extractor ϕ and a classification head $\mathbf{c}$ as $\mathbf{f} = \phi \circ \mathbf{c}$. Using the entire training dataset (of N samples), we train the network $\mathbf{f}$ using the classical ERM technique. We denote the model trained with this setting as $\mathbf{f_{erm}} = \phi_{\mathbf{erm}} \circ \mathbf{c_{erm}}$.

We explored two options for generating segmenting masks: a) through manual annotation (bounding box inputs to Segment Anything (SAM) [1], a foundational model) and b) generate a set of masks and corresponding scores that the human can select to reduce annotation burden.

To generate the set of masks, we utilize SAM in automatic mode without any prompts. To provide a score for each mask, we utilize xGradCAM [14] to a pixel-wise attribution $\mathbf{A}$ for a given prediction. Dice Loss between the xGradCAM output and each SAM-generated mask provides a score that humans can utilize to select the optimal causal mask(s). In the presence of spurious correlations, xGradCAM often saliently attributes the spurious features, and in these cases, the mask with the least score is often the causal mask.

Stage 2: Causal Feature Alignment: We utilize the generated segmentations on the small subset of data $\mathcal{D}'$ to force the feature extractor to provide only causal features, even on images with varied backgrounds. Note that the network is not trained to predict the mask; the mask is used only to derive foreground-only causal images. In this step, only the feature extractor ϕ is trained while the classifier is frozen $\mathbf{c_{erm}}$ (Algorithm 1). We deliberately keep the classifier fixed, as **(a)** causal features are known to perform well with the trained ERM classifier and **(b)** the classifier is trained on a large dataset. Since this causal representation learning stage aims to enforce representations of original images to causal features, we proceed with the above choice of fixed ERM classifier. This stage is trained with a combination of the following objectives:

- Alignment Loss: Cosine Loss between features of original images and foreground-only images $\mathbf{L_{cos}}$
- Classification Loss: Cross Entropy Loss between predictions of the original image and target $\mathbf{L_{CE}}$

While the alignment loss aims to align the representations of the original images with the embeddings of causal features, the classification loss ensures that the discriminative ability of the features from the original images is not compromised.

We modify the feature extractor representations using the following loss function:

$$\mathbf{L}(\phi(\mathbf{x}), \phi_{erm}(\mathbf{x_c}), \mathbf{y}) = \overbrace{\mathbf{L_{cos}}(\phi_{\mathbf{erm}}(\mathbf{x_c}), \phi(\mathbf{x}))}^{\text{Alignment Loss}} + \underbrace{\lambda\, \mathbf{L_{CE}}(\mathbf{c_{erm}}(\phi_{\mathbf{erm}}(\mathbf{x})), \mathbf{y})}_{\text{Classification Loss}} \quad (2)$$

where $\phi^* = \arg\min_\phi \mathbf{L}(\cdot)$. The cosine embedding loss $\mathbf{L_{cos}}$ is defined as $\mathbf{L_{cos}}(\mathbf{a}, \mathbf{b}) = 1 - \mathbf{a} \cdot \mathbf{b} / ||\mathbf{a}|| \cdot ||\mathbf{b}||$ while the cross-entropy loss is $\mathbf{L_{CE}}(p(x), y) =$

Algorithm 1: CFA Training

Training Data: Dataset $\mathcal{D}$ of images and labels $\{(\mathbf{x}^i, \mathbf{y}^i)\}_{i=1}^N$

Stage 1: ERM Training

while *not converged* **do**

  $\mathbf{x}, \mathbf{y} \sim \mathcal{D}$; `// Sample datapoint`

  Update weights $\mathbf{w}$ of $\mathbf{f_{erm}} \equiv \phi_{\mathbf{erm}} \circ \mathbf{c_{erm}}$ with:

  $\nabla_{\mathbf{w}} \mathbf{L_{CE}}(\mathbf{f_{erm}}(\mathbf{x}), \mathbf{y})$; `// Backprop.`

end

Causal Feature Generation $\mathcal{D}'$ (a subset of $\mathcal{D}$) for which we will obtain segmentation maps

for $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}'$ **do**

  $\mathbf{A} = xGradCAM(\mathbf{x}), \{\hat{\mathbf{s}}\}^L = \mathbf{f_{sam}}(\mathbf{x})$

