1. Appendix

1.1. Colored MNIST Dataset

We closely followed the steps from [3] to re-create the semi-synthetic Colored MNIST dataset. Digits "1" and "5" are assigned classes 0 and 1, respectively. Shapes from class 0 are superimposed on a background of color red, whose range is $R_0 = [(115, 0, 0] - [256, 141, 0]$ while shapes from class 1 are superimposed on color green, whose range is $R_1 = [(0, 115, 0) - (141, 256, 0)]$. The overlap in background colors $[(115-141, 115-141, 0)]$ is deliberate to ensure that while the background is highly correlated (0.76), the correlation is not 1.

1.1.1 Weak-CMNIST: Colored MNIST Dataset For Evaluating Effect of Size of Spurious Features

We created a semi-synthetic dataset to study the impact of the size of spurious features on a model's propensity to utilize these features. In each data sample, unlike in the case of colored MNIST dataset, where the digit is placed to cover a large portion of the image, we place the digit in the bottom right quadrant ((14×14) of the (28×28) image (See Figure 1). The authors in [2, 3] empirically showed that the spurious features (generally simpler) are replicated many times in the representations learnt by the network. Further, [2] showed empirically that a network trained on samples with both spurious and core features infers with a satisfyingly high accuracy when presented with samples containing only core features. In these studies, spurious features are treated as binary variables with the possibility of either presence or absence. Our dataset, Weak-CMNIST enables us to further investigate the effect of a reduced number of spurious features on the network.

2. Foundation Models on Waterbirds dataset

We utilize the recently released foundation model, Segment Anything (SAM) [1] to obtain the foreground for a subset of images in the Water dataset. However, since the SAM model is agnostic to the type of object under segmentation, utilizing a weak input in the form of either a bounding box or click is important to be aware of the foreground. We utilize a bounding box as input, generating the bounding boxes from the segmentations with slight jitter in the coordinates of the bounding box to simulate human input. A few sample images along with predicted masks overlaid is shown in Fig. 2

3. Results on Weak-CMNIST

The detailed results on Weak-CMNIST are presented in the table below:

Dataset	Original Model	CFA
Original	97.6	97.4
fg-only	85.2	85.4
mixed-rand	72.2	**78.4**

Table 1. Our method successfully learns to ignore the background in the Background Challenge by improving the *mixed-rand* subset.

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.

[2] 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.

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

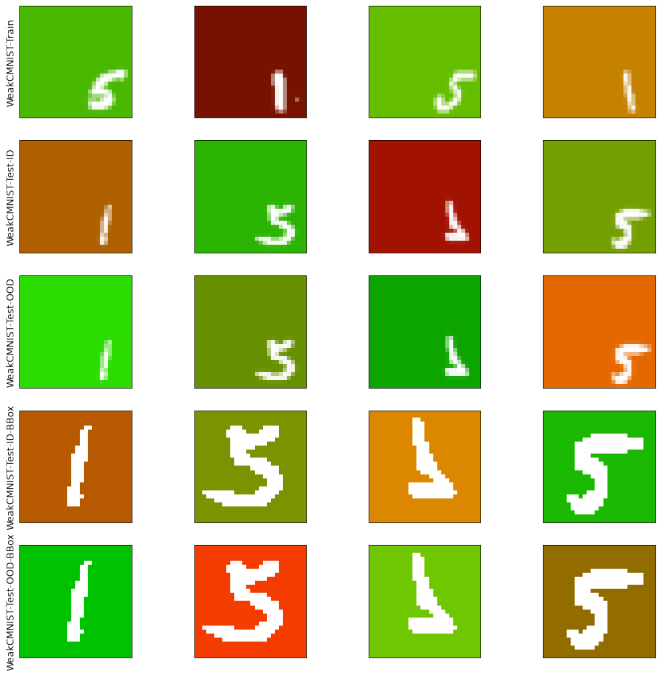

Figure 1. [Best viewed in color] The different Variations of Colored MNIST generated to empirically test the effect of size of spurious features on the effect of model performance. Row 1 (Weak-CMNIST-Train) are example images utilized for training. Note the correlation between digits and background colours. The second row (Weak-CMNIST-Test-ID) displays examples from an in-domain test set. Row 3 (Weak-CMNIST-Test-OOD) in the following row denotes examples from an out-of-domain distribution. Note that the correlation between background colors and digits is tampered with in this dataset. The subsequent row, Row 4 (Weak-CMNIST-Test-ID-BBox) denotes samples where the availability of weak labels in the form of a bounding box enables us to limit the region of interest to the digit. In this dataset, the digit's correlation with the background is maintained as in the case of the training dataset. Finally, the last row (Weak-CMNIST-Test-OOD-BBox) is a case of limiting the image to the shape through a bounding box. However, the correlation of the digit and background colour is corrupted making this an out-of-domain case.

Figure 2. [Best viewed in color] The segmentations on the Waterbirds dataset were obtained by utilizing a weak prompt like bounding box. The segmentation masks are overlaid in blue.