Segmentation as Auxiliary Information for Reducing Spurious Correlations

Zev Nicolai-Scanio*
Department of Computer Science
Harvard University
Cambridge, MA 02138
znicolaiscanio@college.harvard.edu

Catherine Yeo*
Department of Computer Science
Harvard University
Cambridge, MA 02138
cyeo@college.harvard.edu

Abstract

Informally, a spurious correlation can be seen as the unexpected dependence of a model on an aspect of the input data that which should not matter. In this paper we build on existing work to develop a proof-of-concept machine learning pipeline more robust against learning spurious correlations. The approach we present uses targeted data augmentation and modifies model architectures to take advantage of auxiliary structural information, specifically image segmentation masks when performing the task of image classification. Our experiments empirically demonstrate that segmentation as input outperforms the baseline, especially for extreme shifts in confounder correlation.

1 Introduction

Models learning spurious correlations that may be present in the training data can have serious detrimental effects to models' abilities to generalize. These issues are of pressing interest to practitioners, because if left uncorrected, they can hinder model efficacy during deployment. Thus there is a current need and use for machine learning models and training procedures that are robust against learning such spurious correlations. However, it is important that attempted solutions do not drastically increase model complexity or require a large amount of additional human effort, say in intricate data labeling or cleaning tasks. In this paper we present and evaluate early work to develop a proof-of-concept machine learning pipeline more robust against learning spurious correlations using targeted data augmentation and lightweight changes to model architectures to take advantage of simple auxiliary structural information.

In particular, we focus on an example image classification problem where there has been existing work on documenting spurious correlations and where there exist well specified baseline results on the negative impacts that these spurious correlations have on model performance. Building upon the existing data sets and methodology established in this previous work, we test two methods for reducing learned spurious correlations.

The first method consists of targeted data augmentation. The idea is that this augmentation will disrupt the model from learning the spurious correlations during training time. In particular, we do this by applying different forms of pixel-level data augmentations to the part of the image responsible for the spurious correlation.

The second method consists of model architecture modification to allow the inclusion of related structural auxiliary information along with the input image. The idea is that the model will internally

*All authors contributed equally to this research and are listed in alphabetical order.

leverage this to learn to avoid the spurious correlations. In particular, we do this with image segmentation information. We also propose and discuss additional potential methods such as requiring the model to perform an additional task during training that requires structural understanding of the image - and we present an argument in support of trying this approach in further work. Namely, that the model will learn internal representations that are less susceptible to spurious correlations.

The paper connects with current work in the area in a few ways. First, it tests the usefulness of segmentation as auxiliary information, which to the best of our knowledge is novel in the field. Secondly, it contributes to the discussion of when auxiliary information is best used as input features or output targets, which is mentioned in [7]. We hope that our proof of concept here can provide a useful, practical example for further work in combating spurious correlations and help contribute to better performance and more trustworthy models without significantly hindering inference-time performance.

2 Related Work

Veitch et al. [5] observes how the causal structure of a problem carries implications for distributional robustness and thus introduces counterfactual invariance as a means to check for spurious correlations. Building on their work, Makar et al [3] uses auxiliary labels at training time to build models that mitigate shortcut learning and improve the generalization of estimators under distribution shifts. As detailed more in sections 3 and 4, we will be similarly using the Waterbirds dataset in our empirical experiments. We will build on Veitch et al. [5] and Makar et al [3] by considering image segmentations as auxiliary labels.

Xie et al. [7] combines the strengths of auxiliary inputs and outputs and establishes the In-N-Out algorithm, which trains a model with auxiliary inputs, pseudolabels all the in-distribution inputs, pre-trains the model on out-of-distribution auxiliary outputs, and finetunes the model with the pseudolabels. In doing so, In-N-Out outperforms auxiliary inputs or outputs alone on both forms of error.

3 Datasets

The primary dataset used for experimentation is the Waterbirds dataset, as first constructed in [4], where bird photographs from the Caltech-UCSD Birds-200-2011 dataset [6] are combined with image backgrounds from the 2017 Places dataset [8]. Each bird is labeled as either a waterbird or landbird and placed on either a water background or a land background. There are 4795 training examples.

How the Waterbirds dataset was constructed makes it natural to consider and observe segmentation, as the outline of the bird relative to background is automatically well defined during the construction of the data. The pipeline for data generation can be seen in Figure 1.

We split our dataset into training, validation, and test sets and used the same splits across all our experiments. We constructed our test sets across confounding percentages from 50% to 95% in five-percent intervals, where the confounding percentage is the correlation between the bird type and background type. Equivalently, given predicted label Y (i.e. the foreground object) and an auxiliary label V (i.e. the background image), the confounding percentage is $P(Y|V)$, which we observe at test time.