  $\{\mathbf{U}\}^{\mathbf{L}} = \mathbf{L_{Dice}}(\mathbf{A}, \hat{\mathbf{s_i}})$; `// Compute scores for each mask output by SAM`

  $k = \text{argmin}_{i \in L}\, \mathbf{U_i}\ \mathbf{m} = \hat{\mathbf{s}_k}$;

  `// Algorithmic suggestion for causal mask`

  `/* Human can select a mask` $\hat{\mathbf{s}_i}$ `different from algorithmic suggestion, where` $i \neq k$ `*/`

end

$\phi \leftarrow \phi_{\mathbf{erm}}$; `// Initialize`

Stage 2: Causal Feature Alignment

while *not converged* **do**

  $(\mathbf{x}, \mathbf{y}, \mathbf{m}) \sim \mathcal{D}'$; `// Sample datapoint`

  $\mathbf{x_c} \leftarrow \mathbf{x} \cap \mathbf{m}$; `// Casual Features`

  Update weights $\mathbf{w}_\phi$ of ϕ with:

  $\nabla_{\mathbf{w}_\phi} \mathbf{L}(\phi(\mathbf{x}), \phi_{erm}(\mathbf{x_c}), \mathbf{y})$; `// Backprop.`

end

Return: $f_{wgo} \equiv \phi \circ c_{erm}$; `// Final model`

$-\sum_{y \in \mathbf{Y}} \mathbb{1}_{y=\mathbf{y}}(\log(p(x))$. λ is used for changing the weight of classification loss.

Inference The resulting model, which we denote as $\mathbf{f_{wgo}} = \phi^* \circ \mathbf{c_{erm}}$, is the final model that is used for inference. We would like to highlight that the model automatically ignores the spurious features, and no segmentation is performed at the inference step.

3. Prior Art

Comparison with Deep Feature Reweighting (DFR) [12]: While both our method and DFR are two-stage algorithms, with the first stage being common ERM training. In the second stage, DFR uses the group labels on a validation dataset to obtain a group-balanced dataset and retrains only the classification layer while keeping the feature extractor fixed. However, we believe the requirement of a "small" group balanced validation in DFR is misleading due to the heavy sample complexity requirement of samples from minority groups. For instance, in the Waterbirds dataset, the prevalence of samples from the minority group(s) is $<$*10*%. In the validation set, *50*% of samples belong to the minority group due to the requirement of equal group weighting. The validation dataset size is *20*% of training data, which translates to the validation set possessing the same number of minority group samples as the training dataset. Indeed, we observe a significant performance degradation (*92*% to *75*%) when the validation set size is reduced from *20*% to *10*%.

In practice, on top of having expertise in identifying groups apriori, collecting sufficient data for feature reweighting from each minority group may pose more of a challenge than simply annotating bounding boxes. Moreover, the data requirement increases exponentially with respect to the number of attributes inducing groups. Let the number of attributes be n_a and the number of classes n_c. If each attribute is restricted to be a categorical variable with k values, the number of samples required would be $k^{n_a} \times n_c \times x$, where x is a percentage of samples relative to the training dataset required in validation. In contrast, the sample complexity of our review task scales linearly concerning the number of classes $(n_c \times x)$.

Relationship to methods utilizing group information While methods utilizing group information can also reduce reliance on spurious features [6, 7, 9, 10, 16], we would like to highlight a couple of drawbacks. These methods require prior knowledge of groups and spurious attributes, which we posit is highly specialized knowledge (like the chest tube example in pneumothorax detection). Secondly, these methods are optimized for improving worst-group accuracy for groups (even if under-represented) in the training set. They cannot guarantee performance across unseen groups, an inevitable practical necessity. This failure mode is highlighted through a novel test dataset utilizing the popular Waterbirds dataset. Overall, we believe these observations make a compelling case to develop methods free of predefined knowledge of induced groups.

4. Experiments

4.1. Datasets

Coloured MNIST dataset: We utilize a binarized version of the semi-synthetic coloured MNIST proposed by [15]. In this dataset, digits "1" and "5" are mapped to classes 0 and 1 respectively. The training dataset is composed of samples with digit "1" correlated strongly with red with pixel intensities in the range $(R_0 = [(115, 0, 0) - (256, 141, 0)]$ and digit "5" with green $(R_1 = [(0, 115, 0) - (141, 256, 0)])$. The test set comprises samples where this spurious correlation with background color is broken.