	Land Bird	Water Bird
Train	3705	1090
Validation	910	289
Test	4510	1284

Table 1: The number of land bird and water bird images in the training, validation, and test dataset splits.

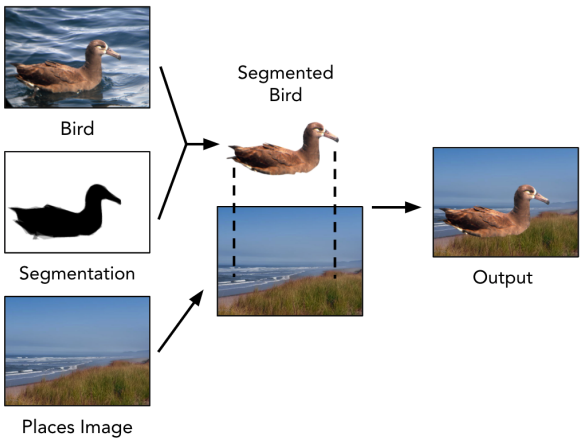


Figure 1: Our data generation pipeline, synthesizing the Caltech-UCSD Birds-200-2011 dataset and the 2017 Places dataset.

4 Experiments

We conduct a series of experiments first to establish a baseline and then to test and benchmark our methods. Each of the methods is described in detail in the following subsections.

The baseline setup is modeled after the empirical setup in [3], where we take our constructed Waterbirds dataset to examine how a model trained under a specific distribution can generalize to various test distributions. We use our version of the Waterbirds dataset as outlined in Section 3, however, any differences compared to [3] are minor; results in Section 4 for the baseline are overall similar to those obtained in [3].

In all of our experiments, the model's architectural backbone is a ResNet-50 model [2] pre-trained on ImageNet which we then finetune on our datasets. Any modifications to the model are explained in detail in the subsection for each relevant experiment.

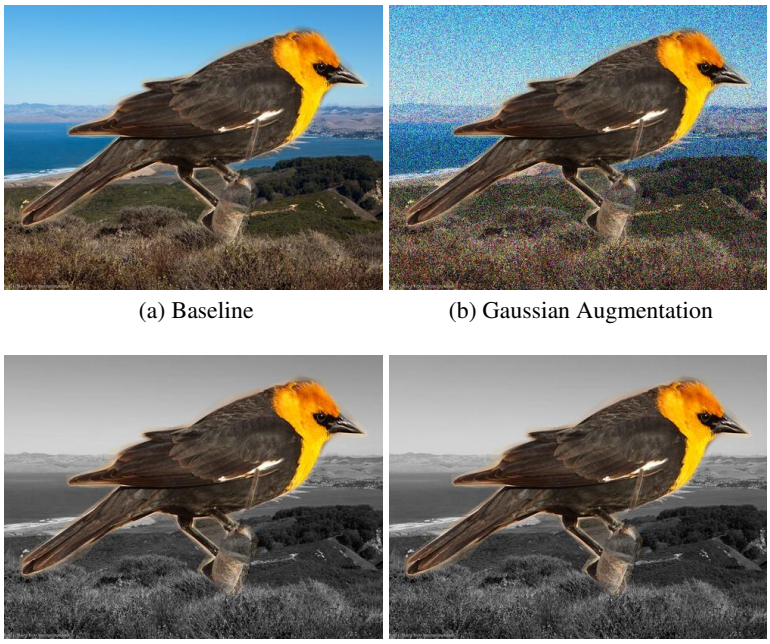
4.1 Targeted Data Augmentation

The first method consists of targeted data augmentation. Our hypothesis behind this experiment is that specific data augmentations will disrupt the model from learning the spurious correlations during training time.

We do this by applying different forms of pixel-level data augmentations to the part of the image responsible for the spurious correlation. Concretely, during the construction of the dataset, we apply the data augmentation to the background image from the Places dataset before the segmented bird is added.

The first data augmentation that we tested was to apply Gaussian noise. However, while Gaussian noise is very noticeable to the human eye, we observed that it does not change certain macro properties of the image, such as the fact that land images contain a large amount of green, while water images contain a large amount of blue. To this end, we also applied a grayscale augmentation and, furthermore, since standard grayscale conversion weights the red, green, and blue channels differently, an even-weighted "decolored" augmentation.

We apply the augmentation during training but do not apply them during test time. We believe that this is realistic, since data labeling could specify the difference between the bird and the backdrop, but we would not have access to this information during deployment. Figure 2 below shows examples of these augmentations.



(a) Baseline (b) Gaussian Augmentation

(c) Grayscale Augmentation (d) Decolored Augmentation

Figure 2: Examples of a segmented yellow-headed blackbird being placed on a water backdrop image from the Places dataset as well as with three different augmentations (Gaussian noise, grayscale, decolored) applied to the same background image. (Note: that the grayscale and decolored backdrops look extremely similar to the human eye.)

4.2 Segmentation as Input

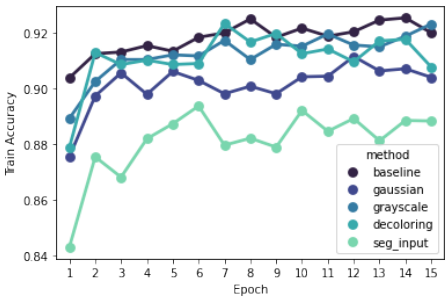
The second method consists of modification to the model architecture to allow the inclusion of related structural auxiliary information along with the input image. Our hypothesis here is that the model will internally leverage the auxiliary information to learn to avoid the spurious correlations.

In particular, we do this with image segmentation information. We believe that using such information both during train time and test time as it requires segmentation but not labeling of bird versus background. To use this additional information, we modified the model to have an additional channel in the first convolution layer, while otherwise using the pre-trained weights as before. We then finetune the model on our dataset.

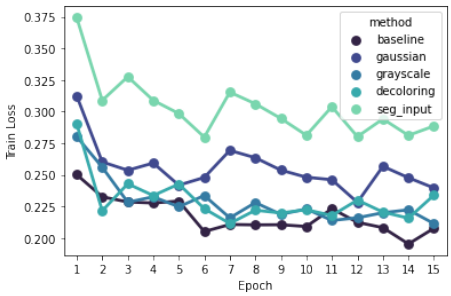
5 Results

We evaluate our performance on the following five methods: the baseline setup, the baseline setup with Gaussian noise applied to the image backgrounds, the baseline setup with grayscale applied to the image backgrounds, the baseline setup with "decoloring" (our evenly weighted form of grayscale) applied to the image backgrounds, and the segmentation as input setup.

In all the runs performed on both the training and validation datasets, for all five methods, segmentation as input had the lowest accuracy (ranging between 84% to 90%) and highest loss, while the baseline setup had the highest accuracy (ranging between 90% and 93%) and lowest loss. We observe that the accuracy and loss results are far closer across the five methods in the validation runs than in the training runs. Figure 3 illustrates the progression of training accuracy and loss over a time span of 15-epochs, and Figure 4 depicts the progression of validation accuracy and loss over the same epoch span.

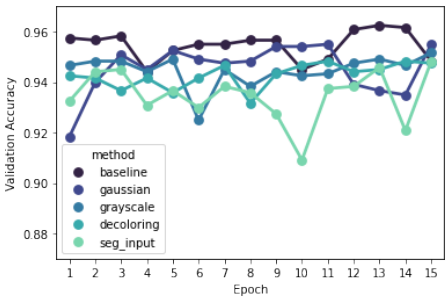


(a) Training accuracy over time span of 15-epochs.

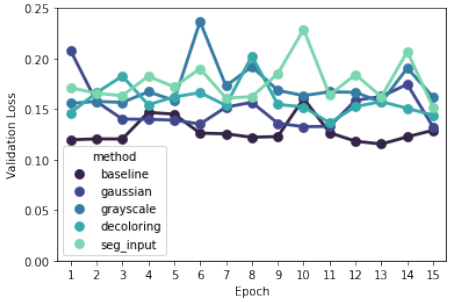


(b) Training loss over time span of 15-epochs.

Figure 3: The progression of training accuracy (a) and loss (b) over a time span of 15-epochs. Segmentation as input had the lowest training accuracy and highest loss, while the baseline method performed the best.



(a) Validation accuracy over time span of 15-epochs.



(b) Validation loss over time span of 15-epochs.

Figure 4: The progression of validation accuracy (a) and loss (b) over a time span of 15-epochs. Segmentation as input again had the lowest validation accuracy and highest loss, although far closer to the remaining methods (as compared to the training run), while the baseline method again performed the best.

In our experimental runs conducted on the test dataset, segmentation as input significantly outperformed all four other methods, with higher accuracy and lower loss, for confounding percentages between 50% to 80% (see Figures 5 and 6). This indicates that including the additional segmentation mask input channel thus improves our results, providing us optimism that auxiliary structural information (i.e. the segmentation mask) can serve as useful knowledge on what part of the image our model should care about.