ImageNet Backgrounds Challenge: Next we validate our method's ability to ignore the spurious background feature

on the challenging ImageNet-9 Backgrounds challenge [8]. This dataset consists of multiple subsets with varying backgrounds to evaluate the impact of the background (typically a non-causal feature) on the classifier's predictions. The dataset is a subset of the Imagenet dataset consisting of 9 coarse-grain classes. Further, to test the reliance of background on the model's prediction, multiple dataset variations are created: (1) *Original*: the images from ImageNet used without modification. (2) *fg-only*: Images with only the foreground present and background zero. (3) *Mixed-rand*: Images with backgrounds replaced by random backgrounds from any class.

Waterbirds dataset: We then utilize the method on the Waterbirds dataset [16] where the target feature is type of bird (waterbird v/s landbird) but confounded by the background (land v/s water). This results in 4 groups of images: waterbirds on water background ($\mathcal{G}_1$), waterbirds on the land background ($\mathcal{G}_2$), land birds on water background ($\mathcal{G}_3$) and land birds on the land background ($\mathcal{G}_4$). ($\mathcal{G}_1$) and ($\mathcal{G}_4$) correspond to majority group while ($\mathcal{G}_2$) and ($\mathcal{G}_3$) are minority groups. We obtain state-of-the-art results on worst-group accuracy (accuracy on the minority groups), thereby demonstrating the applicability of our approach on a benchmark dataset. The distribution of the samples in groups ($\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{G}_3$, $\mathcal{G}_4$) are 73%, 4%, 1% and 22% respectively.

4.2. Baselines

While there are a large number of methods that broadly attempt to improve the worst-case scenario, we compare our method to a few baseline algorithms:

- *ERM* represents conventional training without any safeguards for minority groups.
- *Group-DRO* [16] is the state-of-the-art which uses group information on the training dataset and up-weights worst-group examples during training.
- *Just Train Twice* (*JTT*) [9] is a method that automatically infers the minority group examples on train data but requires group labels on the validation data to tune hyper-parameters.
- *SUBG* utilizes ERM on a random subset of the data where the groups are equally represented [19].
- *Spread Spurious Attribute* (*SSA*) [6] is a method that utilizes the group labels on validation data with a semi-supervised approach that propagates the group labels to training data when group information is unavailable.
- *Deep Feature Reweighting* (*DFR*) [5] is a simple method that utilizes features from an ERM-trained model on a group-balanced validation set to retrain merely the classifier.

4.3. Results

Colored MNIST Dataset: We obtain the baseline empirical risk minimization (ERM) model with a four-layered CNN model. The test accuracy of the ERM model on test data when the spurious correlation in the form of background is collapsed results in a meager *59.1%* as opposed to *98%* if a test dataset was constructed while sustaining the spurious correlations exhibited in the training dataset. We utilize our CFA algorithm to finetune the ERM algorithm utilizing segmentation masks on a subset of training data (*5%*) and obtain an impressive *98%* outperforming previously known state-of-the-art algorithm (ERM+FRR) on this dataset by a large margin. While we acknowledge that ERM+FRR does not require any other additional human effort beyond availability of labels, we believe the difference in performance justifies the necessity of human review in causal feature generation.

Following [15], we quantify the features correlating with the causal feature - the shape of the digit and the spurious feature - background color. A feature from the penultimate layer of the neural network is termed a causal feature or spurious feature if it correlates greater than *90%* with the shape and colour feature, respectively. As we observe in Table 1, the ERM model has an exceedingly high correlation with the spurious feature (background). At the same time, the CFA algorithm minimizes the correlation to the spurious background feature.

Algo.	Number Color,Shape	O/P Corr. Color,Shape	Acc. ID,OOD
ERM	26, 4	0.81, 0.61	99.9%, 59.1%
FRR [15]	26, 4	0.71, 0.65	99.6%, 64.9%
CFA	**30**, **0**	**0.19**, **0.72**	99.6%, **99.2%**

Table 1. **CFA Validation on Coloured MNIST**: We first observe that ERM demonstrates impressive performance on an in-domain (ID) test set while suffering when exposed to an out-of-domain (OOD) dataset where spurious correlations are absent. CFA learns to focus on the causal feature shape while nearly completely ignoring the background features through the correlation metrics compared to ERM and ERM with FRR [15], thereby translating to impressive OOD performance.