As part of the analysis investigating why the different types of augmentation performed the way that they did, we constructed a histogram (see Figure 7) of the distribution of distances between land and water background images in the baseline versus under the different augmentations. Specifically, we chose our distance metric to be the pixel-wise Euclidean distance between randomly sampled pairs of land and water backgrounds for each method or augmentation.

In general, one may expect that if an augmentation was reducing the distances between the groups, and thus making the difference less distinct, the model may have less strong of a signal from the spurious correlation. However, we see that while the augmentations do shift the distribution towards smaller distances, the augmentations do not perform better than the baseline setup (for training, validation, and test data).

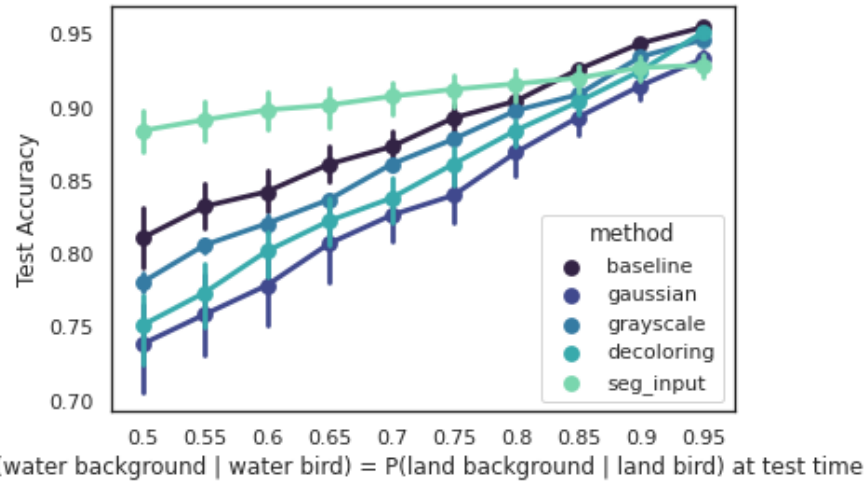


Figure 5: Test accuracy over confounding percentages of 50% to 95%, evaluated at five-percent intervals. The x-axis shows $P(Y|V)$ at test time.

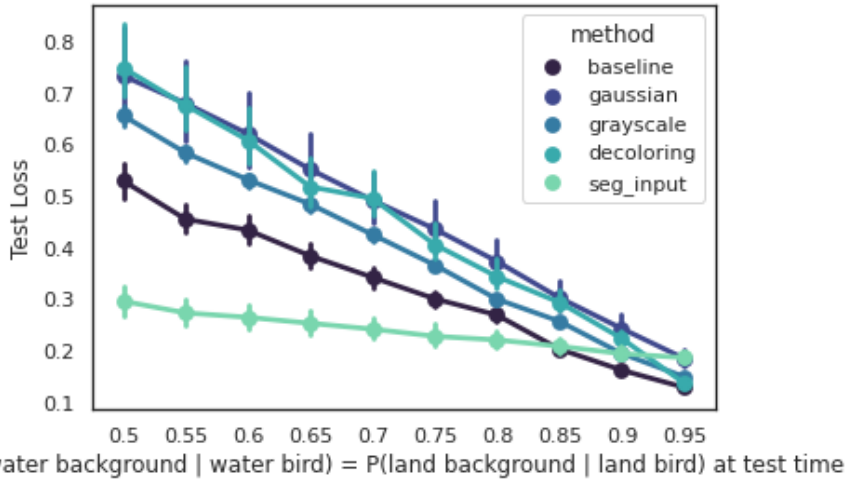


Figure 6: Test loss over confounding percentages of 50% to 95%, evaluated at five-percent intervals. The x-axis shows $P(Y|V)$ at test time.

One potential reason for this phenomenon is that the two types of background images (water vs. land background image) are still different enough that effectively it has no difference in terms of the model's ability to learn the spurious correlation.

Another potential confounding factor is that while the augmentations reduce the distance between the two place types, they may also reduce the variation within the types themselves, which could lead to the model more easily learning the spurious correlation. This could potentially be tested by visualizing the way that the images are clustered using dimensionality reduction techniques.