ImageNet Backgrounds Challenge: To obtain the baseline ERM model, we finetune an Imagenet-pre-trained ResNet-50 with the *train* split of *Original*. We observe that the accuracy of the finetuned model is *97%* on *Original*, *72%* on *Mixed-rand* and *85%* on *fg-only*. We hypothesize that the performance of *fg-only* represents the entitlement of the classifier based only on the foreground features. For finetuning the algorithm, the causal masks were generated through a) bounding boxes input to SAM (CFA-BB) and b) human review and selection of generated masks (CFA-Min).

We test the Algorithm 1 on the *Mixed-rand* variation to check for improvement in Background reliance. We note

Dataset	Original Model	CFA-BB	CFA-Min
Original	97.6%	97.4%	97.2%
fg-only	85.2%	85.4%	85.6%
mixed-rand	72.2%	78.4%	77.8%

Table 2. Our method successfully learns to ignore the background in the Background Challenge by improving the *mixed-rand* subset.

that our intervention improves the accuracy to *78%* from *72%* without the knowledge of *Mixed-rand* data creation for training. Additionally, we observed that the performance difference between CFA-BB and CFA-Min was only marginal. This experiment demonstrates the generality of our method to ignore spurious features manifested as background.

State-of-the-art results on Worst-Group Accuracy on Waterbirds Dataset: We trained a conventional ERM model with the group-imbalanced training data by finetuning an ImageNet-pre-trained Resnet-50. To finetune the ERM model, we generated masks for causal features in 2 ways: a) obtaining segmentations through bounding box input to SAM (CFA-BB), and b) human review and selection of generated masks (CFA-Min). We chose the subset of the dataset for generating segmentation masks randomly in equal measure from the two classes. We note that this strategy results in very few samples from minority groups being chosen. Secondly, in *96%* of the cases presented for human review in CFA-Min, our algorithm predicted the correct mask and required modifications only in the remaining few cases.[1]. We observed that in spite of the reduced annotation effort, the performance does not degrade significantly compared to annotation through bounding box input to SAM. We finetuned the model for 80 epochs and observed that the worst-group accuracy increases monotonically, thus obliterating the careful hyper-parameter tuning requirement (Fig. 3). This starkly contrasts all state-of-the-art methods based on group information where the validation dataset is heavily used for model selection [19].

To re-emphasize the points made in the above sections, we do not utilize any information about groups – either predefined knowledge of groups or sample-wise group labels on a subset of train/validation data. Further, we perform no careful hyper-parameter finetuning on the validation dataset for model selection, as is the usual case in group-based methods thus far. Hence, we provide a plug-and-play method that can eliminate reliance on spurious features for classification that is, in principle, with a causal learning approach.

CFA Works on Novel Challenging Dataset: Additionally, since our method utilizes segmentation - an explicit form

[1]The code will be updated in Github upon acceptance.

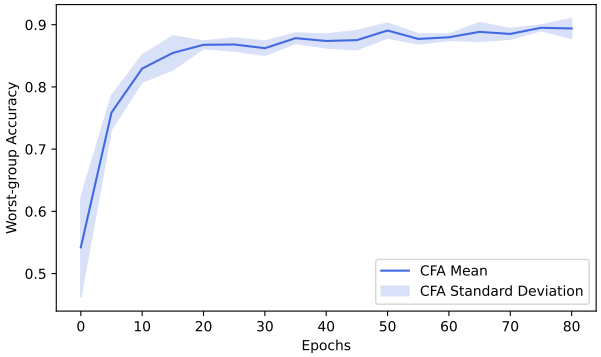

Figure 3. The plot demonstrates a near monotonic increase in worst-group accuracy during CFA on the waterbirds dataset, thus eliminating the need for hyperparameter tuning using validation dataset, as is the case for group-based methods.

Method	Accuracy	
	Worst(%)	Mean(%)
ERM	$72.6/85.5_{\pm 1.0}$	97.3
JTT [9]	$86.7/85.6_{\pm 0.2}$	93.3
Group-DRO [16]	$91.4/87.1_{\pm 3.4}$	93.5
SUBG [19]	$89.1_{\pm 1.1}$	-
SSA [6]	$89.0_{\pm 0.55}$	$92.2_{\pm 0.87}$
DFR [5]	$92.9_{\pm 0.2}$	$94.2_{\pm 0.4}$
CFA-BB (our method)	$\mathbf{93.02}_{\pm 0.1}$	$95.2_{\pm 0.4}$
CFA-min (our method)	$92.62_{\pm 0.1}$	$95.2_{\pm 0.4}$

Table 3. We demonstrate that our method outperforms all other methods on a dataset where the foreground is sufficient for classification. The two numbers in the worst group accuracy represent numbers from different sources, highlighting the variability of the methods ([9], [19]) Further, we do not require any model selection using a validation dataset. Mean accuracy corresponds to overall test accuracy.

Figure 4. Example images from the new test set created with novel backgrounds to test CFA's ability to utilize only causal features. In the above examples, mountains and dams are the backgrounds used.

Figure 5. Weak Labels For Supervising CFA

of causal supervision, our approach sustains performance when inferred with newer types of background. Since group knowledge is necessary for strategies that optimize worst-group accuracy, they do not generalize when presented with images with completely unseen backgrounds. To verify this claim, we create a new testing set (examples in Fig. 4) with novel backgrounds like mountains, dams etc. These backgrounds need not necessarily correspond to water or land. Group-based methods like DFR fail to sustain impressive performance in this out-of-domain (OOD) dataset and obtain a maximum accuracy of *84*%, down from *92*% with known groups. On the contrary, our method provides similar accuracy of *93*% even on the new challenging test set. This experiment demonstrates that CFA improves worst-group accuracy and can sustain performance even when inferred on out-of-domain datasets. This impressive performance can be attributed to the extraction of causal features, which are shown to be robust to domain changes.

5. Discussion

We perform a series of ablation experiments to tease out the different factors that contribute to the performance of our algorithm and better understand the learning process.

5.1. Granularity of Causal Supervision

Effect of Specificity of Segmentation: We have experimented with utilizing inaccurate segmentations in the form of weak annotations on Colored MNIST dataset in the following two ways (See Fig. 5) without converting them to dense annotations: a) cropping out the pixels not under background for input, thereby effectively reducing the number of spurious features and resizing the image and b) Substituting the background pixels with 0 or mean value to avoid utilizing these features for discrimination. However, these methods only improve the worst-group accuracy from *59*% for ERM models to *63*%, which is not substantial, necessitating the requirement to convert weak labels to pixel-wise annotations.

Effect of Number of Segmented Samples: We varied the number of segmented samples from 10% to 40% on the Waterbirds and backgrounds challenge experiments. As reported in Fig. 6, we observe that the performance increases significantly when the number of segmented samples in-

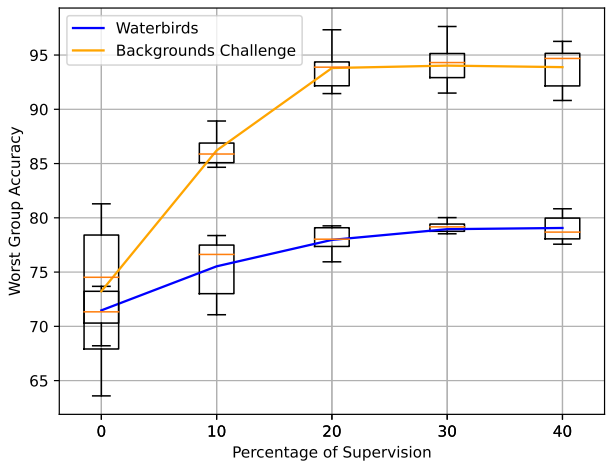

Figure 6. Performance (inter-quartile range) of CFA as a function of the percentage of segmented samples in the training dataset.

creases from 10% to 20% but starts saturating thereafter, demonstrating that a small number of causal localization is enough to restrict the feature extractor from using background features.