6 Discussion and Future Work

In this paper, we proposed and implemented two methods to reduce the learning of spurious correlation by machine learning models and tested them in a synthetic yet representative experimental context. Despite various different configurations, the augmentation-based method was unable to outperform the baseline. The method that modified the model to take advantage of auxiliary segmentation data,

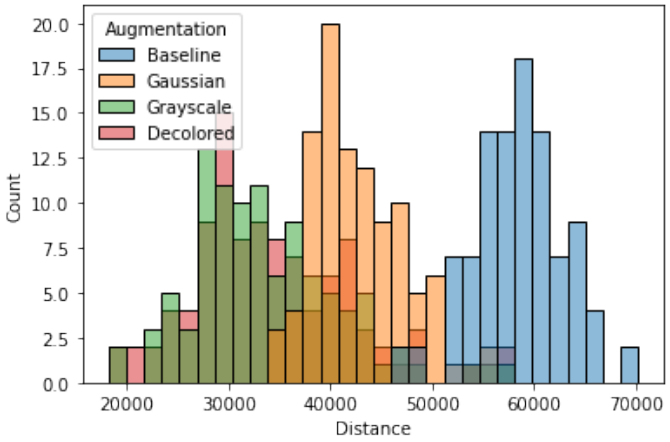


Figure 7: Histograms demonstrating the pixel-wise Euclidean distance between randomly sampled pairs of land and water backgrounds for each method or augmentation.

however, was able to significantly outperform the baseline, especially for extreme shifts in confounder correlation.

In the following subsections we discuss ideas for further work including an additional method to try and experiments to additionally investigate the current methods.

6.1 Segmentation as Output

We are interested about what happens when we consider segmentation masks as an output, instead of only as an input. Our hypothesis is that the additional segmentation task will result in internal representations that are more robust to the confounder.

To achieve this, a multi-headed model needs to be constructed. Segmentation architectures generally consist of both a convolutional network and a deconvolutional network. To remain consistent with the rest of this paper's experiments, one can use ResNet-50 as the convolutional network representation, which takes care of both the segmentation and the classification, and then add a deconvolutional network for the remainder of the segmentation process. Google's DeepLabv3 [1] model does exactly this, proposing a semantic segmentation architecture that is built using ResNet blocks, so DeepLabv3's model architecture seems ideal to reference and adapt for this future experiment.

6.2 Real World Stress Testing

It could also be useful to run similar experiments to this paper but conducted on a real-world, non-synthetic data set. For example, in this paper we used ground truth segmentation masks that were part of the data generation pipeline. In comparison, in deployed contexts, practitioners would most likely be working with the outputs of either a custom or pre-trained image segmentation model. While we believe that any differences would be negligible, such additional experiments confirming these results could be valuable.

6.3 Understanding Model Properties

Investigation as to the differences in what and how the models are learning between the baseline and segmentation as input methods could help to explain the success of this approach and aid in the design of future methods. For example, sensitivity analysis or saliency maps could help to determine whether the model with access to the auxiliary segmentation information is interacting differently with the visual features of the bird than the baseline model without access to this information.

References

[1] Chen, L., Papandreou, G., Schoff, F., & Adam, H. (2017) Rethinking Atrous Convolution for Semantic Image Segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[2] He, K., Zhang, X., Ren, Sq., & Sun, J. (2015) Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., & D'Amour, A. (2022) Causally Motivated Shortcut Removal Using Auxiliary Labels. 25th International Conference on Artificial Intelligence and Statistics (AISTATS).

[4] Sagawa, S., Koh, P.W., Hashimoto, T.B., & Liang, P. (2021) Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. International Conference on Learning Representations.

[5] Veitch, V., D'Amour, A., Yadlowsky, S., & Eisenstein, J. (2021) Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests. Conference on Neural Information Processing Systems.

[6] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011) The Caltech-UCSD Birds-200-2011 Dataset. Computation & Neural Systems Technical Report, 2010-001.

[7] Xie, S.M., Kumar, A., Jones, R., Khani, F., Ma, T., & Liang, P. (2021) In-N-Out: Pre-Training and Self-Training using Auxiliary Information for Out-of-Distribution Robustness. International Conference on Learning Representations.

[8] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017) Places: A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

A Appendix

A list of each participant's contribution if it is a group project:

- Zev: developed data pipeline to generate all datasets and augmentations and performed distance histogram analysis.
- Catherine: built ML pipeline to finetune ResNet-50 on all datasets, expanded the model to take in segmentations as a fourth-channel input, and generated data visualizations.
- Both equally contributed to creating the poster, writing this paper, and conducting code review (i.e. debugging).

This project does not overlap with ongoing research work for either of us.

Skills acquired in this project:

- Creative and persistent problem solving.
- Division of a research problem into multiple levels of experimentation.
- Project iteration and development in response to preliminary experimental results.
- Synthesis of ideas across papers and research domains, in both the research question construction and solution development processes.
- Building upon and interfacing with existing academic and industry code bases.