Effect of Number of Segmented Samples in Worst-Group: CFA does not utilize group information to determine the samples on which segmentation input is required for the finetuning step. In all our experiments, the samples for segmentation were chosen randomly. This implies that the number of samples from minority groups used for segmentation is extremely small. In the Waterbirds challenge, less than *1*% of data from the worst group were annotated with segmentation. This shows that the success of CFA does not rely on the type of samples on which segmentations were obtained. This flexibility can enable CFA to play a big role in practical applications where it is hard to obtain clear group information. In contrast, group-based methods require identifying samples from each group, which implies one has to continue annotating until a minimum number of samples from all the groups are obtained, which in many cases can be the entire dataset.

5.2. Quality of CFA Features

Saliency of CFA Features: We utilize the Grad-CAM [14] method to visualize the salient regions in the image that the classifier utilizes for prediction. For each image $\mathbf{x}$, two Grad-CAM images are computed, one from $\phi_{\mathbf{erm}}(\mathbf{x})$ and $\phi_{\mathbf{CFA}}(\mathbf{x})$. A few random Grad-CAM examples are shown in Fig. 8, and it can be noted that the saliency map using a model trained using CFA is more causally relevant than the ERM model.

Alignment of Features due to CFA: On the Waterbirds dataset, we visualize the t-SNE plot of the features on the test split (unseen) of the data to verify if the features on the

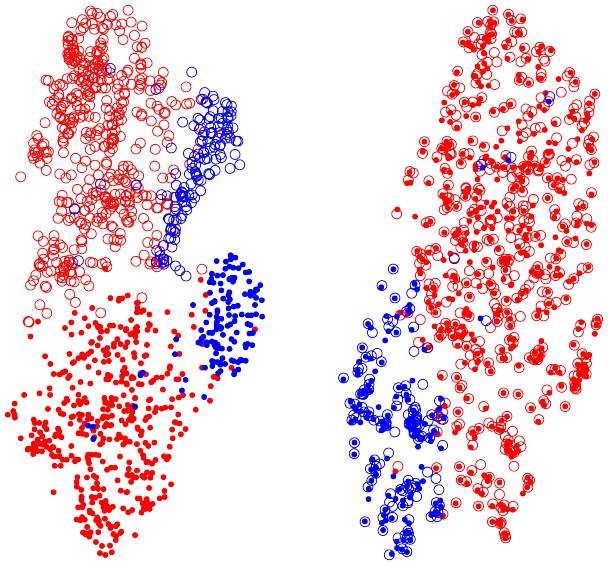

Figure 7. [Best viewed in color] We plot 2D tSNE embeddings of the original image with empty circles and corresponding causal features with solid dots. Red and blue correspond to the two classes - waterbirds and landbirds. The left and right images correspond to representations from ERM and CFA, respectively. We observe that (i) CFA aligns representations of original images with causal counterparts and (ii) improves class discrimination.

original image have aligned with features from the input with only causal features. As seen in Fig. 7, the CFA algorithm increases the alignment between foreground-only images and original images. We also note increased discrimination after alignment. Also, the CFA algorithm utilizes 20% of samples only with causal inputs, showing the ability to extend causal feature extraction to all (test) samples.

6. Conclusion

In this paper, we introduced a method called Causal Feature Alignment (CFA), a method to ignore the spurious background features by utilizing spatial localization of causal features on a subset of training data instead of the group labels used predominantly in the literature. Our experimental analysis shows that CFA can successfully reduce reliance on background by improving accuracy in the worst group across various benchmark datasets.

This is a markedly different research direction compared to the existing popular group-based methods. Further, this has significant advantages in terms of ease of annotations as, (i) it does not require domain expertise to identify groups and instead relies on user input, and (ii) the number of samples impacts accuracy but not the type of samples. It is also a relatively cognitively less intense task to delineate

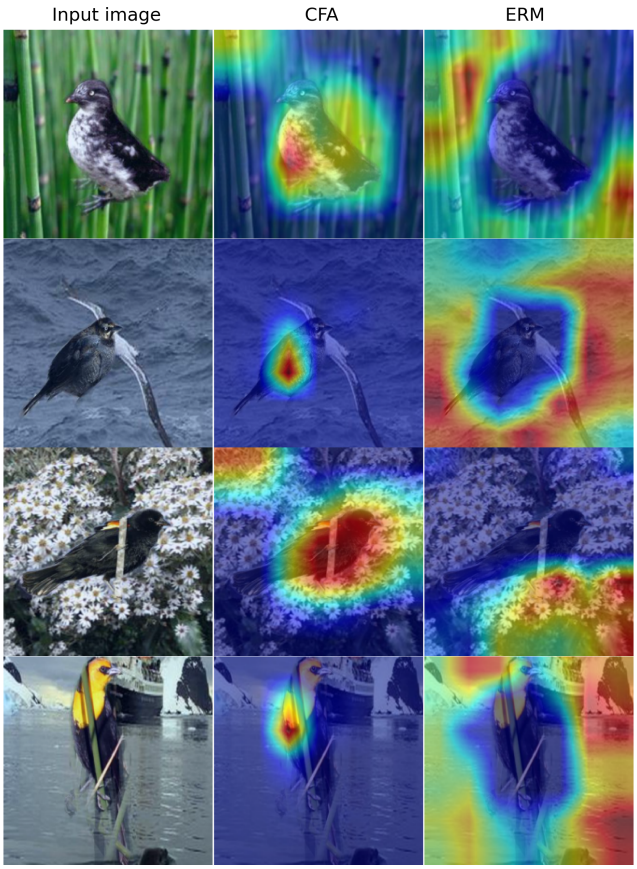

Figure 8. [Best viewed in color] xGradCAM saliency map for images from the waterbirds test set with models trained using CFA and ERM. The red color in the map corresponds to the model's notion of salient features. It can be clearly seen that models trained with ERM often focus on the background (confounder) while predicting, whereas using CFA, the models learn to focus on the bird.

the foreground from the image instead of annotating samples for group labels.

We acknowledge that our method is limited to only problems where the spurious feature is spatially disjoint from the causal feature. Other spurious features manifested in the form of brightness, saturation, etc., cannot be easily overcome using our method. Thus, it is important to develop methods for other manifestations of spurious features that utilize relatively lower cognitive input from humans.

References

[1] Kirillov. A., Mintun. E., Ravi. N., Mao. H., Rolland. C., Gustafson. L., and R. Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[2] Krizhevsky. A., I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[3] Gulrajani. I. and Lopez-Paz. D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2020. 2

[4] Serna. I., A. Morales, Fierrez. J., and N. Obradovich. Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *Artificial Intelligence*, 305:103682, 2022. 1

[5] P. Izmailov, P. Kirichenko, N. Gruver, and A. G. Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022. 5, 6

[6] Nam. J., Kim. J., Lee. J., and J. Shin. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. In *International Conference on Learning Representations*, 2021. 4, 5, 6

[7] Nam. J., Cha. H.and Ahn. S., Lee. J., and Shin. J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 4

[8] Xiao. K., Engstrom. L., Ilyas. A., and A. Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. 2, 5

[9] E. Z. Liu, Haghgoo. B., A. S. Chen, Raghunathan. A., Koh. P. W., and S. Sagawa. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 4, 5, 6

[10] Zhang. M., Sohoni. N. S., Zhang. H. R., C. Finn, and C Ré. Correct-n-contrast: a contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, pages 26484–26516. PMLR, 2022. 4

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

[12] Kirichenko. P. and A. G. Izmailov. P. & Wilson. Last layer retraining is sufficient for robustness to spurious correlations. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. 2, 4

[13] J. Pearl. *Causality*. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009. 2

[14] Selvaraju. R., Cogswell. M., Das. A., Vedantam. R.and Parikh. D., and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3, 7

[15] Addepalli. S., Nasery. A., Radhakrishnan. V. B., Netrapalli. P., and Jain. P. Feature reconstruction from outputs can mitigate simplicity bias in neural networks. In *The Eleventh International Conference on Learning Representations*, 2022. 4, 5

[16] Sagawa. S., Koh. P. W., Hashimoto. T. B., and Liang. P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 4, 5, 6

[17] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020. 1

[18] J Winkler, C Fink, F Toberer, A Enk, T Deinlein, R Hofmann-Wellenhof, Luc Thomas, A Lallas, A Blum, W Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019. 1

[19] Idrissi. B. Y., M. Arjovsky, M. Pezeshki, and D. Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, pages 336–351. PMLR, 2022. 5, 6